Artificial Intelligence and Pattern Evidence: A Legal Application for AI

Patrick Juola

Duquesne University, Pittsburgh, PA, 15282, USA,
juola@mathcs.duq.edu, home page: http://mathcs.duq.edu/~juola

Abstract. Artificial intelligence changes everything, and almost no jobs will be immune. The application of AI to the practice of law is well-known and well-understood. In this paper, we present some aspects of the related disciplines of forensic science and specifically the development and analysis of "pattern [and impression] evidence." We show that pattern evidence has a great need for AI. We discuss several applications in detail but focus mostly on the application of AI-based text analysis technology to forensic linguistics.

Keywords: AI and law, forensic science, text analysis

1 Introduction

Artificial intelligence (AI), and specifically the joint data science and machine learning revolutions, are fundamentally changing the way that the world works. Where the industrial revolution radically changed the nature of blue-collar jobs by handling much of the physical labor, AI promises to radically change white-collar jobs by handling much of the routine intellectual labor. For example, search engines perform much of the job of reference librarians or research assistants; speech-to-text systems handle the work of stenographers, and machine learning systems can do routine tasks such as interpret radiology images.

The practice of law has been greatly altered as well [1]. Document review in civil litigation can be an expensive and time-consuming task involving manual inspection of hundreds of thousands of documents. "Technology Assisted Review" systems [31] use text classification to "learn" what the characteristics of relevant documents are, and perform with better than human-level accuracy in a small fraction of the time and cost [16]. Patent search technology similarly makes it very easy to find prior work related to a new or disputed invention. Contract analysis (determining, for example that a proposed contract doesn't contain a choice of law provision) is more accurate, more efficient, and cheaper if done by a computer. Ainsworth [1] identified several other areas where AI is likely to have a major effect in the near future.

In this paper we focus on not just the routine practice of law, but on a related area, that of developing scientific ("forensic") evidence. We argue that forensic science is an important and underserved area ripe for the application of artificial intelligence, one that is capable of producing great benefits to society at large.

2 Forensic Science

An important interdisciplinary subfield within the law is that of forensic science, which can be loosely defined as the application of science for the purposes of a court of law. As the American Academy of Forensic Sciences puts it:

> The forensic sciences are used around the world to resolve civil disputes, to justly enforce criminal laws and government regulations, and to protect public health. Forensic scientists may be involved anytime an objective, scientific analysis is needed to find the truth and to seek justice in a legal proceeding. Early on, forensic science became identified with law enforcement and the prosecution of criminal cases—an image enhanced by books, television, and movies. This is misleading because forensic science is objective, unbiased, and applies equally to either side of any criminal, civil, or other legal matter.[1]

With the rise of television shows such as CSI or NCIS, the use of forensic science as a tool in criminal investigation is well-understood by the general public, but the actual use of forensic science is much broader. For example, forensic genealogists [20] are rarely involved directly in criminal cases, but may be used to locate heirs for probate, help find missing heirs, assist adoptees with finding their birth parents, or even help DNA experts locate relatives of a person of interest for genetic studies. Forensic chemists may determine the composition of a material to see if a product is safe to sell. Forensic engineers can reconstruct accidents to aid investigators in determining events, their causes, and damages. As will be discussed later, even linguistics has its forensic applications. Forensic linguists have been asked to investigate documents relevant to criminal cases [7, 15], but have also contributed to the resolution of commercial disputes [35] or even helped to resolve questions of identity in immigration hearings. [26]

As with other types of "expert" evidence, the use of forensic science is governed by the local rules of law. In the United States,[2] the admission of expert evidence in Federal courts is governed by the Federal Rules of Evidence as interpreted through three famous cases.[3] FRE 702 lays down four criteria that must be met before an expert may testify. In broad terms, these criteria are simply that the testimony must be helpful, reliable, and factually well-grounded. *Daubert*, in turn, provides (explicitly Popperian [44]) guidelines for judges to assess whether or not these criteria are met. Canadian law[4] of course does not rely on US law, but has explicitly adopted similar criteria. The joint legal system of England

[1] AAFS, https://www.aafs.org/home-page/students/choosing-a-career/what-is-forensic-science/, accessed 29 April, 2019.

