Articles:

Assessing AI Output in Legal Decision-Making with Nearest Neighbors

Timothy Lau* and Alex Biedermann†

ABSTRACT

Artificial intelligence ("AI") systems are widely used to assist or automate decision-making. Although there are general metrics for the performance of AI systems, there is, as yet, no well-established gauge to assess the quality of particular AI recommendations or decisions. This presents a serious problem in the emerging use of AI in legal applications because the legal system aims for good performance not only in the aggregate but also in individual cases. This Article presents the concept of using nearest neighbors to assess individual AI output. This nearest

* Federal Judicial Center, Research Division, Thurgood Marshall Federal Judiciary Building, Washington DC. – The views expressed in this Article are of the author alone and do not represent the views of the Federal Judicial Center.

† University of Lausanne, Faculty of Law, Criminal Justice and Public Administration, School of Criminal Justice, 1015 Lausanne-Dorigny (Switzerland). Visiting Researcher at Northwestern University Pritzker School of Law, Chicago IL. – Alex Biedermann gratefully acknowledges the support of the Swiss National Science Foundation through grant BSSGI0_155809.

The authors thank David Kaye of Penn State Law, Edward Imwinkelried of UC Davis, Ronald Allen of Northwestern University Pritzker School of Law, Brandon Garrett of Duke University School of Law, Ian Evett CBE of Principal Forensic Services, and Clare Lau of Johns Hopkins Applied Physics Laboratory for their helpful comments and suggestions.

neighbor analysis has the benefit of being easy to understand and apply for judges, lawyers, and juries. In addition, it is fundamentally compatible with existing AI methodologies. This Article explains how the concept could be applied for probing AI output in a number of use cases, including civil discovery, risk prediction, and forensic comparison, while also presenting its limitations.

Table of Contents

I. INTRODUCTION

The accelerating adoption of AI to generate recommendations or decisions is starting to have a serious impact on the economy and in our daily lives. Unsurprisingly, it has prompted endless discussions in the legal literature, from how it may change the legal practice to how substantive law should adapt to widespread AI usage. What is missing, however, is a discussion about how to assess AI output in the legal context. This is especially important because legal decision-makers will be called upon to look into the quality of decision-making assisted or automated by AI.

The technical literature has discussed some of these issues. Mathematical metrics such as "recall" and "precision" exist for the evaluation of the quality of AI output.[1] But those metrics are aggregate measures for evaluating the *overall* performance of AI. They are not always applicable to the legal context, which often requires evaluation of the quality of a *single* recommendation or decision.[2]

Consider, for example, a judge receiving a recidivism assessment about a particular convict from an AI-driven risk prediction[3] instrument ("RPI"). Although the overall quality of the AI recidivism assessments is important information, it is no less valuable for the judge to have some sense of the quality of the particular assessment for that particular convict.

In addition, these metrics may not be intuitive to legal decision-makers. Things may change in the future, but any fair evaluation of judicial systems today must conclude that the decision-makers from jurors to judges are, as a whole, not highly skilled in interpretation of data and AI output in particular. It is therefore helpful to have metrics for evaluating AI output that are easy for laypersons to understand.

The proposal of alternative metrics, however, is not a simple task.[4] After all, metrics need to be technically feasible and consistent with the methodologies of AI.

In this Article, we propose the notion of the nearest neighbor analysis ("NNA") to assess AI output geared for the legal system at an individual level.[5] Suppose that an AI makes a recommendation or renders a decision

---

1. *See, e.g.*, STUART RUSSELL & PETER NORVIG, ARTIFICIAL INTELLIGENCE: A MODERN APPROACH 869 (3d ed. 2010); KEVIN D. ASHLEY, ARTIFICIAL INTELLIGENCE AND LEGAL ANALYTICS: NEW TOOLS FOR LAW PRACTICE IN THE DIGITAL AGE 114 (2017); Kai Ming Ting, *Precision and Recall*, *in* ENCYCLOPEDIA OF MACHINE LEARNING AND DATA MINING 990, 990 (Claude Sammut & Geoffrey I. Webb eds., 2d ed. 2017).

2. *See* David L. Faigman et al., *Group to Individual (G2i) Inference in Scientific Expert Testimony*, 81 U. CHI. L. REV. 417, 420 (2014) (discussing "the challenge of reasoning from group data to decisions about individuals").

3. Although ubiquitously used to refer to AI output, the term "prediction" is usually misleading because it suggests a deterministic assertion about the future. But AI output involves uncertainty—most often in a probabilistic sense—so the output is more appropriately, though less stylishly, termed "prevision" or "forecast." *See* EUR. COMM'N FOR THE EFFICIENCY OF JUSTICE, EUROPEAN ETHICAL CHARTER ON THE USE OF ARTIFICIAL INTELLIGENCE IN JUDICIAL SYSTEMS AND THEIR ENVIRONMENT 30 (2018) [hereinafter EUROPEAN ETHICAL AI CHARTER]; Alex Biedermann et al., *Prediction in Forensic Science: A Critical Examination of Common Understandings*, 6 FRONTIERS PSYCHOL. 737 (2015), *available at* https://bit.ly/2wLuuQE.

4. Though the ultimate hope in AI research is to create explainable AI that is capable of defending individual decisions or recommendations, such AI does not exist today and likely will not be available over the medium term. *See* U.S. GOV'T ACCOUNTABILITY OFFICE, GAO-18-142SP, ARTIFICIAL INTELLIGENCE: EMERGING OPPORTUNITIES, CHALLENGES, AND IMPLICATIONS 18–19 (2018).

5. There may be times when the legal system will be required to evaluate AI output that is not geared towards the legal process. For example, a court may be asked to look into a crash of a self-driving car caused by a particular turn taken by the car. Although the focus

for a particular legal problem, and the legal decision-maker seeks to understand the trustworthiness of the AI output. In simplistic terms, NNA seeks to assist with this inquiry by searching for precedent cases with features similar to the legal problem at hand so as to allow for a comparison.

The concept of nearest neighbors is old and well-established in fields such as physics and computer science.[6] Here, we explain how nearest neighbors can be incorporated into legal analysis in a way comparable to the established use of precedent cases by lawyers. The added value of NNA is that it provides a focused and empirical examination of specific AI output beyond standard aggregate AI performance metrics. Our perspective also has the benefit of being explainable in principle and implementable with the many AI algorithms already in existence.[7] Our purpose is not to promote NNA as an exclusive assessment template, as it is by no means an all-encompassing, perfect concept. However, we propose nearest neighbors as additional information that legal users of AI could ask for that may be helpful in their decision-making.

Some of the mainstream discussion and controversy portray AI as a threat to legal practice, in particular, regarding the extent to which AI may or ought to influence legal proceedings and decision-making. In this Article, we take a different view. We start by recognizing the reality that AI is *already* being used and referred to at various instances in the legal process. This, for us, is a reason to focus on how to suitably comprehend AI output at an individual level. Although many take a rather defensive approach by challenging algorithms on aspects such as robustness, error rates, bias, and transparency, we choose a more proactive perspective. We seek to encourage participants in the legal process to request additional AI output—specifically, nearest neighbor data—that is not commonly delivered or requested today. Scrutinizing AI and its output in legal applications at an individual level should contribute to a better informed, more transparent, and more defensible practice, which is preferable to passive attitudes that allow AI to "take over" the legal arena.

---

of this Article is on AI output for legal applications, the concept of NNA may also be used by legal decision-makers for evaluating AI output in other contexts as well.

6. *See* DAVID L. POOLE & ALAN K. MACKWORTH, ARTIFICIAL INTELLIGENCE: FOUNDATIONS OF COMPUTATIONAL AGENTS 321 (2d ed. 2017); KEVIN P. MURPHY, MACHINE LEARNING: A PROBABILISTIC PERSPECTIVE 16 (2012).

7. One benefit of NNA is that the analysis can be conducted without the need for source code. An aspect related to the use of algorithms in the law, but not dealt with in this Article, is the availability and review of computer source code, especially in cases where the code is used to generate evidence against defendants. For a general discussion of this subject matter, see, for example, Edward J. Imwinkelried, *Computer Source Code: A Source of the Growing Controversy Over the Reliability of Automated Forensic Techniques*, 66 DEPAUL L. REV. 97, 111–30 (2016).

This Article is structured as follows. Part II provides a brief overview of AI and its incorporation into legal decision-making. Part III begins by explaining problems related to conventional methods of evaluating AI output. Part III also explores the concept of a nearest neighbor using simple illustrations, and then discusses how NNA can be used to assess AI output in general. Part 0 then provides a comparative perspective by investigating some selected legal applications that rely to varying extents on the use of AI and that exhibit different stages of technical development. Further, Part IV provides suggestions for using NNA in evaluating AI output within these specific areas.

II. AI USAGE IN LEGAL DECISION-MAKING

A. Definition of AI

Any discussion of AI must first acknowledge that there is no universally accepted definition of AI.[8] It is possible to define AI from a pragmatic point of view, based on whether the AI is used for purposes that require "intelligence:" "a computerized system that exhibits behavior that is commonly thought of as requiring intelligence."[9] Another way to define AI is based on whether the AI is actually "intelligent:" "A set of scientific methods, theories and techniques whose aim is to reproduce, by a machine, the cognitive abilities of human beings."[10]

In this Article, we will use the former, that is, the pragmatic definition of AI. This use-based definition of AI is relatively broad. On one hand, it captures the many types of applications which everyone at this time would think of as AI, such as self-driving cars and poker-playing programs. On the other hand, it includes *algorithms*[11] that may seem relatively "unintelligent," such as brute force optimizations, or may seem like routine "data processing," such as solving equations.

A broad, use-based definition of AI is helpful when discussing AI for legal applications. The general users of AI within the legal field are not likely to be particularly technologically savvy. They may treat AI tools as "black box" oracles without discrimination as to the actual sophistication of the underlying algorithm because they lack the training or capability to do so. For example, such users may use an AI tool based on logistical regression as they may use an AI tool based on neural networks. They may

8. *See* NAT'L SCI. & TECH. COUNCIL, COMM. ON TECH., EXEC. OFFICE OF THE PRESIDENT, PREPARING FOR THE FUTURE OF ARTIFICIAL INTELLIGENCE 6 (2016).

9. *Id.*

10. EUROPEAN ETHICAL AI CHARTER, *supra* note 3, at 69.

11. An "algorithm" is a sequence of instructions that a computer can execute, starting from an initial point and input information and leading to a final state. *See, e.g.*, Conor Ryan, *Computer Algorithms in* ENCYCLOPEDIA OF PHYSICAL SCIENCE AND TECHNOLOGY 507, 507 (3d ed., 2003), *available at* https://doi.org/10.1016/B0-12-227410-5/00840-1.

do so even though the two tools differ greatly in terms of technology and may yield different results under certain circumstances. This Article is concerned with providing general suggestions to legal users of AI about how to evaluate AI output in legal applications, and therefore uses a definition of AI consistent with their use of AI.

B. Gradations of Automation and the Need for Assessment Tools

Within legal applications, the task that requires intelligence is decision-making. And in this field, AI are computer tools to assist or automate decision-making. Before we begin, we need to more fully contextualize what AI is used for in legal decision-making. From a conceptual level, decision-making can be thought of as comprising two steps, one of *evaluation* and one of *decision*. Evaluation is the assessment of the relative plausibility or probability of competing propositions based on the available evidence. After evaluation comes *decision*, the acceptance of a proposition as a conclusion. Absent a *decision criterion*, no defensible conclusion can be made.

The need to distinguish between evaluation and decision arises from uncertainty.[12] Where there is uncertainty, legal decision-makers cannot make decisions based their understanding of the evidence alone. Instead, they must also consider their preferences among decision consequences.[13] As a result, a decision criterion fundamentally includes value judgments.

Consider a situation in which there is a fingermark of unknown source ("U") found on the crime scene. There is a defendant whose reference fingerprint ("K") is available. The following figure illustrates the conceptual difference between evaluation and decision in the context of determining whether the defendant *is* the source of U.

12. If there were no uncertainty at the time a decision had to be made, there would be no decision problem. It would be possible to directly select the course of action that would lead to an accurate outcome or, more generally, the best decision consequence given the known state of affairs.

13. *See* Eric J. Horvitz et al., *Decision Theory in Expert Systems and Artificial Intelligence*, 2 INT'L J. APPROXIMATE REASONING 247, 253–54 (1988).

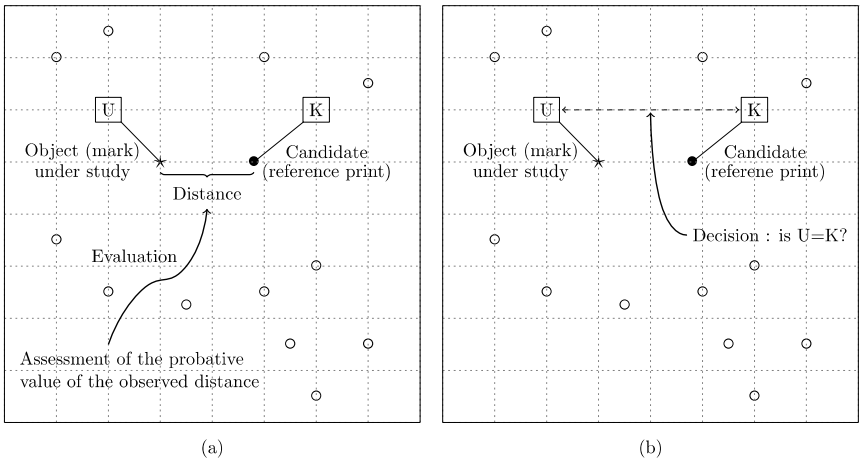


Figure 1: Illustration of the difference between *evaluation* and *decision* in the context of forensic comparison.

In the evaluative step, the differences and similarities between U and K are measured. These measurements are assessed for whether and to what extent they support the hypothesis that the defendant is the source of U. One may, for example, be drawn to think that the defendant is the source of U because one observes many similarities and few unexplainable differences between U and K.

In the decision step, the legal decision-maker must decide whether the defendant *is* the source of U. In doing this, the decision-maker must answer the key question of whether he or she *should* do so. The answer depends on the observed distance between the collected mark and the reference print. But it also incorporates value determinations. For example, the decision-maker must consider the drawback associated with a false identification compared to the drawback associated with a false non-identification. To wrongly associate defendant with U is after all to "miss" the true source of the fingermark. If the loss or damage resulting from a false identification were particularly severe, the decision-maker might decide *not* to identify the defendant as the source of U, despite a strong belief in the truth of the proposition.[14]

14. *See* Alex Biedermann & Joëlle Vuille, *Understanding the Logic of Forensic Identification Decisions (Without Numbers)*, 2018 SUI-GENERIS S. 397, S. 404–05 (2018); *see also* Simon A. Cole & Alex Biedermann, *How Can a Forensic Result Be a "Decision"?: A Critical Analysis of Ongoing Reforms of Forensic Reporting Formats for Federal Examiners*, 57 HOUS. L. REV. 551, 566–70 (2020).

This complication of decision under uncertainty attends virtually all situations of legal decision-making. Though the fundamental ingredients of decisions under uncertainty are widely agreed, opinions on how to make decisions in light of these ingredients are diverging.[15] Formal methods based on decision theory, which incorporates probability as a measure of uncertainty, were proposed as early as the 1960s[16] and have been heavily debated since then.[17] Whatever decision model is used, it is still necessary to choose and implement a particular decision framework, otherwise no decision is made.

This distinction between evaluation and decision may seem extremely formalistic. But it is crucial to understand the distinction in connection with the present state of AI technology. AI scientists recognize assisted or automated decision-making itself as a spectrum, defined in terms of the amount of human input necessary to complement the AI output. For example, the Society of Automotive Engineers ("SAE") has provided the following diagram to illustrate the graduations of assistance and automation[18] in the context of driving:

15. *See* Ronald J. Allen, *Artificial Intelligence and the Evidentiary Process: The Challenges of Formalism and Computation*, 9 ARTIFICIAL INTELLIGENCE & L. 99, 108–14 (2001).

16. *See generally* John Kaplan, *Decision Theory and the Factfinding Process*, 20 STAN. L. REV. 1065 (1968) (explaining how decision theory may be used in legal factfinding in criminal trials).

17. *See, e.g.*, Richard S. Bell, *Decision Theory and Due Process: A Critique of the Supreme Court's Lawmaking for Burdens of Proof*, 78 J. CRIM. L. & CRIMINOLOGY 557, 557–58 (1987); D.H. Kaye, *Clarifying the Burden of Persuasion: What Bayesian Decision Rules Do and Do Not Do*, 3 INT'L J. EVIDENCE & PROOF 1, 1–4 (1999); Richard O. Lempert, *Modeling Relevance*, 75 MICH. L. REV. 1021, 1021–22 (1977).

18. As seen in the diagram, the SAE only regards automation at "level 3" or above as "automated." The definition of AI used in this Article is broader.

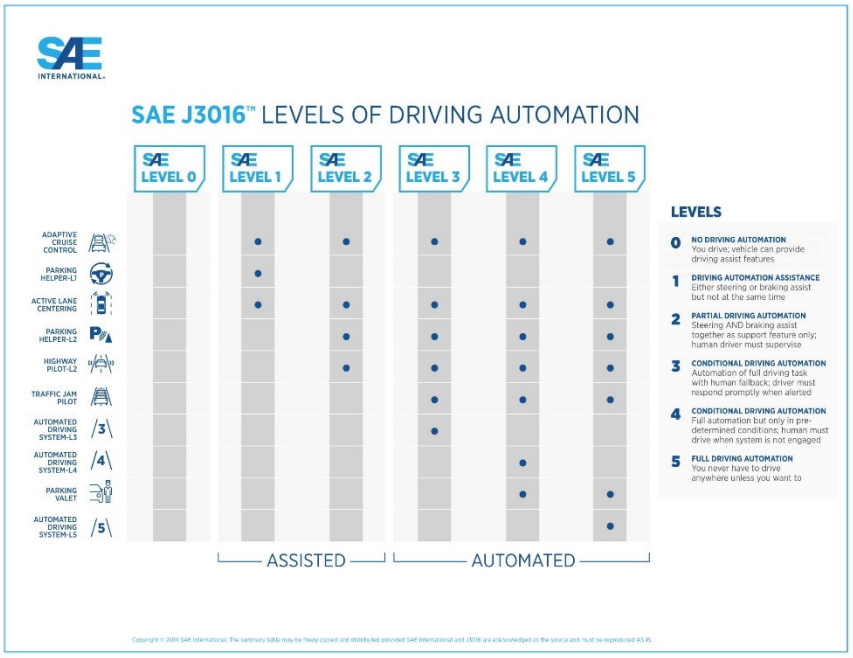


Figure 2: SAE automated driving level definitions (SAE J3016).

Within the legal field, there is at this time no similar agreed definition of the gradations of automation. However, it is not difficult to adapt the concept to legal work. Take the drafting of a legal document, for example. At the lower levels of automation is a software tool that provides case law research, such as the more modern iterations of WestLaw, LexisNexis, and Casetext. Such AI can only provide *evaluative* assistance. It is not ready to make any *decision* without human input. The output is in the form of *recommendations* for human decision-makers to use in their own legal drafting work. At the higher end of automation, there may be a software tool that actually performs the legal drafting and leaves the humans with a "minimum" number of edits and approvals. The AI moves beyond evaluative assistance and closer towards an automated maker of *decisions*.

The level of automation is, in some sense, about the level of deference humans are willing to entrust to AI.[19] After all, a tool that provides recommendations can be transformed instantly into a tool that makes decisions by blind acceptance of all the recommendations. Thus, allowing

19. The notion of deference has been discussed in legal literature in the context of understanding the role of specialized expertise in the legal process. *See generally* Ronald J. Allen & Joseph S. Miller, *The Common Law Theory of Experts: Deference or Education?*, 87 NW. U. L. REV. 1131 (1993).

AI to make decisions instead of recommendations would mean deferring to the value system either built into the AI or into AI use processes.[20]

It is difficult to articulate with any precision how much decision-making responsibility should be given to AI.[21] At the very least, the amount of delegated responsibility should take into consideration the level of automation of the AI, the difficulty of the problem presented to the AI, and the importance of the problem.

For example, this Article discusses the example of civil discovery, where AI output is not all individually reviewed by humans. With respect to the production of millions of responsive documents from one litigant to another, the probability that any single document in a document collection will decide the entire case is slight. The technical difficulty of determining whether an individual document is responsive or not is not particularly high. To that extent, it may be acceptable to cede to AI the decision-making authority over whether to produce individual documents, even if the AI is not particularly sophisticated.

In contrast, forensic science may have strong and close implications on the welfare, life, and liberty of the accused. The technical difficulty of the evaluation could be considerable. It may be necessary to confine AI, however trustworthy, to do no more than provide an evaluative function.

This Article does not aim to resolve where to strike the line between using AI for evaluation and for decision. Instead, it acknowledges the divide merely as a starting point, and proceeds to describe NNA as a concept for assessing AI recommendations created in the evaluative step of decision-making and also for reviewing AI automated decisions.

---

20. *See* POOLE & MACKWORTH, *supra* note 6, at 439.

21. Humans can certainly entrust decision-making authority to AI even when the AI was not designed for such a purpose. This problem is well-publicized in the context of driving. In its accident report concerning a fatal crash of a Tesla car under autopilot into a semitrailer, the National Transportation Safety Board ("NTSB") stated:

> To summarize the discussion of the . . . driver's actions before the crash, he used the Autopilot system on roadways for which it was not designed . . . and had extended periods of hands-off driving and other indications of lack of engagement/awareness before the crash . . . . Both driver behaviors strongly indicate that, although the Tesla owner's manual provided information and warnings on these subjects, the driver either did not know of or did not heed the guidance in the manual. Therefore, the NTSB concludes that the Tesla driver's pattern of use of the Autopilot system indicates an overreliance on the automation and a lack of understanding of system limitations. Also, the NTSB concludes that the Tesla driver was not attentive to the driving task . . . .

NAT'L TRANSP. SAFETY BD., ACCIDENT REPORT NTSB/HAR-17/02 PB2017-102600, COLLISION BETWEEN A CAR OPERATING WITH AUTOMATED VEHICLE CONTROL SYSTEMS AND A TRACTOR-SEMITRAILER TRUCK NEAR WILLISTON, FLORIDA MAY 7, 2016, 35–36 (2017).

C. Nature of AI Output

AI may seem like magic for readers unfamiliar with the concept. However, its process is fundamentally based on precedents[22] and obeys this fundamental process:

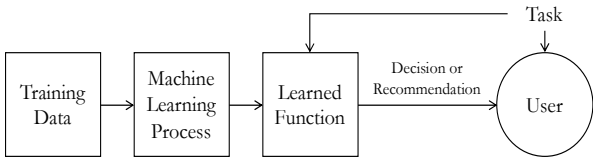
1) collect assessments or outcomes for known cases, ideally *ground truth* responses; and
2) synthesize a response to the new situation based on the known responses to the known cases.

In this process, a *decision boundary*,[23] which separates one class from another, is formed. Take, for example, an AI designed for identifying cats in pictures. The system will be supplied with many pictures, all *tagged* as either having cats or not having cats, which will serve as the ground truth and will be used as its *training set*. When given a new picture and called upon to decide whether there is a cat in the picture, the AI would decide whether the picture has a cat based on the training set.[24] This general concept of using AI to make recommendations or decisions driven by precedents has already been widely implemented in the legal realm.[25]

Lawyers and judges are well familiar with this process flow. When confronted with a legal problem, they search through the precedent for

---

22. The following diagram summarizes the process flow of a typical, modern AI machine learning process:



See U.S. GOV'T ACCOUNTABILITY OFFICE, *supra* note 4, at 19 (with adaptations).

23. "Decision boundary" is an accepted technical term of art in the field of data science. It will therefore be used in this article whether the context is evaluation or decision.

24. We emphasize that this is a strongly simplified description and that practical applications face many challenges and anomalies. For an example, see Anh Nguyen et al., *Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images*, *in* COMPUTER VISION AND PATTERN RECOGNITION (CVPR '15), IEEE 427, 432–34 (2015).

25. In the field of forensic science, AI is used to generate candidate suggestions based on input images. *See* Che-Yen Wen & Jing-Yue Yao, *Pistol image retrieval by shape representation*, FORENSIC SCI. INT'L, Dec. 2005, at 35, 35 (discussing the use of input images of pistols); Daniel Moreira et al., *Pornography classification: The hidden clues in video space-time*, FORENSIC SCI. INT'L, Nov. 2016, at 46, 47 (discussing the use of input images to classify videos as pornography). Law enforcement agencies also use AI to, for example, identify potential investment advisor misconduct in new regulatory filings based on examples of previously examined filings. *See* Scott W. Bauguess, Acting Dir., Div. of Econ. & Risk Analysis, Sec. & Exch. Comm'n, Keynote Address at OpRisk North America 2017 (June 21, 2017) (transcript *available at* https://bit.ly/34tHI14).

similar cases.[26] They then compare the legal problem at hand with the precedent cases to look for similarities or differences in the underlying facts. They use the outcomes of the precedent cases to guide them to a recommendation or decision. Although they do not use formalized terms such as *decision boundary*, they do refer to their own actions as "line-drawing." Our discussion here is therefore directed to adapting a mode of reasoning that has been at the core of the legal system for centuries towards analysis of AI output.