[2] The author bases most but not all of his practice in the US.

[3] *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579 (1993); *General Electric Co. v. Joiner*, 522 U.S. 136 (1997); *Kumho Tire Co. v. Carmichael*, 526 U.S. 137 (1999) — collectively often called the *Daubert* trilogy

[4] R. v. Mohan, [1992] 2 S.C.R. 9; R. v. J. (J.-L.), [1999] 130 C.C.C. (3d) 541 (Que. C.A.)

and Wales [54] has developed a four-requirement system ("assistance," "relevant expertise," "impartiality," and "evidentiary reliability") with a similar effect. Many jurisdictions that do not provide explicit reliability requirements (e.g., Scotland [34]) are regarded as problematic and the American *Daubert* standard has been explicitly cited as a model to follow.

In theory, these rules should ensure that the evidence that reaches the judge and the jury is well-founded and reliable. Indeed, under *Daubert*, decision-makers should even know the "known or potential rate of error" as measured through testing. A particular treatment of evidence that has been shown under a variety of controlled conditions — "supported by appropriate validation," in the words of *Daubert* — to yield a correct answer in a high percentage of cases can and should be given more credence than a simple "we've always done it this way." Even the evidence used by police in investigating a crime should meet these standards, as otherwise the police would risk having a case dismissed for lack of reliable evidence.

Unfortunately, this is not always, or perhaps even commonly, the case. A well-known counterexample is that of Brandon Mayfield [42], who "had been identified [...] as the source of a (partial) fingerprint found on a bag of detonators" connected to a terrorist attack in Madrid, Spain. No fewer than three separate FBI experts independently confirmed that Mayfield had made the (single) fingerprint. At the same time, however, the Spanish National Police laboratory had found that Mayfield was *not* the source of these prints. A court-appointed and supposedly independent expert agreed with the FBI. The mere fact of this discrepancy, if made public, should have suggested that the FBI's determination was neither reliable nor factually well-grounded. Nor, given the risk of proceeding against an innocent man, was it in any way helpful.

Later investigation [42] suggested several causes of this erroneous finding. The FBI suggested technical explanations: "the poor quality of the digital image of LFP 17[5], lack of access to the original fingerprint on the bag of detonators, and the similarity of LFP 17 to Mayfield's fingerprint." Later investigators were not so kind: "Several panelists concluded that the initial examiner failed to conduct a complete analysis of LFP 17 before conducting the IAFIS search, which in turn caused him to disregard important differences in appearance between LFP 17 and Mayfield's known prints. Several panelists cited overconfidence in the power of IAFIS and the pressure of working on a high-profile case as contributing to the error. Some panelists stated that the verification was 'tainted' by knowledge of the initial examiner's conclusion." Formal analysis by the Office of the Inspector General concurred: "we concluded that the examiners committed errors in the examination procedure, and that the misidentification could have been prevented through a more rigorous application of several principles of latent fingerprint identification." In other words, this misidentification was caused primarily by human error.

[5] the FBI's copy of the partial print in question

3 Pattern Evidence

Fingerprint identification is one type of what forensic scientists term "pattern [and impression] evidence," defined by the United States National Institute for Science and Technology as:

> any markings produced when one object comes into contact with another object, such as fingerprints, shoeprints, toolmarks, and tire treads. It also includes pattern analysis, such as is used when evaluating handwriting, typewriting, and writing instruments [6].

Fingerprints are perhaps the best-known example of pattern evidence, but other application areas include ballistics markings, toolmarks, bite marks, bloodstains, shoe prints, and questioned documents. Even the study of biometrics, such as facial recognition, can be a form of pattern evidence. Central to this field are the assumptions that, first, (as Ainsworth and Juola [2] put it) "a creator's characteristics are reflected in his or her creation, such that patterns displayed in the creation can provide evidence regarding the identity of it's maker." For example, the pattern of ridges on a fingertip are assumed to be reflected in the pattern of lines in a finger*print* it creates. Similarly, the properties of a typewriter should be reflected in documents typed on it, the properties of shoes and tires should be reflected in their prints, and the properties of teeth should be reflected in their bitemarks.