Humans exhibit a great deal of variation in how they extend precedent to answer new cases. Lawyers and judges, for example, may look for patterns, identifying rules and standards. They may apply doctrines such as the maxims of equity. Where the new case matches or almost exactly maps onto a particular case in the precedent, they may decide based on that one case alone.

AI recommendation or decision is, in this way, not all that different from human decision. As with humans, there is great variation in how each AI methodology "extends" the training set to deal with situations that do not fall directly within the training set.[27] Techniques such as neural networks look for "features" within the training set that can best account for the known correct responses and then look for these "features" in new cases. Watson, an IBM computer well-known for winning *Jeopardy!*, used an "expanded corpus" built from a merger of "informative nuggets" rather than pattern recognition.[28] Mathematics, physics, and other scientific knowledge can also be used to extend the training set to handle new situations. More advanced machines, like self-driving cars, may use a combination of methods. How to do this "better" is the work of research.

In short, the fundamental process of AI recommendation or decision-making is not magic and can be analogized to human behavior. Moreover, both recommendation and actual decision-making can be blended with the logic of decision analysis, specifically, decision theory.

III. NEAREST NEIGHBORS

A. Problems with Conventional Methods of Evaluating AI

26. *See* Edward H. Levi, *An Introduction to Legal Reasoning*, 15 U CHI. L. REV. 501, 501 (1948).

27. The training set can itself be extended by the technique of data augmentation or enrichment. *See, e.g.*, Zahraa S. Abdallah et al., *Data Preparation*, *in* ENCYCLOPEDIA OF MACHINE LEARNING AND DATA MINING 318, 321 (Claude Sammut & Geoffrey I. Webb eds., 2017).

28. *See* David Ferrucci et al., *Building Watson: An Overview of the DeepQA Project*, AI MAG., Fall 2010, at 59, 69 (2010).

Output

In real-world applications, recommendation or decision-making based on precedents is rarely ever perfect. This is true either because precedents are wrongly applied or because existing precedents are insufficient to cover a new, unexpected problem. By nature, there must be some entity with the authority to make the ultimate call.

Nonetheless, it is a practical necessity that, at least some of the time, recommendations or decisions be made based on precedents. After all, it would not be efficient for the final authority to make all the calls itself. No one in the legal field would think it appropriate for the highest court to decide all issues in all cases within the jurisdiction. Instead, everyone accepts the practicality of having lower courts apply precedents. Likewise, the idea behind AI assistance in or automation of decision-making is to reduce the amount of human work needed to make decisions.

But the question of how recommendations or decisions should be evaluated is a difficult problem. To that end, it is necessary to distinguish between aggregate measures of overall performance and measures of the quality of an individual recommendation or decision. This problem is well-recognized in the field of forensic science:

> [T]he overall performance of an expert—as monitored across repeated exercises (tests) under controlled conditions—is not a direct measure of the 'goodness' of a particular decision (to be) made in a given case at hand. Stated otherwise, general performance indicators of an expert, including any technical system or device that is used during the analysis process, do not answer the question of how 'good'—or accurate (when an underlying truth state can be considered)—an individual decision is.[29]

This distinction exists in the evaluation of recommendations or decisions made by human experts as well as AI.

The importance of this distinction can be observed with a hypothetical. Let us imagine an appellate court reviewing a district court's decision. The appellate court has, on the one hand, the opinion that contains the reasoning of the district court. On the other hand, the appellate court has the reversal rates of that district court. Which should the appellate court consider in its review? Even if reversal rates were valid measures of performance, no one would tolerate appellate review by such

29. *See* Alex Biedermann et al., *A formal approach to qualifying and quantifying the 'goodness' of forensic identification decisions*, 17 L., PROBABILITY & RISK 295, 299 (2018); *see also* Faigman et al., *supra* note 2, at 420.

overall performance metrics without regard to the particularities of the decision.

The same logic is true whenever a legal decision-maker has occasion to evaluate AI output. When a court is presented with an RPI assessment of a particular convict's potential for recidivism, it would at best be helpful, but not entirely sufficient, to know the overall performance of the RPI. It would seem appropriate for the court to have some indication about how well the RPI would perform with regard to particularities of the convict. Similarly, when assessing the trustworthiness of a forensic expert's AI-assisted identification of a low-quality fingermark,[30] a court may wish to know about the performance of the process with comparable trace materials rather than information about the general performance.[31] And when a court is tasked to consider why a "smoking gun" document was not produced in an AI-driven document review process, the court may be interested to learn about why that specific document was not produced as opposed to the general measures of the document review process.

These particularities present a deep challenge. As of now, there are many well-established and accepted metrics of overall performance of decision-making. Table 1 summarizes the four basic metrics that are used for evaluations of performance against ground truths.[32]

		Actual condition	
		Positive	Negative
Asserted condition	Positive	True positive (TP)	False positive (FP)
	Negative	False negative (FN)	True negative (TN)

30. In this Article, a *fingerprint* is understood as a print—usually of good quality—taken from a person of interest under controlled conditions. *See* CHRISTOPHE CHAMPOD ET AL., FINGERPRINTS AND OTHER RIDGE SKIN IMPRESSIONS 317 (Max Houck ed., 2016). Examples include an inked print on a ten-print card or a scan of a finger's ridge patterns. A fingerprint is distinguished from a *fingermark*, which refers to an impression—of varying quality—left by ridge skin on a receptor surface by an unknown person. *See id.* A fingermark may be visible or latent and require an enhancement and/or lifting technique in order to be recorded and further processed. *See, e.g.*, Didier Meuwly, *Forensic Use of Fingerprints and Fingermarks*, *in* ENCYCLOPEDIA OF BIOMETRICS 723, 734 (Stan Z. Li & Anil K. Jain eds., 2015).

31. While high quality fingerprints of a given person exhibit only small variations so that current algorithms are able to perform reasonably well and can be run in fully automated mode, the processing of partial, blurry or smeared fingermarks requires more human intervention. *See* Davide Maltoni & Raffaele Cappelli, *Fingerprint Recognition*, *in* HANDBOOK OF BIOMETRICS 23, 31 (Anil K. Jain et al. eds., 2008).

32. *See, e.g.*, POOLE & MACKWORTH, *supra* note 6, at 279.

Table 1: Simple example of a confusion matrix for outcomes that are classified into the two categories of "positive" and "negative."

These four metrics form the basis of more elaborate metrics,[33] which exist in different fields under varying names,[34] such as sensitivity, specificity, positive predictive value, among others.[35]

But there is nothing similar in rigor and ease of use for assessing the quality of an individual AI output. The problem comes, in part, from the inability of many existing AI systems to explain themselves[36] and from the inadequacy of AI determination of the quality of its own assessments.[37] For example, the reader may recall Watson, an IBM computer, and its comprehensive victory over the best human *Jeopardy!* players in a 2011 exhibition match. At the "Final Jeopardy" round of the first night of the two days of competition, Watson and its human competition were given the following clue under the category of "U.S. Cities:" "Its largest airport was named for a World War II hero; its second largest, for a World War II battle." The correct answer was, "What is Chicago?" The airports named for the World War II hero and battle were, respectively, O'Hare and Midway. Watson at that point had a large, insurmountable lead over the human contestants. Still, it famously blundered, with "What is Toronto???," making reference to a city which is not even in the United States.

In response to the intense public interest, IBM provided this explanation for Watson's mistake:

> How could the machine have been so wrong? David Ferrucci, the manager of the Watson project . . ., explained . . . that several things probably confused Watson. First, the category names on *Jeopardy!* are tricky. The answers often do not exactly fit the category. Watson, in

33. *See* Kai Ming Ting, *Confusion Matrix*, *in* ENCYCLOPEDIA OF MACHINE LEARNING AND DATA MINING 260, 260 (Claude Sammut & Geoffrey I. Webb eds., 2d ed. 2017).

34. The divergence in names of the metrics originates from the different fields independently arriving at the same concepts to evaluate performance. The metric known as "precision" in the field of information retrieval, for example, is known as "positive predictive value" in the field of epidemiology.

35. *See, e.g.*, David H. Kaye, *The Validity of Tests: Caveat Omnes*, 27 JURIMETRICS J. 349, 351 (1987); Aakifa Aamir & Robert G. Hamilton, *Predictive Value Model for Laboratory Tests: Diagnostic Sensitivity, Diagnostic Specificity, Positive and Negative Predictive Value, Efficiency, Likelihood Ratio ([positive and negative]), Incidence and Prevalence*, *in* ENCYCLOPEDIA OF MEDICAL IMMUNOLOGY 581 (Ian R. Mackay et al. eds., 2014).

36. This is an active area of AI research.

37. As discussed further in Section IV.C, some applications in forensic science attempt to remedy this by focusing on assessing the probative value of a given comparison score using score distributions for competing versions of an event of interest.

> his training phase, learned that categories only weakly suggest the kind of answer that is expected, and, therefore, the machine downgrades their significance. The way the language was parsed provided an advantage for the humans and a disadvantage for Watson, as well. "What US city" wasn't in the question. If it had been, Watson would have given US cities much more weight as it searched for the answer. Adding to the confusion for Watson, there are cities named Toronto in the United States and the Toronto in Canada has an American League baseball team. It probably picked up those facts from the written material it has digested. Also, the machine didn't find much evidence to connect either city's airport to World War II. (Chicago was a very close second on Watson's list of possible answers.) So this is just one of those situations that's a snap for a reasonably knowledgeable human but a true brain teaser for the machine.[38]

The substantive explanation provided by IBM is, in all likelihood, an accurate description of the "thinking" process of Watson. It may make sense from a machine point of view. But it cannot be considered a satisfying explanation as to why "Toronto" was a good answer to the question. Associating "U.S. city" with "American League baseball" is simply not human logic. AI does not "reason" like a human and, until the technology of AI explanation greatly improves, any explanation of how it arrived at its answer will always seem cryptic and unusable.

Measures such as confidence may be useful but are also problematic. To begin with, "confidence" in statistics is a conflictual notion in its own right. Indeed, it has been argued that adding a probability assertion with another value thought to assert the "confidence" in the assigned probability would amount to "an infinite regress of beliefs about beliefs about beliefs."[39]

Also, asking AI about its confidence is akin to asking a human expert how confident he or she is in a particular decision. Confidence is relative, based on the knowledge of the expert, and may be unique to particular cases. The testimony of an expert who declares he is "completely confident" in the results is not necessarily easier to trust than that of an expert who declares that she is "quite confident." The human expert must accompany the confidence statement with explanation to be believed.[40]

---

38. *See* Steve Hamm, *Watson on Jeopardy! Day Two: The Confusion over an Airport Clue*, BUILDING A SMARTER PLANET (Feb. 15, 2011, 7:30 PM), https://bit.ly/2T0lakm.

39. DENNIS V. LINDLEY, UNDERSTANDING UNCERTAINTY 115 (2006).

40. In the case of forensic science, uncertainty in the assessment of comparative examination results is already considered an inherent part of the assessment procedure. A forensic expert is supposed to report the "best" assessment for the case at hand. It would then look redundant to add another level of assessment that attempts to qualify the

The same is true of AI. The confidence score is calculated according to the internal mechanism and knowledge of the AI. Standing alone in isolation, these measures have no meaning. Here is, for example, the explanation of Watson's confidence measures provided by IBM:

> [Watson] must rank the hypotheses and estimate confidence . . . . We adopted a machine-learning approach that requires running the system over a set of training questions with known answers and training a model based on the scores. . . . For more intelligent ranking . . . ranking and confidence estimation may be separated into two phases. In both phases sets of scores may be grouped according to their domain (for example type matching, passage scoring, and so on.) . . . . [C]ertain scores that may be crucial to identifying the correct answer for a factoid question may not be as useful on puzzle questions.[41]

Watson's confidence level for Toronto as an answer "was about 30%."[42] How is 30% to be evaluated against 20% or 40%, especially when the confidence is different for "different question classes"?

The confidence numbers, though an objective number calculated from a formula, are based on Watson's knowledge and are meaningful only within Watson's framework. Additionally, as with human experts, the confidence level of an AI needs to be accompanied with reasons to be useful to humans. Unfortunately, this simply returns us to the earlier discussion about the difficulty AI has in explaining itself.

B. Nearest Neighbors in the Training Set

Although it may not be able to provide explanations, what AI can provide in principle is the closest precedent in its training set, that is, the nearest neighbor precedent. Prior to an exploration of nearest neighbors, however, it is necessary to discuss the concept of distance itself. The subject matter of distance is extremely complicated, but here we will illustrate with simple examples.

1. Distance and Nearest Neighbor in Geometry

We will begin with a review of the concept of vectors, which is necessary to understand how AI evaluates distance. Most readers may recall the concept of the Euclidean vector, that is, graphically speaking, an arrow connecting two points. Figure 3 illustrates three vectors in the Cartesian coordinate system.

"goodness" of the expert's "best" assessment. Nonetheless, the forensic expert would be required to explain the particular assessment given and its inherent sources of uncertainty.

41. David Ferrucci et al., *supra* note 28, at 74.

42. Hamm, *supra* note 38.

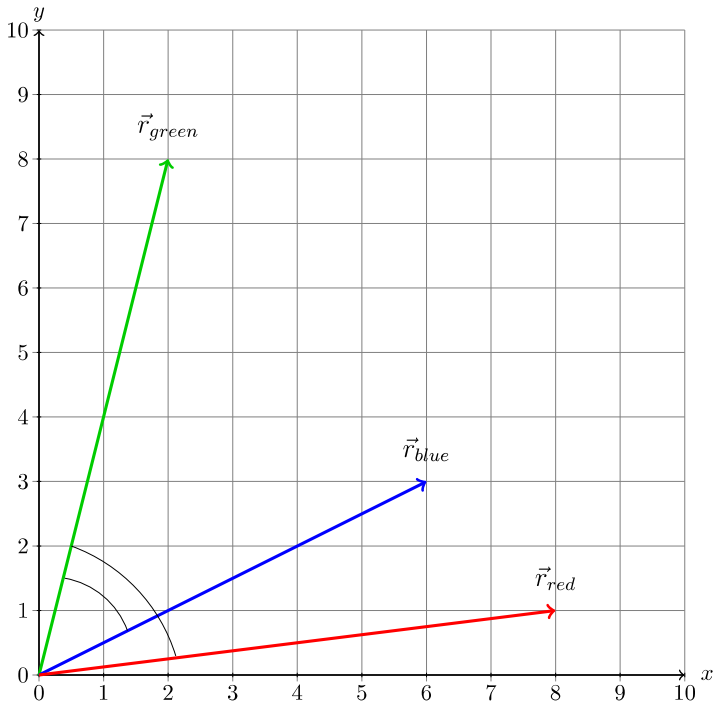


Figure 3: Three Euclidean vectors and the angles between them.

The three vectors can be expressed in Cartesian coordinates, [x, y], and their lengths, according to the most familiar *Euclidean* distance,[43] calculated.

43. This is, in technical parlance, the L_2 *norm*. The reader is encouraged to refer to a mathematical text for other Euclidean distances such as the L_1 *norm* and the L_∞ *norm*.

Vector	Cartesian Coordinates	Length[43]	Normalized Coordinates
	$[x, y]$	$l = \sqrt{x^2 + y^2}$	$[x/l, y/l]$
$\vec{r}_{red}$	[8, 1]	8.0623	[0.9923, 0.1240]
$\vec{r}_{blue}$	[6, 2]	6.3246	[0.9487, 0.3162]
$\vec{r}_{green}$	[2, 8]	8.2462	[0.2425, 0.9701]

It is sometimes useful to normalize the vectors to all have the same length of 1. That way all the vectors in some sense have the same importance. The way to do this is to divide the original vector coordinates by the length of the vector. We could then obtain the following distance between the vectors using the Euclidean distance, where Σ indicates summation.

$$\begin{aligned} d_{red-blue} &= \sqrt{\sum_i (red_i - blue_i)^2} \\ &= \sqrt{(0.9923 - 0.9487)^2 + (0.1240 - 0.3162)^2} \\ &= 0.1971 \end{aligned}$$

The distances between all three vectors are presented in the following table.

Euclidean Distance	$\vec{r}_{red}$	$\vec{r}_{blue}$	$\vec{r}_{green}$
$\vec{r}_{red}$	0	0.1971	1.1305
$\vec{r}_{blue}$	0.1971	0	0.9624
$\vec{r}_{green}$	1.1305	0.9624	0

Based on this definition of distance, out of blue and green, blue is the nearer neighbor to red. This makes sense geometrically from the diagram.

The Euclidean distance is one of many possible distance metrics, but there are also non-Euclidean definitions of distance, such as the *hamming distance*.[44]

2. Distance and Nearest Neighbor in Textual Data

The above mathematics should be no surprise to anyone who recalls their high school geometry class. The concept of the vectors and dot products, however, can be used to generate a mathematic measure of distance of *data* as well. Let us consider an example of textual data:

Lightly Row, Gently Row

Row, Row, Row Your Boat

Speed, Bonnie Boat

It is possible to convert the phrase data to *feature vectors*—compact summary representations of the phrases' main characteristics. One way to do this is to count the number of times each word appears:[45]

Phrase		Lightly Row, Gently Row	Row, Row, Row Your Boat	Speed, Bonnie Boat
Frequencies	Boat	0	1	1
	Bonnie	0	0	1
	Gently	1	0	0
	Lightly	1	0	0
	Row	2	3	0
	Speed	0	0	1
	Your	0	1	0
Length	$l = \sqrt{a^2 + b^2 + \ldots}$	2.4495	3.3166	1.7321

The vectors can then be normalized based on their lengths:

44. The reader is encouraged to refer to mathematical texts for such examples.

45. This sort of textual representation of sentences is known in natural language processing as a *bag-of-words* model. *See, e.g.*, RUSSEL & NORVIG, *supra* note 1, at 866.

Phrase		Lightly Row, Gently Row	Row, Row, Row Your Boat	Speed, Bonnie Boat
Normalized Frequencies	Boat	0	0.3015	0.5773
	Bonnie	0	0	0.5773
	Gently	0.4082	0	0
	Lightly	0.4082	0	0
	Row	0.8165	0.9045	0
	Speed	0	0	0.5773
	Your	0	0.3015	0

We can then obtain the distance between the vectors using the Euclidean distance. As an example, this is the distance between "lightly row, gently row" and "row, row, row your boat."

$$d_{LRGR-RRRYB} = \sqrt{\sum_i (LRGR_i - RRRYB_i)^2}$$
$$= \sqrt{\begin{matrix}(0-0.3015)^2 + (0-0)^2 + (0.4082-0)^2 + (0.4082-0)^2 \\ +(0.8165-0.9045)^2 + (0-0)^2 + (0-0.3015)^2\end{matrix}}$$
$$= 0.7231$$

The Euclidean distances between all three phrases are presented in the following table.

Euclidean Distance	Lightly Row, Gently Row	Row, Row, Row Your Boat	Speed, Bonnie Boat
Lightly Row, Gently Row	0	0.7231	1.4141
Row, Row, Row Your Boat	0.7231	0	1.2852
Speed, Bonnie Boat	1.4141	1.2852	0

Based on this definition of distance, out of "row, row, row your boat" and "speed, bonnie boat," "row, row, row your boat" is the nearer neighbor to

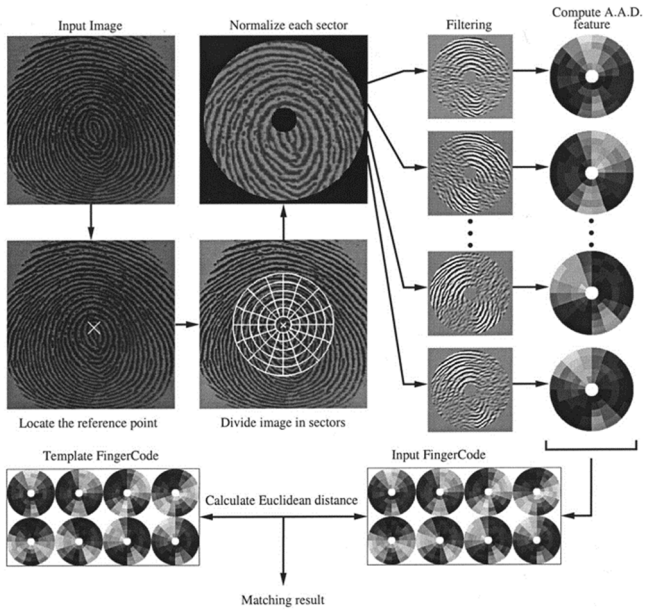
"lightly row, gently row." In addition, out of "lightly row, gently row" and "row, row, row your boat," "row, row, row your boat" is the nearer neighbor to "speed, bonnie boat."

The mathematical results confirm our intuitive knowledge. "Row, row, row your boat," while sharing the word "boat" with "speed, bonnie boat," is more similar to "lightly row, gently row" than it is to "speed, bonnie boat" because the two phrases heavily feature "row." Likewise, "speed, bonnie boat" is more similar to "row, row, row your boat" than to "lightly row, gently row" because the two sentences share the word "boat" while "speed, bonnie boat" and "lightly row, gently row" have no word in common.

3. Distance and Nearest Neighbor in Other Data

The type of analysis seen above is used for other types of data as well. Consider, for example, the diagram shown in Figure 4, drawn from a paper on fingerprint comparison.[46]

Figure 4: Visual summary of Jain et al.'s procedure for analyzing local texture in a fingerprint and definition of a feature vector that can be



compared to a reference collection of FingerCodes.[47]

46. *See* Anil K. Jain et al., *Filterbank-Based Fingerprint Matching*, 9 IEEE TRANSACTIONS ON IMAGE PROCESSING 846, 848 (2000).

47. *See id.*

Fingerprints may seem very different from text, but the general idea of extracting[48] and comparing features that underlie so-called fingerprint feature extraction and comparison algorithms is quite similar to what we discussed above.[49] In the approach illustrated in Figure 4, feature vectors, referred to as "FingerCodes," are extracted from input fingerprint images. To assess the similarity between a pair of FingerCodes, the Euclidean distance is calculated.

More generally, applications of distance calculations between feature vectors are found in many areas of forensic science.[50]

4. Recapitulation

Though it is difficult to generalize across all applications of AI, we will provide some common observations before we proceed to discuss how to evaluate the results of AI in the legal context. As discussed before, AI synthesizes a response to a new situation based on the correct or asserted responses to known cases. Some comparison of the new situation to the known situation is necessary. Typically, this comparison is conducted mathematically using feature vectors. How to generate a feature vector depends greatly on the application. For example, generating feature vectors for fingerprint comparison in a forensic analysis will be very different from that for textual comparison in civil discovery. Also, in many ways, these methods will be proprietary.

But once the feature vectors are generated, the methods of mathematically computing distance are generally quite similar. There may be many choices of distance definitions, as described above, and the choice of distance metric may be governed by the application. What is noteworthy is that there will usually be some sort of distance calculation. This is common to many AI methodologies. Accordingly, it is not outside the realm of technical possibility to demand the operators of AI in legal

48. Of course, the actual generation of the feature vectors in this application is very different from the generation of feature vectors for text discussed in Section 2. The various algorithms that have been described in the literature rely on different approaches to the comparison task. Some approaches focus on extracting configurations of minutiae such as features of ridges in fingerprints, while others do not use minutiae, relying instead on texture information for example.

49. However, conducting fingerprint comparisons with machines is considered a very difficult problem because images taken from a particular finger do exhibit intra-source variations. *See* DAVIDE MALTONI ET AL., HANDBOOK OF FINGERPRINT RECOGNITION 167 (2d ed. 2009).

50. For an illustration, see generally Charles E.H. Berger, *Objective Ink Color Comparison Through Image Processing and Machine Learning*, 53 SCI. & JUST. 55 (2013) (assessing and evaluating the similarity of ink colors).

decision-making to provide the nearest neighbor within the precedent cases.

C. Using NNA to Evaluate AI Output

Having examined how nearest neighbors can, at least in principle, be obtained from AI data, we now discuss NNA—the use of nearest neighbors to evaluate AI recommendations or decisions.

At the outset, the use of nearest neighbors as an analytical concept is not unfamiliar to judges and lawyers. Lawyers regularly cite "on point" cases in support of their position. At the same time, the duty of candor to the tribunal requires lawyers to "disclose to the tribunal legal authority in the controlling jurisdiction known to the lawyer to be *directly adverse* to the position of the client."[51] Lawyers may not simply ignore such adverse but "on point" cases; instead, "[they] may challenge the soundness of the other decision, attempt to distinguish it from the case at bar, or present other reasons why the court should not follow or even be influenced by it."[52] The use of nearest neighbors as a way for judges and legal practitioners to evaluate AI output is not revolutionary in terms of what they already do.