The second assumption is that the process of forensic analysis can identify these properties with sufficient accuracy, consistency, and reliability to be "helpful, reliable, and factually well-grounded" as described above. Unlike the previous assumption, this is an empirical and technical issue, as it reflects the analysts' ability to identify and measure aspects of the pattern and to connect the dots to a specific individual.

Unfortunately, critical analysis of these assumptions has in many cases been lacking. [40, 49] For many forensic sciences, especially long-standing practices such as latent fingerprint analysis, the validity of the method has simply been assumed and accepted as such by courts. Recent scandals "have called increasing attention to the question of the validity and reliability of some important forms of forensic evidence and of testimony based upon them" [49]. In particular, some studies that have been done have shown [49]:

- "a 2002 FBI re-examination of microscopic hair comparisons the agency's scientists had performed in criminal cases, in which DNA testing revealed that 11 percent of hair samples found to match microscopically actually came from different individuals;
- "a 2004 National Research Council report, commissioned by the FBI, on bullet-lead evidence, which found that there was insufficient research and data to support drawing a definitive connection between two bullets based on compositional similarity of the lead they contain;

[6] NIST, Pattern and Impression Evidence, https://www.nist.gov/oles/pattern-and-impression-evidence, accessed 29 April 2019

- "a 2005 report of an international committee established by the FBI to review the use of latent fingerprint evidence in the case of a terrorist bombing in Spain, in which the committee found that 'confirmation bias'—the inclination to confirm a suspicion based on other grounds—contributed to a misidentification and improper detention; and
- "studies reported in 2009 and 2010 on bitemark evidence, which found that current procedures for comparing bitemarks are unable to reliably exclude or include a suspect as a potential biter."

and further that " FBI examiners had provided scientifically invalid testimony in more than 95 percent of cases where [microscopic hair analysis] was used to inculpate a defendant at trial."

A major further issue is that "expert witnesses have often overstated the probative value of their evidence, going far beyond what the relevant science can justify. Examiners have sometimes testified, for example, that their conclusions are '100 percent certain;' or have 'zero,' 'essentially zero,' or 'negligible,' error rate. As many reviews—including the highly regarded 2009 National Research Council study—have noted, however, such statements are not scientifically defensible: all laboratory tests and feature-comparison analyses have non-zero error rates" [49]. For example, a fingerprint expert testified in *U.S. v. Baines* (2009) that the FBI had an error rate of fewer than one instance in ten million cases [7] In light of these issues, it is clear that the current state-of-the-art in forensic science is problematic in a way that artificial intelligence may be able to address.

The typical structure of a problem in pattern evidence can be understood as a data classification problem. In a typical case (without the use of computer assistance), the analyst will look at one or several "known" samples as well as the "unknown" or "questioned" samples of interest—these samples might be documents, fingerprints, ballistic markings on a bullet, and so forth. From these samples, the analyst will identify features of interest in the samples. In fingerprint analysis, a typical feature (termed "minutia") might be the ending of a fingerprint ridge, a spot where a ridge bifurcates, a "dot" (a ridge of relatively small dimension), a "delta," a scar on a fingertip, and many others, including the inevitable "miscellaneous" category. [32, 56] Associated with these minutia are their locations. The analyst has tremendous discretion about which features to select, and will usually [32, 56] focus on a relatively small (8–12) number of features for a given comparison; for example, [6] marks fingerprint cases for review but will still accept the results unless at least twelve features are used in the analysis. The analyst will then determine which, if any, of these features are shared among samples. If a sufficiently large number of features are shared and a sufficiently small number of features are not shared (again, "sufficient" is often left to the discretion of the analyst), the result is declared to match.

The techniques of forensic DNA analysis, which PCAST has declared "excellent examples of objective methods whose foundational validity has been properly established," follows this structure as well, with some improvements. The

[7] *U.S. v. Baines* 573 F.3d 979 (2009) at 984, cited in [49].

feature set studied comprises 13 specific genetic loci that can be isolated and identified; two samples match if and only if the alleles at these loci are the same. The random match probability (the likelihood of a false positive error) can be calculated based on empirical demographic measurements and the mathematics of population genetics. Key to this precision is the fact that the features used are standardized and defined in advance, the fact that the measurements themselves can be taken with high precision, and the fact that the match criteria are well-understood and objective. DNA is often considered the "gold standard" of forensic science in part for these reasons.