1. Identification or Individualization

The first question in NNA of a particular AI output is whether the AI output is in the form of an identification or an individualization. Individualization is the assignment of an object to a category that consists of a single unit.[53] In the process of individualization, an object is compared to the precedents that have already been categorized.[54] It is then placed in a category based on similarities to the precedents determined to be in that category and differences from precedents determined to be in other

51. MODEL RULES OF PROF'L CONDUCT r. 3.3(a)(2) (AM. BAR ASS'N 1983) (emphasis added).

52. ABA Comm. on Ethics & Prof'l Responsibility, Informal Op. 84-1505 (1984).

53. *See* John I. Thornton & Joseph L. Peterson, *The General Assumptions and Rationale of Forensic Identification*, *in* 4 MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY 1, 11 (David L. Faigman et al. eds., 2006-2007 ed. 2006). This Article assumes that individualization is indeed possible, though as a decision and not as a reporting format for forensic examiners. *See also supra* notes 14 and 46.

54. Usually, the precedent cases are categorized by humans or are categorized by AI under human supervision. We leave aside here the more complicated case in which precedents are categorized by AI without any human review at all.

categories.[55] Identification is the placement of an object in a category consisting of like units.[56]

Both individualization and identification are categorizations and can be conveniently illustrated through forensic science applications. Take, for example, the statement that a bullet was a .45 ACP. That statement is an identification because .45 ACP ammunition is produced in a great variety of forms and compositions. Also, there are many different guns that can fire this type of ammunition. However, the statement that a specific gun fired the bullet in question is, strictly speaking, an individualization.[57]

Throughout this Article, we emphasize this distinction between identification and individualization to ensure the clarity of concepts. We do note that the two concepts are commonly confused. For example, an individualization is often colloquially called an identification, as in the term "fingerprint identification."

Underlying identification is the assumption of likeness, meaning that objects can be categorized together based on the existence of a common set of properties.[58] In contrast, underlying individualization is the assumption of discernible and ascertainable uniqueness.[59] Specifically, the object defining the single unit has a distinctive set of properties that no other reference object within the same general category can share.[60]

---

55. This process does not necessarily require one-to-one comparisons with the various category representatives. It may be sufficient to determine the features of the object under study and then assign that object to a category if it satisfies the set of features which *define* that category. For example, forensic document examiners may determine the type of toner used by a black and white electrophotographic printing system by examining the physical properties of the toner present on a questioned document without a direct comparison to a reference item. *See* Alex Biedermann et al., *Analysis and Evaluation of Magnetism of Black Toners on Documents Printed by Electrophotographic Systems*, 267 FORENSIC SCI. INT'L, Oct. 2016, at 157.

56. *See* Paul L. Kirk, *The Ontogeny of Criminalistics*, 54 J. CRIM. L., CRIMINOLOGY, & POLICE SCI. 235, 236 (1963); Thornton & Peterson, *supra* note 53, at 11.

57. Identification can be an end goal in itself. However, identification can be seen as an intermediate step towards individualization. *See* Kirk, *supra* note 56, at 236. Alternatively, it can be seen as a limiting case of identification. *See* David H. Kaye, *Identification, individualization and uniqueness: What's the difference?*, LAW, PROBABILITY & RISK, July 7, 2009, at 86 [hereinafter Kaye, *Identification, individualization and uniqueness*]. Consider the determination that a bullet is fired from a gun belonging to a set G consisting of *n* guns. The determination remains an identification until sufficient defining features have been determined to narrow *n* to only 1.

58. An identification of any sophistication would also consider the differences of the object with objects already determined to be in another class. *See* Kaye, *Identification, individualization and uniqueness*, *supra* note 57, at 87 (explaining that "all identifications are classifications").

59. *See* C. Champod, *Overview and Meaning of Identification/Individualization*, *in* ENCYCLOPEDIA OF FORENSIC SCIENCES 303 (Max. M. Houck eds., 2013).

60. *See* Kirk, *supra* note 56, at 236 (noting that "[a] thing can be identical only with itself, never with any other object, since all objects in the universe are unique"). Note that

The nearest neighbor takes on a very different meaning depending on the purpose of the analysis. When one is identifying an object, presumably the nearest neighbor precedents to that object belong to the same category. If the AI correctly assigned the categorization to the object under study, then its categorization should correspond to that of the nearest neighbor precedents.[61] If the category of the nearest neighbor precedents differs from the category assigned by the AI to the object under study, then there may be cause for concern.

When one is individualizing an object, however, each of the precedents is its own category. Because individualization seeks to associate the object under study with only one of the precedents and thereby assign it to the category associated with that one precedent, all of the other precedents can be seen as competing categorizations. The precedents that are nearest neighbors to the object, therefore, are the ones determined by the AI to be the best competing categorizations. In other words, the first nearest neighbor is the best candidate for an individualization, the second nearest neighbor is the second best, and so on.[62]

2. Proximity

The previous section briefly touched on the idea that there may be divergence in the determination of proximity, or "degree of match," between AI and humans.[63] Just because AI deems a precedent to be a

---

this is an idealistic view. In practice, there are inevitable variations between objects assigned to a particular unit category. *See id*. For example, striation marks on bullets fired with the same gun will vary, to some extent, from one bullet fired to the next. This is due to the fact that the properties of the inner surface of a gun barrel evolve over time depending on factors such as the extent of use of the gun, the cleaning habits, and storage conditions. Even reference prints made by the same finger under controlled conditions vary slightly in aspects such as clarity and exact spatial arrangement of minutiae due to differences in pressure and angle of application. Much of the controversy about forensic science over the last decade concerned the foundations of individualization, and whether practitioners are actually capable of reliably achieving individualizations. *See* Kaye, *Identification, individualization and uniqueness*, *supra* note 57, at 88–90 (explaining by statistical example that an expert is "very likely" to individualize a bullet to a gun, but may not do so with 100% accuracy).

61. This assumes that the system has a rich enough knowledge base regarding the scope of categories and category representatives. For a thorough discussion on nearest neighbor precedents, see *infra* Section III.C.2. (discussing proximity).

62. This is an abstract idea of individualization, however. As we will explain later through examples, there are multiple obstacles, intrinsic to the objects under study, that adversely affect the performance of such procedures. In particular, ultimate "calls" made by human operators regarding individualization may be made with candidates other than the first nearest neighbor. *See infra* Section III.C.2. (discussing proximity).

63. For that matter, different AI programs operating from different definitions of distance will differ from each other in their determinations of proximity.

nearest neighbor to the particular object at hand does not mean that the precedent is actually close to the object in human eyes. For example, it is possible that the AI does not have a precedent case that corresponds to the current object or, in other words, the current object is a matter of "first impression" to the AI. Also, AI is using its own metrics to judge distance, which may not directly align with human impressions of distance. Phrased differently, AI may "see" something that distinguishes the nearest neighbor precedent from the present case, even if the distinguishing factors are not obvious to humans.

In the context of identification,[64] there is also the possibility of divergence in the category associated with the nearest neighbor precedent and the AI-assessed category for the current object. In general, one would expect that the AI would place the object into the same category associated with the nearest neighbor. But that may not always happen. Notwithstanding proximity of a current object to a nearest neighbor precedent, the other precedent cases may direct the AI to place the current object into a category different from that of the nearest neighbor.

Human perception of distance is *generally* taken as the gold standard, given the current state of AI development. Take, for example, the problem of glasses for facial recognition software. Few humans would believe that the faces shown in Figure 5 belong to different people.

Figure 5: Example of image perturbation, presented by Sharif et al.,[65] based on an image of actor Owen Wilson.

However, some AI facial recognition algorithms fail to relate the two, finding a lack of proximity that does not exist for the human observer.

Nevertheless, AI assessment should be approached with due respect. AI can "see" patterns where humans cannot and thereby outperform

64. This problem does not occur in the context of individualization because each precedent has its own individual category. *See* Kirk, *supra* note 56, at 236. The AI assessment for the current object is the category associated with the first nearest neighbor precedent.

65. *See* Mahmood Sharif et al., *A General Framework for Adversarial Examples with Objectives*, ACM TRANSACTIONS ON PRIVACY AND SECURITY, June 2019, at 16, 16.

humans. For example, AI can identify indications of Alzheimer's before the symptoms are perceptible to humans.[66]

3. Proximity in Individualization Problems

Having introduced the distinction between identification and individualization and the concept of proximity, we are now ready to combine the concepts together. We will start with the problem of individualization. Individualization seeks to associate the object under study with only one of the precedents and thereby assign it to the category associated with that one precedent. At present, AI systems that support individualization processes will generally present a menu of recommendations, all of which are near neighbors of the object under study, for its human operators to pick. The precedent picked by the humans is the ultimate decision.

There are two NNAs that can be helpful to probe the decision. The first is to consider the nearest neighbors of the object under study. If one accepts the top recommendation—the first nearest neighbor of the object under study—it would be extremely useful to consider the other nearest neighbors, as illustrated in Figure 6(a). If there is a second nearest neighbor that is indistinguishable to human eyes, or nearly so, this should prompt caution against making an individualization decision with the first nearest neighbor. At the least, the human decision-maker should carefully consider arguments as to why the second nearest neighbor has not been retained. In contrast, if the second nearest neighbor looks completely unlike the first nearest neighbor, it may also be a cause for concern that the AI does not have sufficient precedent cases and categories. If one does *not* accept the top candidate, however, there should be articulable reasons for skipping over that first nearest neighbor. The further one goes down the list of nearest neighbors in picking a match, the more one should be skeptical.

The second NNA, illustrated in Figure 6(b), also considers the nearest neighbors of the precedent picked by the human within the precedent set. Unlike the previous NNA, which searches for the nearest neighbors of the current object, this analysis searches for the nearest neighbor of the selected precedent and is concerned with the nearest neighbors of a nearest neighbor.

66. *See* Sulantha Mathotaarachchi et al., *Identifying incipient dementia individuals using machine learning and amyloid imaging*, 59 NEUROBIOLOGY AGING, Nov. 2017, at 82–84.

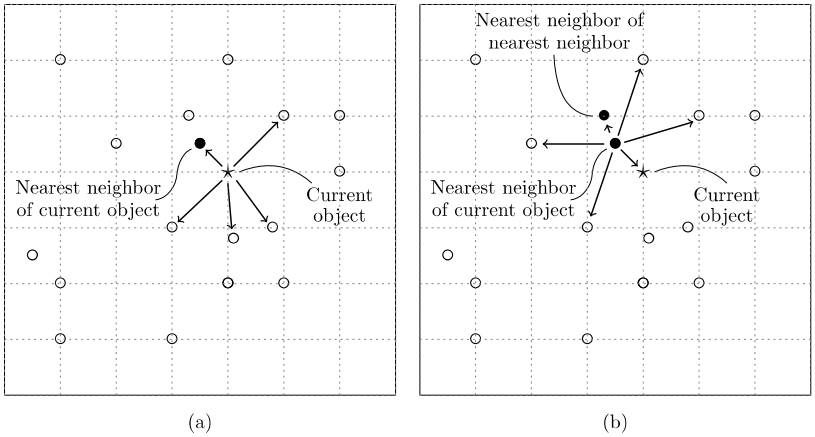


Figure 6: (a) Illustration of nearest neighbors of a current object, with the plain dot representing the nearest neighbor, that is, the selected precedent. (b) Illustration of the nearest neighbors of a selected precedent.

The nearest neighbors of the selected precedent can then be compared to the object under study in order to see whether these nearest neighbors look proximate to the object under study. We can recognize two extremes:

(1) when the nearest neighbors of the selected precedent do not appear to be proximate to the object under study; and

(2) when the nearest neighbors of the selected precedent appear as proximate to the object under study as the selected precedent itself.

In situation (1), the selected precedent can be considered an isolate. On the one hand, it may indicate that there is indeed something unique to that precedent that does not conflict with the individualization decision. On the other hand, one should keep in mind the size of the precedent set within which the search is conducted, which may be too limited.

In situation (2), the selected precedent can be considered to be in a "crowded" field. It may indicate reason to be skeptical of making an individualization decision. Most scenarios would fall somewhere in between the two extremes.

4. Proximity in Identification Problems

Let us now examine how NNA can be used in identification problems. We can recognize three extreme situations:

(1) when the nearest neighbor precedent is proximate to the object under study and when the category of the nearest neighbor precedent corresponds to the AI assessment for the object under study;

(2) when the nearest neighbor precedent is proximate to the object under study and when the category of the nearest neighbor precedent does *not* correspond to the AI assessment for the object under study; and

(3) when the nearest neighbor is not proximate to the object under study and has the same or a different classification label.

Figure 7 provides schematics of these three situations in the context of a binary decision[67] involving a sharp decision boundary.[68]

67. That is, a system with only two categories. In reality, there can be multiple categories.

68. In reality, the decision boundaries can be gradual and not sharp. For example, one need not designate between "fast" or "slow"; one can have "fast," "not very fast," and "slow."

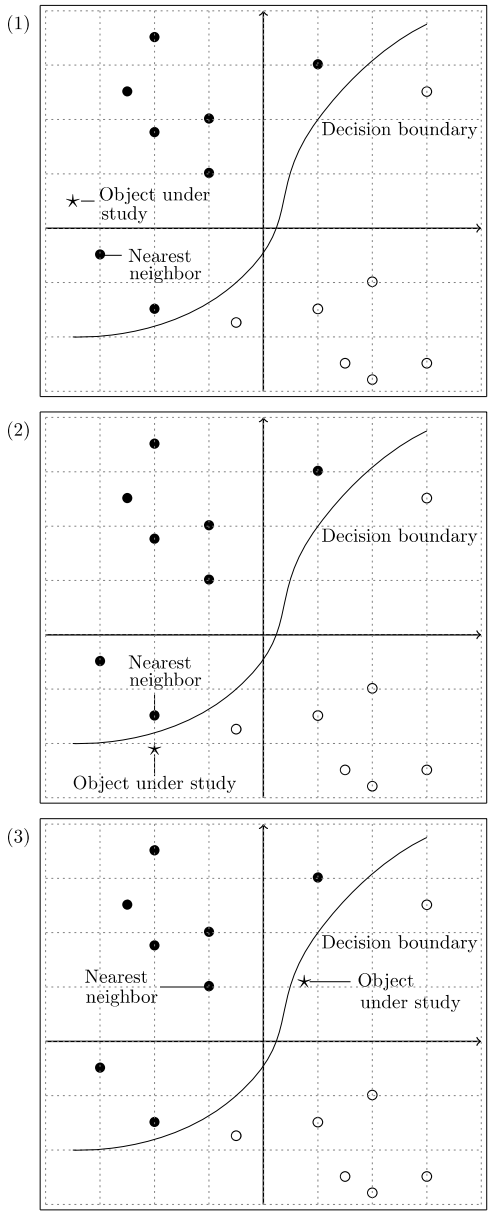


Figure 7: (1) Illustration of an AI assessment for a current object that corresponds to the tagged outcome of the nearest neighbor. (2) Illustration of an AI assessment where the nearest neighbor tag does *not* correspond to the current object. (3) Situation in which the nearest neighbor is "far" away. Note that the tag of the nearest neighbor in the latter case may correspond to or be different from that of the object under study.

Situation (1) represents the easiest case. If the AI output for a proximate current object corresponds to the category of the nearest neighbor precedent, then the AI output may, in this instance, be regarded as trustworthy as the category of the nearest neighbor precedent. The AI output may then be adopted or followed based on that basis.

Let us now consider situation (2). The fact that the category for the proximate nearest neighbor precedent is *different* from the AI output for the object under study indicates that the current case is a borderline case. There is also the possibility that the AI may not be drawing the decision boundary well. Consequently, there is a need to question the AI output or to bolster the training of the AI to refine the drawing of the decision boundary. It is also possible that the classification scheme is not granular enough. For example, a new category may be needed for the object under study.

Finally, we consider situation (3), where the distant nearest neighbor category may be different from one of the objects under study. This is a sign that the current case is a matter of "first impression" to the AI. The question then becomes one of whether to trust the AI. The AI output for the object under study may not be wrong; however, the AI output should be treated with skepticism.

It should be noted that situations (1), (2), and (3) are extreme situations. It is easy to imagine intermediate situations where, for example, the nearest neighbor is not proximate but is not too far from the current case. The treatment of such intermediate situations would require appropriate adaptations of the treatment of the extreme situations.

IV. POTENTIAL OF NNA IN LEGAL DECISION-MAKING INVOLVING AI

A. Civil Discovery

The discussions in this section will start with the use of AI in civil discovery because it is readily accessible and relatively uncontroversial. Under U.S. rules of civil procedure, a party may obtain documents "relevant to [his or her] claim or defense" by propounding document requests.[69] The responding party is then generally required to produce non-privileged documents responsive to the requests.[70] Traditionally, human attorneys representing the responding party review the document sets to determine what documents to produce. In large cases, this human review of documents can be extremely time-consuming and costly.

69. FED. R. CIV. P. 26(b)(1).
70. *Id.*

AI technologies, falling under the more general umbrella of technology assisted review ("TAR"), are now being used to assist in this process.[71] The idea of TAR is not to remove humans from reviewing documents, but rather to reduce the burden of human review. The process of training the AI follows the standard template of decision by precedence, as seen in the following description:

> For the software to begin classifying documents as to relevance, documents that are representative of relevant content must be identified and submitted to the computer [by the responding party]. For many supervised machine learning methods, documents that are representative of nonrelevant content must also be identified and submitted [by the responding party]. Once a set of relevant and nonrelevant examples have been submitted, the software analyzes their features and builds a predictive model, a classification system that categorizes or ranks documents in the TAR set.[72]

The precise reduction of human effort with the use of TAR depends on the particular review. As an unwritten rule of thumb, the typical TAR process can reduce the number of documents to be reviewed by humans by a third.

Although there are many TAR processes which differ from each other, it should be noted that they are close to automated decision-making. While human reviewers will review some of the documents reviewed by TAR to check for the validity of the processes,[73] they will not review every single AI output for each of the documents. Instead, they will accept the bulk majority of the responsiveness determinations made by AI.

At the time of writing, no U.S. court has ever found any specific TAR review process to be invalid.[74] That is not surprising, given that the traditional metrics of recall and precision are commonly accepted metrics for such a determination.[75] However, to the best of our knowledge, no court has encountered a case where a TAR process is called into question for failing to produce a specific responsive document or documents. Likewise, no court has encountered a case of malicious TAR usage, where the human reviewers taught the AI to avoid a specific category of

71. For more about the details of TAR, see TIMOTHY T. LAU & EMERY G. LEE III, TECHNOLOGY-ASSISTED REVIEW FOR DISCOVERY REQUESTS 2–5 (2017). *See also* BOLCH JUD. INST., TECHNOLOGY ASSISTED REVIEW (TAR) GUIDELINES 1–5 (2019).

72. BOLCH JUD. INST., *supra* note 71, at 10.

73. *See id*. at 25–26.

74. *See id*. at v. Note, however, that courts have also resisted compelling unwilling parties to use TAR. *See* Hyles v. New York City, No. 10 Civ. 3119, 2016 US Dist. LEXIS 100390, at *9 (S.D.N.Y. Aug. 1, 2016).

75. *See* BOLCH JUD. INST., *supra* note 71, at 5.

responsive documents or hid documents that the AI specifically identified as responsive.

Therefore, let us consider a situation where the responding party used a TAR process to produce documents and the propounding party found, through independent means, a document that should have been produced. The document was reviewed by the TAR process, identified as unresponsive, and was not produced by the responding party. The court is then asked by the propounding party to sanction the responding party for the failure.

In such a situation, it may be helpful to conduct an NNA. Specifically, the nearest neighbor precedent to the document in question within the training set can be identified and the categorization of the precedent examined. In view of the fact that TAR processes are identification tools to categorize documents as responsive or unresponsive to a specific document request, the principles outlined above in Section 4 can be applied:

(1) If the nearest neighbor precedent is similar to the document in question and the nearest neighbor is classified as unresponsive, then there may be reason to think that the training conducted by the responding party was deficient.

(2) If the nearest neighbor precedent is similar to the document in question and the nearest neighbor is classified as responsive, then there is reason to think that the document in question is a borderline case. In such an instance, the failure to produce the document may reflect an inherent limitation in the system's operability. Inaccurate outcomes are to be expected even in systems which generally perform to satisfaction.

(3) If the nearest neighbor precedent is not at all similar to the document in question, then the document in question is an outlier with regard to the training data. It may indicate that the training set was not representative of the overall document population. Alternatively, it may be that the document in question is an outlier within the entire document population altogether.

The NNA-based inspection can be extended by considering other nearest neighbor precedents beyond the first nearest neighbor precedent.

B. Risk Prediction

RPIs are actuarial tools used to assess whether individuals pose a criminal risk. Within the United States, RPIs have been used by probation

officers to determine how to supervise offenders and, more recently, by courts to help determine sentencing.[76] There are many similar tools which vary in complexity and sophistication. One RPI that has drawn attention in the media is COMPAS,[77] which provides "risks scales for general recidivism, violent recidivism, and pretrial misconduct."[78] The system is used in some state jurisdictions, such as Wisconsin[79] and Michigan.[80] In this discussion, COMPAS will be taken as representative of RPIs.

Many RPIs are not transparent.[81] COMPAS, for example, is proprietary; the inner workings of the software are not publicly known.[82] It is not entirely clear how the system is trained. The official documentation gives no information on the data used to develop the general recidivism scale. However, the official documentation does state that the violent recidivism scale is based on a "large sample of probation and presentence investigation . . . cases," but no further detail is provided.[83] As for the underlying algorithm, the documentation states that "[t]he methods used to develop both [recidivism] risk scales are described in various books on regression modeling and machine learning," giving citations but not entering into specifics.[84]

Although little is known about the internals, there is information about how the user interacts with COMPAS. The end user, usually a

---

76. *See* State v. Loomis, 881 N.W.2d 749, 752–53 (Wis. 2016).

77. *See, e.g.*, Ed Yong, *A Popular Algorithm Is No Better at Predicting Crimes Than Random People*, ATLANTIC (Jan. 17, 2018), https://bit.ly/2TcrHak; Jason Tashea, *Courts are Using AI to Sentence Criminals. That Must Stop Now*, WIRED (Apr. 17, 2017, 7:00 AM), https://bit.ly/2HY1Aiv. COMPAS is an acronym that stands for "Correctional Offender Management Profiling for Alternative Sanctions," but the RPI is rarely referred to with its full name. *See* MICH. DEP'T OF CORRS., FIELD OPERATIONS ADMINISTRATION, ADMINISTRATION AND USE OF COMPAS IN THE PRESENTENCE INVESTIGATION REPORT 2 (2017), *available at* https://bit.ly/2PpmcEg.

78. NORTHPOINTE, INC., PRACTITIONER'S GUIDE TO COMPAS CORE 26 (2015), *available at* https://bit.ly/2w5ZB90.

79. *See Loomis*, 881 N.W.2d at 752–53.

80. MICH. DEP'T OF CORRS., *supra* note 77, at 9.

81. On the notions of openness and transparency in relation to the use of algorithms in legal contexts, see Jason M. Chin et al., *Open Forensic Science*, J. LAW & BIOSCIENCES, July 2019, 284–85 and Megan T. Stevenson & Christopher Slobogin, *Algorithmic Risk Assessments and the Double-Edged Sword of Youth*, 96 WASH. U. L. REV. 681, 703–05 (2018). For arguments in support of public availability, see, for example, Brandon Garrett & John Monahan, *Assessing Risk: The Use of Risk Assessment in Sentencing*, 103 JUDICATURE, no. 2, Summer 2019, at 1, 47–48.

82. *See* Julia Dressel & Hany Farid, *The Accuracy, Fairness, and Limits of Predicting Recidivism*, 4 SCI. ADVANCES, no. 1, Jan. 2018, at 3.

83. NORTHPOINTE, INC., *supra* note 78, at 28.

84. *Id.* at 13; *see also* NORTHPOINTE RES. & DEV. DEP'T, COMPAS SCALES AND RISK MODELS VALIDITY AND RELIABILITY: A SUMMARY OF RESULTS FROM INTERNAL AND INDEPENDENT STUDIES 4 (2010) ("Standard logistic regression was used to predict recidivism with the full set of variables in each candidate set.").

probation officer, fills out a questionnaire of "137 questions that are either answered by defendants or pulled from criminal records."[85] The officer then enters the questionnaire into the system, and COMPAS provides a risk score.