By contrast, bitemark analysis is almost purely subjective and highly inaccurate. [48] While it is possible to measure parameters of human teeth, standards of practice (as with fingerprints) do not specify which features should be analyzed. Furthermore, skin distortion makes it impractical if not impossible to reliably analyze a bitemark, even to determine if a specific injury is a mark produced by human teeth. As a result, PCAST "finds that bitemark analysis is far from meeting the scientific standards for foundational validity" and, damningly, "considers the prospects of developing bitemark analysis into a scientifically valid method to be low. We advise against devoting significant resources to such efforts."

4 Case studies: Forensic Linguistics and Authorship Analysis

Forensic linguistics [14, 9, 43, 36] is a field not reported on by PCAST, but it shows many of the traits of other pattern evidence. Due to its historical connection both with statistical analysis [39] and with computational linguistics [22, 51, 30, 45], though, it is also a good example of how artificial intelligence can be deployed to improve evidence. A typical evidentiary task [7, 15] is to analyze a questioned document to determine its authorship by inspection of the writing style. For example, a document that speaks of parking a "lorry" on the "pavement" in front of an "ironmonger" (instead of a "truck" on the "sidewalk" in front of a "hardware store") is likely to have been written by a speaker of Commonwealth English, as opposed to US English. A more realistic example [50, 17] derives from a kidnapping case, where the ransom note said to put the money in a trash can "on the devil strip" at a particular corner. The examining linguist noted that "devil strip" is a highly regionalized term—used only near Akron, Ohio, USA—and correctly suggested that the police should look for connections with the Akron area.

In general, language is a highly underconstrained system in that there are many ways to express the same thought. Linguist Malcolm Coulthard writes that a person's personal writing style "will manifest itself in distinctive and cumulatively unique rule-governed choices for encoding meaning linguistically in written and spoken communications they produce." [8] Thus, the creation (a document) will reflect the choices of the creator, and at least in theory, these choices can be detected and compared as features to other choices/features in other documents. While some of these choices require highly expert domain

knowledge (as with the "devil strip" example), others can be expressed in fairly straightforward mathematical and algorithmic terms. For example, statisticians since the 19th century have suggested that average word length [38, 37] would be a useful way to detect authorship, on the assumption that some people tend to use bigger words than others. In a now-classic study, Mosteller and Wallace [39] used Bayesian statistics on a carefully chosen set of short, common words to attribute the authorship of *the Federalist papers*. The use of such "stop words" or "function words" has proven to be a widely used [55, 5, 4, 19, 12], robust, and informative feature set that enables accurate authorship determination. Other successful features include the use of character or word n-grams [52] among more than 1000 different feature sets that have been proposed [47].

One key insight that AI has brought to forensic linguistics [2] is that by formalizing analysis procedures and instantiating them in computer algorithms, evaluation of proposed methods is easier, more accurate, and more comprehensive [57, 23]. For instance, the use of *average* word length (in conjunction with statistical methods like t-tests) is not regarded as particularly accurate [18]. As one author [25] put it:

> If you actually get a group of documents together and compare how different they are in average word length, you quickly learn two things. First, most people are average in word length, just as most people are average in height. Very few people actually write using loads of very long words, and few write with very small words, either. Second, you learn that average word length isnt necessarily stable for a given author. Writing a letter to your cousin will have a different vocabulary than a professional article to be published in *Nature*.

However, a simple tweak can greatly improve accuracy; instead of averaging word lengths, one treats word length as a probability distribution (for example, $p_i\%$ of the words in the sample are i letters long) and samples can be compared for similarity using any number of distance metrics (e.g., Euclidean distance, Kolmogorov-Smirnoff distance, etc.) By treating length as a probability distribution instead of a single average, more information is available to the analyst[8] which can be shown empirically to improve performance.

As a tutorial example, it is difficult to improve on Binongo's [4] attribution of the 15th book of *Oz*. The long-running series that started with *The Wonderful Wizard of Oz* changed authors after L. Frank Baum's death, but the authorship of *The Royal Book of Oz* remained in dispute. Was it, as publishers initially claimed, based on a mostly complete draft left by Baum, or was it largely the work of the author hired to continue the series?