It must be emphasized that, at this time, no RPI is known to completely automate decision-making. After all, RPIs are concerned with prospective risk management.[86] With respect to sentencing, for example, courts also need to consider the role of sentencing as retrospective punishment of crime, which is not a factor considered by RPIs.[87] Instead, the scores RPIs generate are evaluative suggestions that are used to guide the decision-making. The documentation for COMPAS, for example, explicitly acknowledges the possibility of disagreement with the scores by stating that, "[s]ometimes the COMPAS risk score for a particular person does not match the practitioner's expectations or clinical judgment regarding the level of risk posed by that person."[88] Indeed, it anticipates that disagreement will occur with some frequency:

> It is also important to note that we would expect staff to disagree with an actuarial risk assessment (e.g. COMPAS) in about 10% of the cases due to mitigating or aggravating circumstances which the computer is not sensitive to. In those cases[,] staff should be encouraged to use their professional judgment and override the computed risk as appropriate—documenting it in COMPAS with the Override Reason—for monitoring by supervisory staff.[89]

The literature is replete with discussions about whether RPIs are accurate or fair.[90] This Article does not delve into controversies over their statistical foundations, their operational suitability, nor their relative merit with respect to individual clinical evaluations.[91]

---

85. Julia Angwin et al., *Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks*, PROPUBLICA (May 23, 2016), https://bit.ly/32sIlaa.

86. *See* PAMELA M. CASEY ET AL., USING OFFENDER RISK AND NEEDS ASSESSMENT INFORMATION AT SENTENCING: GUIDANCE FOR COURTS FROM A NATIONAL WORKING GROUP 4–5 (2011).

87. *See id.* at 5–6.

88. NORTHPOINTE, INC., *supra* note 78, at 28.

89. *Id.* at 31.

90. *See* Anthony W. Flores et al., *False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks,"* FED. PROB., Sept. 2016, at 38, 44; Angwin et al., *supra* note 85; Dressel & Farid, *supra* note 82, at 3.

91. For a detailed critical discussion, see Peter B. Imrey & A. Philip Dawid, *A Commentary on Statistical Assessment of Violence Recidivism Risk*, 2 STAT. & PUB. POL'Y 25, 39–41 (2015).

What is important to note, as is repeated throughout this Article, is that a bad system can make a good recommendation and a good system can make a bad recommendation. None of the discussion in the literature helps evaluate the quality of a particular recommendation provided by an RPI. Regardless of the validity of RPIs, there needs to be some way to evaluate the quality of their recommendations at the individual level. Recurrent claims that these instruments can be "validated" before use and "revalidated" over time simply miss this basic point. What NNA can do here is provide an inherently empirical check on individual candidate conclusions and, hence, is neither a rival nor a substitute to standard measures of validity.

Consider the standard use case of a probation officer or a judge being given a score for the recidivism risk of a particular defendant. NNA can be used to evaluate the score as follows. The nearest neighbor of a current defendant is the defendant in the training data, who is most similar to the current defendant based on the distance metric inherent within the RPI. Whether the nearest neighbor did or did not recidivate is a known fact. The probation officer or the judge can, therefore, use this known fact to help assess the RPI score for the current defendant.

It should immediately be clear that RPIs are identification tools, used specifically to place the defendant within a particular risk category. For example, the documentation for COMPAS states that: "[r]isk assessment is about predicting group behavior (identifying groups of higher risk offenders) - it is not about prediction at the individual level. Your risk score is estimated based on known outcomes of groups of offenders who have similar characteristics."[92]

In accordance with the principles outlined in Section 4, we can consider the following situations:

(1) If the nearest neighbor is indeed similar to the current defendant and the known fact of the recidivism is actually not in conflict with the RPI score, then the nearest neighbor inspection does not raise a doubt about the RPI score.

(2) If the nearest neighbor is similar to the current defendant and the known fact of the recidivism disagrees with the RPI score, then there is reason to think that the current defendant is a borderline case. In such an instance, there may be reason to conduct further inquiry about the current defendant.

92. NORTHPOINTE, INC., *supra* note 78, at 31.

(3) If the nearest neighbor is not at all similar to the current defendant, then there may be reason to disregard the RPI score altogether, as the system would need to be regarded as having no knowledge about the recidivism status of individuals close enough to the defendant.

As is the case of civil discovery, the NNA can be extended by considering other nearest neighbor precedents beyond the first nearest neighbor precedent.

The use of nearest neighbor, to a degree, reduces the need for a definitive settlement of RPI validity. At present, decision-makers who are skeptical about RPIs give the scores they receive little weight, while those who believe in the validity of the tools trust the scores as useful information. This divergence in treatment of RPI scores cannot be considered an optimal outcome, as it results in uneven justice.

C. Forensic Comparison

NNA can also be useful in forensic science comparison.[93] Forensic science can be thought of as "the application of scientific or technical

93. The lay reader should not confuse machine-supported forensic comparison with identity verification. *See* MALTONI ET AL., *supra* note 49, at 3. For example, fingerprints may be used to identify a criminal from a crime scene or to verify a user to unlock a smart phone. Both applications may involve similar algorithmic architectures, but they serve rather distinct purposes. In the forensic comparison context, the machine output is typically a list of candidate sources for the input biometric data, ranked according to the degree of similarity between the questioned biometric data and the biometric features of each candidate. As explained in the main text, a human forensic examiner then "boils down" the list of candidates and—if possible—reports one candidate as the potential source of the questioned mark. As such, the machine is essentially a sorting device. In contrast, in the identity verification mode, a biometric system would typically provide a binary decision—a categorical association or exclusion of the input biometric features to a given reference entry. For an overview see, e.g., Christophe Champod & Damien Dessimoz, *Linkages Between Biometrics and Forensic Science*, *in* HANDBOOK OF BIOMETRICS 425 (Anil K. Jain et al. eds., 2008).

The two applications involve fundamentally different design demands. In classic forensic identification, a one-to-many search is conducted. For example, a fingermark retrieved from a crime scene is compared against a database of ten-print cards. In contrast, verification involves a focused one-to-one comparison. For example, a person's input fingerprint on a phone scanner is compared against the reference fingerprint previously taken by that exact same scanner stored in the phone. The problem of verification is fundamentally "easier," involving better input data and simpler binary decisions of acceptance or rejection of identity claims. This is not the case for procedures seeking to "identify" the source of a crime scene fingermark.

In addition, the social consequences of error are vastly different. A person wrongly rejected by a fingerprint biometric verification system may have to call a technician to unlock the system after human verification of his or her identity. But these inconveniences pale in comparison to the consequences of a bad forensic comparison. A person wrongly associated with a fingerprint at a crime scene will at least be the target of a criminal probe, and at worst may be put to trial or sent to prison.

practices to the recognition, collection, analysis, and interpretation of evidence for criminal and civil law or regulatory issues."[94] Forensic science examinations, especially in the so-called comparison disciplines,[95] generally involve comparing objects of unknown source from a crime scene with reference items.[96] The work is conducted to direct investigatory efforts at the known owners of the reference objects or to assist legal decision-makers in reaching ultimate determinations of guilt regarding owners.

Some forensic comparison focuses on identification, such as drug identification and taxonomy in wildlife. NNA for these types of forensic comparisons are conceptually similar to NNA for other identification tasks, which is covered by the previous sections about civil discovery and risk prediction. To avoid repetition, this section focuses on forensic comparisons that pertain to *individualization*.[97] In these comparisons, the objective is not to place an object within a category of like objects, but to associate it with a specific source.[98]

To do their work, forensic science examiners do not simply "match" objects. This is because there is virtually never a perfect congruence or correspondence between compared objects.[99] Instead, examiners focus on

94. PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH., EXEC. OFFICE OF THE PRESIDENT, FORENSIC SCIENCE IN CRIMINAL COURTS: ENSURING SCIENTIFIC VALIDITY OF FEATURE-COMPARISON METHODS 1 (2016).

95. Not all forensic science tasks involve comparisons. Examples that do not involve comparisons in the traditional sense include accident reconstruction, blood spatter analyses, and autopsies.

96. The nature of these entities may be biometric, such as ridge skin characteristics, DNA, ear, face, and iris, as well as physical traces, that is, marks and impressions left by items such as shoes, tires, firearms, and tools.

97. This Article also focuses on individualization because it is a task that currently attracts most criticisms. Individualization tasks are controversial mainly because they can provide strong evidence to associate defendants with case-relevant stains or marks and thereby can have severe consequences. *See, e.g.*, Simon A. Cole, Symposium, *From the Crime Scene to the Courtroom: A Discouraging Omen: A Critical Evaluation of the Approved Uniform Language for Testimony and Reports for the Forensic Latent Print Discipline*, 34 GA. ST. U. L. REV. 1103 (2018).

98. *See* Kirk, *supra* note 56, at 236; Thornton & Peterson, *supra* note 53, at 11.

99. This is true even of DNA analysis, commonly thought of as involving exact matches. During DNA analysis it may happen that, for various reasons, additional signals—indicative of genetic features—are recorded, while there is also a possibility that some features that should appear in a profile go undetected, leading to an incomplete or partial profile. The occurrence of such phenomena depends on, among other things, the quality and amount of the DNA available for analysis and the performance of the laboratory. *See, e.g.*, Tacha Hicks et al., *A Framework for Interpreting Evidence*, *in* FORENSIC DNA EVIDENCE INTERPRETATION 37, 70 (John S. Buckleton et al. eds., 2d ed. 2016). Nevertheless, there may be situations in which the observed features look "too good to be true" in the context of the case, which should give rise to a suspicion that the evidence has been fabricated or tampered with.

evaluation, which is based on measurement of the objects of interest and the observation of similarities and differences.[100] Examiners characterize the extent of correspondence which may vary from few similarities and many differences to many similarities and few differences. What they seek to do in evaluation is to assess the probative value of observed similarities and differences.[101] It is in this evaluative aspect of examination work that AI can play its role. Simply put, information about objects is fed to AI as input, and AI algorithms provide a measure of the distance between the compared objects. This distance is expressed, for example, in terms of a comparison score.

After evaluation comes *decision*. That is, a legal decision-maker—with the assistance of the report or testimony of examiners—*decides* whether the candidate object (rather than an unknown source) is *in fact* the source of the questioned item. After all, any statement in terms of observed similarities and differences between compared objects, generated by either human examiners or AI, is only a measurement of distance. Current forensic AI systems, even when augmented with additional work done by examiners, do not cover this decisional step in an autonomous way.[102]

For illustration, consider a forensic fingerprint comparison system that searches a partial fingermark collected on a crime scene against a database. 0 provides an example of such a search. As can be seen in Figure 8, such a system will provide a list of "candidates," ranked according to their degree of proximity to the questioned mark. We can call these candidates "nearest neighbors" in the same sense used throughout this Article. Clearly, this ranked list will contain a top-ranked candidate, the first nearest neighbor. But the AI does not "identify" persons as sources of fingermarks used as evidence in criminal proceedings.[103] Rather, human examiners are needed to exercise their judgement to determine which candidate from the AI-generated list, if any, to report as the potential source of the questioned mark. They may take into consideration their own personal inspection of the nature, quality, and quantity of corresponding

100. This is particularly important where the input information is difficult to process, such as pattern evidence like fingermarks and tool-marks. For an example, see the illustration provided in Section III.B.3.

101. Within the forensic area, probative value is elicited by assessing the extent to which the findings in the comparisons are more compatible with one proposition (*e.g.*, the compared items come from the same source) rather than an alternative proposition (*e.g.*, the compared items come from different sources).

102. But, note again, that there are applications that do make decisions autonomously, such as access verification problems. However, these are not core forensic science tasks considered in this section. *See also supra* note 93.

103. For a more detailed illustration, see 0.

features used by the fingerprint searching system. They may also consider any additional features that, due to complexity and level of detail, are not taken into account by the system.[104] Ultimately, the legal decision-maker, based on the work product of the examiner and other pieces of evidence, makes the ultimate *decision*.

These contemporary machine-based approaches are, therefore, not autonomous decision systems.[105] Instead, they are used as sorting systems that provide human examiners with potential candidates, each accompanied with a measure of distance with respect to the input data. Even then, the work of the human examiners is but one piece of evidence that informs the legal decision of a judge or jury.

It is unlikely in the near to medium term that AI systems capable of fully-automating the work of human examiners can be built. In a study conducted by the National Institute of Standards and Technology on the performance of experts and AI in facial recognition, the authors concluded:

> The results of the study point to tangible ways to maximize face identification accuracy by exploiting the strengths of humans and machines working collaboratively. First, to optimize the accuracy of face identification, the best approach is to combine human and machine expertise. Fusing the most accurate machine with individual forensic facial examiners produced decisions that were more accurate than those arrived at by any pair of human and/or machine judges. This human–machine combination yielded higher accuracy than the fusion of two individual forensic facial examiners. Computational theory indicates that fusing systems works best when their decision strategies differ. Therefore, the superiority of human–machine fusion over human–human fusion suggests that humans and machines have different strengths and weaknesses that can be exploited/mitigated by cross-fusion.[106]

What we may expect to see as AI continues to improve is the increasing "cross-fusion" of human and AI expertise. This will allow AI to boost, but

104. For example, the AI may focus on so-called level 2 features (*e.g.*, major ridge path features) while the human examiner may also consider level 3 features (*e.g.*, shape of pores and relative position of pores).