To address this, Binongo extracted the fifty most frequent words from the various books of the *Oz* series as well as other works by the same authors, and calculated the token frequencies of these words. These fifty words include words like " He then applied principal component analysis (PCA) to reduce this

[8] As any poker player can tell you, a three and a nine are not the same as a pair of sixes, despite having the same average.

fifty-dimensional space to an easily plottable two dimensions. The resulting plot showed clearly that the first principal component separated the two candidate authors; without exception, all samples by Baum had positive values along this axis, while all samples by the other author had negative values. Upon further showing that all samples from the *Royal Book* also had negative values, the conclusion is fairly clear that Baum did not write even a substantial fraction of that work. As Binongo expressed it, the "stylistic gulf" between the candidates "confirms . . . [that f]rom a statistical standpoint, this book is much more likely to have been written by Thompson than by Baum."

Another high-profile example is Juola's analysis [27, 25] of the pseudonymously published detective novel *The Cuckoo's Calling*. Using a collection of other detective novels, including J.K. Rowling's *The Casual Vacancy*, Juola performed four separate analyses based on four separate feature sets: (1) word length, (2) character 4-grams, (3) word pairs, and (4) the 100 most frequently used words (as per Binongo's analysis above). As Juola wrote, "Of the four authors, Rowling, and only Rowling, was not eliminated by at least one analysis." The accuracy of this analysis was confirmed when Rowling herself acknowledged authorship after this analysis was made public—a rare example of an actual dispute where the ground truth could be confirmed after the fact.

Although there are many ways of addressing this type of question, there are some typical similarities. A typical authorship attribution experiment might run as follows. After selecting a questioned document of interest and collecting sample documents from each of the candidate authors under consideration, one performs the following steps:

- pre-process the documents to convert them into canonical form (canonicize);
- extract features or events from each document. For example, Binongo identified fifty common words as features, while Juola identified the lengths of individual words as one of his feature sets;
- apply a classification technique to determine which candidate author wrote the questioned document. Binongo used PCA; Juola used nearest neighbor with a specific distance formula. More sophisticated analysis can determine probabilities and confidence measures, or possibly decide that none of the above wrote it.

This is a typical instantiation of a traditional AI problem, that of text classification. In this regard, it is little different than language identification or part-of-speech tagging (where a single instance of a word must be identified with its grammatical category based on the features present in the immediate context).

5 Discussion

What does AI bring to forensic science? As mentioned above, AI technology can improve evaluation and testing of forensic methods. While it is not necessary for forensic science to be perfect (and unreasonable to expect it to be), it is certainly desirable for it to be as accurate as practical, and it is arguably *necessary* for the

inaccuracy to be recognized and measured so that the judge and jury can evaluate an expert's findings appropriately. For this reason, the "known or potential rate of error" associated with a given analysis is one of the *Daubert* factors that affect whether an expert's findings will ever see the inside of a courtroom.

While forensic scientists have long recommended proficiency testing as a method of quality control, the recent controversies have demonstrated this to be insufficient. In broad terms, proficiency testing (certification of one practitioner's methods by another practitioner) at best only ensures agreement, but not accuracy. As PCAST put it, "neither experience, nor judgment, nor good professional practices (such as certification programs and accreditation programs, standardized protocols, proficiency testing, and codes of ethics) can substitute for actual evidence of foundational validity and reliability.[...] [A]n experts expression of confidence based on personal professional experience or expressions of consensus among practitioners about the accuracy of their field is no substitute for error rates estimated from relevant studies" [49]. By contrast, AI researchers, including computational forensic linguists, have a long history of evaluating, testing, and publishing the accuracy rates of their systems. The TREC series of competitive evaluations has been extended to a number of similar "bakeoffs" focused on authorship analysis, including the long-running series of workshops sponsored by the Plagiarism Action Network [21, 24, 29, 53]. As an illustrative example, PAN-2013 [29] included texts in three languages, with multiple problems in each language. The computer was asked to determine whether all documents in a problem had the same author. Overall performance was well above baseline, with eight of eighteen participating teams achieving 70% or better on the English Language problems, and combining all eighteen methods into a single ensemble method scored 86.7% correct. *No comparable results are available for human analysts*; in fact, it may not be practical to do similar testing due to the amount of time a purely paper-and-pencil analysis by a human takes.