105. The fact that such systems are not intended to be used as autonomous decision systems does not mean that they cannot be used as autonomous decision systems. As explained in Section B, deference to systems can convert them from recommendation-makers to decision-makers. This can be considered a type of AI misuse.

106. P. Jonathon Phillips et al., *Face Recognition Accuracy of Forensic Examiners, Superrecognizers, and Face Recognition Algorithms*, 115 PROC. NAT'L ACAD. SCI., June 2018, at 6171, 6174 (internal citation omitted).

not supplant, the work of human examiners by serving as a second set of eyes.

Therefore, we will focus on AI as used in the evaluative step of forensic individualization. For the legal decision-maker, NNA is useful for scrutinizing the trustworthiness of the candidate source selected with the help of AI. In other words, before asking to what extent the observed similarities and differences between the trace of unknown source and the candidate source are probative,[107] it is relevant to inquire about why the particular candidate source rather than another member of the list of candidates has been retained. This is a rarely considered aspect in current fingerprint practice and court adjudication.

As before, we can consider a number of situations, in accordance with the principles outlined in Section 3. To simplify the discussion, we continue to use fingerprints as an example. First, the candidate fingerprint chosen by the examiner among the initial candidate list is a nearest neighbor to the trace. But it is not necessarily the case that the examiners would pick the first nearest neighbor as the candidate.[108] Therefore, it may be useful for the legal decision-maker to inquire about the existence of other nearest neighbors and their ranks compared to that of the selected candidate. The reasons they have been discarded may be illuminating. Similarly, the meaning of the absence of a suitable nearest neighbor needs to considered with respect to the number of data points available to the comparison system. For example, the larger a dataset, the more likely it is to find appropriate nearest neighbors.

If the number of nearest neighbors that the examiner passed over to identify the candidate is larger than the number of nearest neighbors passed over in a typical comparison, this should raise some suspicion as to the suitability of the selected candidate for consideration. Conversely, if the number of nearest neighbors that the examiner passed over to identify the candidate is smaller than the number of nearest neighbors passed over in the typical comparison, this would be something we would expect to see *if* the selected candidate is indeed the source of the questioned item.[109]

107. Ideally, we would want comparison scores to be typical for one proposition, and not the other, and vice-versa. In reality, a neat separation is rarely possible. Some comparison scores may be observed under both competing propositions.

108. See Appendix A for a real case example where the examiner faced several nearest neighbors.

109. It is important to keep in mind that these are qualitative considerations. The examiner will still need to assess the probative value based on the actual comparison score assigned to the selected candidate, that is, considering how the actual score compares with scores from same source comparisons versus different source comparisons.

Second, the nearest neighbors of the candidate fingerprint can be considered. As discussed in Section 3, there can be two extremes in the outcome. The first of these two extremes is that the nearest neighbors to the candidate fingerprint do not resemble the candidate fingerprint in the pertinent details. Encountering such a situation depends, of course, on the size of the pool of fingerprints available for search by the AI as well as the diversity or distinctiveness of their feature configurations. But if the system does not find fingerprints that resemble the candidate fingerprint in the pertinent details, then that is an empirical suggestion that those details are rare or distinctive within the search pool.

The second extreme is that the nearest neighbors of the candidate fingerprint look similar to the candidate fingerprint in terms of the pertinent details. In this situation, the AI output can be thought of as challenging the proposition that the candidate fingerprint and the fingerprint under study are from the same source. After all, the computer system is pulling up good competitors to the candidate fingerprint. As noted previously, it may be valuable in such a case to inquire about the reasons that led the examiner to choose the particular candidate source rather than one of the good competitors.

It must be noted that, even though the NNA process as outlined here can assist in the evaluative step of discriminating between competing propositions,[110] it is not sufficient to allow for categorical assertions about whether or not the trace came from the same source as that of the reference.[111] Making such an assertion to "identify" a particular person as the source of the trace is a *decision* which, as already stated, involves value judgement.[112] That is more than NNA can support.

Nonetheless, NNA could usefully complement current forensic examination protocols. In fact, the most widely known and practiced protocol, the so-called "Analysis, Comparison, Evaluation–Verification"

110. There are already a few available computerized systems that provide assistance in the kind of evaluation described in this Article. An example includes PiAnoS, short for "Picture Annotation System," an open-source software-package developed at School of Criminal Justice at the University of Lausanne. *See* CHAMPOD ET AL., *supra* note 30, at 46. For another example, see Henry J. Swofford et al., *A Method for the Statistical Interpretation of Friction Ridge Skin Impression Evidence: Method Development and Validation*, 287 FORENSIC SCI. INT'L 113, (2018).

111. That is true of the entire comparison process in general. Indeed, some forensic laboratories, including the U.S. Defense Forensic Science Center, have declined to state conclusions that "identify" a particular person as the source of a given fingermark. *See* DEF. FORENSIC SCI. CTR., DEP'T OF THE ARMY, INFORMATION PAPER NO. CIFS-FSL-LP, USE OF THE TERM "IDENTIFICATION" IN LATENT PRINT TECHNICAL REPORTS 1 (2015).

112. *See* Biedermann & Vuille, *supra* note 14, at 399.

("ACE-V"),[113] devotes little attention to how the candidate retained for the comparison process was selected. In particular, ACE-V does not inform about the existence of close competitors to the selected candidate, nor does it encourage examiners to look actively for competing candidates. The example mentioned in Appendix A illustrates that extended searches can lead to surprising turns in actual cases.

V. CONCLUSION

Law is about the life and liberty of individuals. As such, for legal applications, good decision-making in the aggregate is not enough. The legal system aims not only to deliver good performance in the aggregate but also quality outcomes at the granular level: each opinion, sentencing decision, and verdict. As AI systems are increasingly adopted to assist or automate legal decision-making, it is important to have ways to gauge the quality of individual output of AI systems.

AI systems capable of explaining individual recommendations or decisions do not yet exist and may not exist for a very long time. For now, this Article argues that NNA is helpful for probing AI output used in the process of generating candidate conclusions in the *evaluative* step, as part of their own decision-making, and also for reviewing AI automated *decisions*. This Article explains how the methodology is useful for both individualization and identification problems, providing use cases to show what it can do. Though among the primary aspirations of the legal process, the insight into the actual quality of individual decisions remains—in practice—essentially inaccessible. NNA is not a "magic wand" to overcome this intricacy. We encourage lawyers, decision-makers, and regulators to critically observe the nature and explainability of AI output in current and future developments. At this time, however, NNA can be used for a case-specific and practically feasible challenge to AI output used to support legal decision-making. This should be of interest to all participants of the legal process who consume AI output. In addition, as NNA is fundamentally based on the use of comparison to precedents, it should be familiar to and easily adopted by legal professionals who are well acquainted with the classic analogical reasoning of the common law system.

This Article ends with a reminder. NNA may be used to help make up one's mind about *what to think* about AI output, such as whether a

113. *See* PAUL LEE ET AL., NAT'L INST. OF STANDARDS & TECH., NISTIR 8215, FORENSIC LATENT FINGERPRINT PREPROCESSING ASSESSMENT 2 (2018); CHAMPOD ET AL., *supra* note 30, at 34.

convict is a high recidivist risk or whether a recovered fingermark comes from a particular person. What it cannot do, however, is provide direct guidance about *what humans should do* with the AI output. Ultimately, decision-making relies on value judgments on the part of humans. Decisions made by machines are the result of the way in which humans instruct the machines. Human decision-makers must take responsibility for the use of AI and cannot separate themselves from the consequences of particular decisions.

APPENDIX A. EXAMPLE OF A FINGERMARK COMPARISON

Figure 8 shows an example screenshot of a comparison of a poor-quality fingermark against a database of over one million reference fingerprints using a current forensic fingerprint comparison system. The comparison was done for research purposes at the School of Criminal Justice at the University of Lausanne. The image on the top-left in Figure 8 shows a fingermark, the infamous Latent Fingerprint Number 17 ("LFP 17"), that was recovered on a bag containing explosive devices in Madrid following terrorist bomb attacks on several commuter trains on March 11, 2004.[114] LFP 17 has now been determined to belong to Ouhnane Daoud, an Algerian national, whose right middle fingerprint is shown centered on the top in Figure 8.

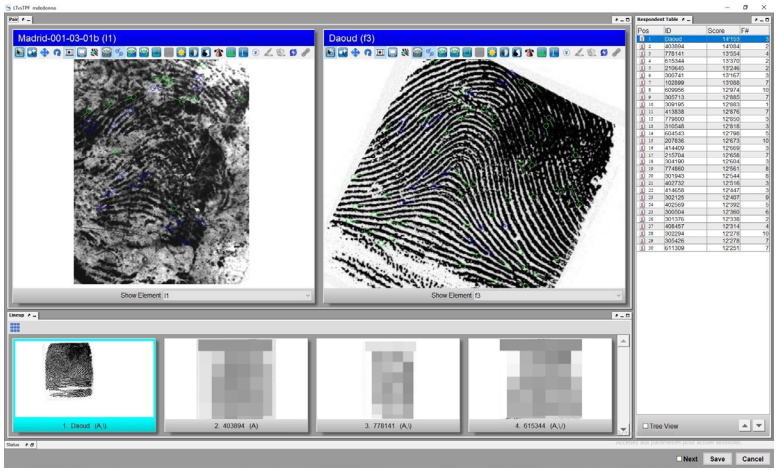


Figure 8: Illustration of the result of a computer-assisted comparison of a fingermark against a database of fingerprint references.[115] The top-left image shows the input fingermark whereas the top-middle (and bottom-left) image shows the reference fingerprint with the highest score (i.e., degree of similarity). Both images are publicly available. The list on the top-right contains the 30 highest scoring candidate fingerprints. Preview-images of the candidate fingerprints ranked 2nd, 3rd, and 4th are masked (images at the bottom).

114. Before LFP 17 was attributed to Ouhnane Daoud by the Spanish National Police, the FBI's Latent Print Unit attributed it to Brandon Mayfield, a U.S. citizen. For a summary of Mayfield's arrest and subsequent litigation, see *Mayfield v. United States*, 599 F.3d 964 (9th Cir. 2010). When LFP 17 was processed by the FBI, the resulting candidate list ranked a reference print from Mayfield in position 4 out 20. *See* OFFICE OF THE INSPECTOR GEN., OVERSIGHT & REVIEW DIV., DEP'T OF JUSTICE, A REVIEW OF THE FBI'S HANDLING OF THE BRANDON MAYFIELD CASE 31 (2006).

115. Courtesy of Marco De Donno and Professor Christophe Champod, School of Criminal Justice, University of Lausanne.

This example illustrates a few key-points regarding the computer-assisted information processing in forensic (fingerprint) comparisons. First, due to the limited quality of the input information and the large number of comparisons conducted,[116] it is not surprising to find closely resembling candidate reference prints, called *look-alikes* in the practice.[117] In our example search, Daoud's fingerprint scored first among the 30 candidates, listed on the right, with a score of 14'153. The second-ranked candidate has a similarity score of 14'083, which is quite close to Daoud's score.

Second, the members of the candidate list of reference prints will generally show lower degrees of similarity when the quality of the input fingermark is poorer.[118] The reason for this is that a low-quality mark offers a feature set with reduced discriminative capacity.

Third, it is thus clear that the limitations inherent in the input information represent a major hindrance to autonomous identification. At this time, computer systems are not allowed to render identification *decisions*; they are only used to help retrieve similar candidates for further one-to-one comparison by human examiners.

Therefore, AI output in the context of forensic fingerprint examination does not amount to an identification decision, but rather a series of similarity assessments for multiple potential candidates. With regard to the topic discussed throughout this Article—how to assess AI output—the question of interest is what to conclude from a particular similarity assessment. This question is discussed in Section III.C.4.

116. Over 1.2 million in the particular search conduced for the purpose of the illustration presented here as compared to the several millions conducted by the FBI in the Mayfield case. *Compare* De Donno, *supra* note 115, *with* OFFICE OF THE INSPECTOR GEN., OVERSIGHT & REVIEW DIV., DEP'T OF JUSTICE, *supra* note 114, at 1.

117. *See* Davide Maltoni et al., *Automated Fingerprint Identification Systems: From Fingerprints to Fingermarks*, *in* HANDBOOK OF BIOMETRICS FOR FORENSIC SCIENCE: ADVANCES IN COMPUTER VISION AND PATTERN RECOGNITION 37, 52 (Massimo Tistarelli & Christophe Champod eds., 2017).

118. *See id.*