AI also provides significant advantages in time and efficiency. Binongo's computer analyzed 14 *Oz* books by Baum, 14 by Thompson, plus the disputed 15th book. It also analyzed six non-*Oz* books by Baum, a non-*Oz* work by Thompson, and an *Oz* book by a third, unrelated author. Juola's analysis, performed within a day, analyzed five novels. Forensic laboratories are well-known to be understaffed, underfunded, and overworked—it is impractical for a human to do this much close reading in a timely fashion.

More generally, PCAST has clearly expressed itself on the need for forensic science "to be repeatable, reproducible, and accurate, at levels that have been measured and are appropriate to the intended application." In addition to helping with the accuracy issue, AI systems are much more likely to be repeatable and reproducible. By contrast, humans are subject to many well-known factors such as "confirmation bias" (allowing judgment to be influenced by other factors than the data analyzed). This was confirmed in fingerprint studies [11, 10], where scientists were asked to re-analyze the data from old cases, but with different case information supplied. For example, the scientists might be told that the prints were from the Mayfield case, and that the FBI had erroneously identified

the given prints as a match. In reality, the prints were from previous examples of the scientist's own casework, and had been certified as a match by the scientists themselves. The majority unknowingly changed their opinion, certifying these cases to be a mismatch. In other words, believing something different about the circumstances surrounding the fingerprints affected their conclusions about the fingerprints themselves. By contrast, Binongo's computer (and Juola's) does not know anything about the circumstances, but only about the feature sets fed into it.

More importantly, though, is the formality that artificial intelligence can bring to the process of forensic analysis. For example, to create a feature set suitable for processing by a computer, it is necessary to define exactly what features are being studied. In forensic linguistics, almost any word or phrase can potentially be a feature ("devil strip"?) and more than a thousand feature sets have been proposed in the literature [47]. Large-scale testing enables researchers to determine the most accurate feature sets and thereby recommend best practices. Similarly, there are many ways of assessing similarity [41] and testing can determine the degree to which, for example, support vector machines are more accurate than simple nearest neighbor algorithms. For DNA comparison, as discussed above, scientists have already established a standard set of loci and comparison methods, but research is ongoing and it is not unreasonable to believe that fifteen years from now, there will be a new recommended and standardized set.

By contrast, "There is not currently any method of defining a 'correct' [fingerprint] markup for any given latent. An examiner's decision of whether a minutia is present in an unclear location is analogous to an examiner's decision as to whether the similarity of two prints is sufficient to make an individualization determination: in either case, the best information we have to evaluate the appropriateness of examiners decisions is the collective judgment of other experts" [56]. The lack of standardization can have significant effects:

> "[I]n the vast majority (>90 percent) of identification decisions, examiners modified the features marked in the latent fingerprint in response to an apparently matching known fingerprint (more often adding than subtracting features). (The sole false positive in [one] study was an extreme case in which the conclusion was based almost entirely on subsequent marking of minutiae that had not been initially found and deletion of features that had been initially marked.)" [49]

Indeed, part of the motivation for the cited research is to help develop such standardized methods; the development, widespread use, and acceptance of AI-based fingerprint comparison methods would help push this along.

Bitemark evidence is in a sufficient state as to be perhaps beyond repair, but similar issues hold for other pattern-based methods. PCAST, for example, has this recommendation about firearms analysis, the study of whether or not a given gun fired a given bullet based on "toolmarks" left by the gun: "'to convert firearms analysis from a subjective method to an objective method. This

would involve developing and testing image-analysis algorithms for comparing the similarity of tool marks on bullets." One possibility, discussed by NIST, is the creation of a 3D image of a bullet, extracting a "signature" automatically from that image, then using that image as a feature set for automatic comparison. NIST writes something similar: "While validation studies of firearms and toolmark analysis schemes have been conducted, most have been relatively small data sets. If a large study were well designed and has sufficient participation, it is our anticipation that similar lessons could be learned for the firearms and toolmark discipline [9]. "We are unaware of any study that assesses the overall firearm and toolmark disciplines ability to correctly/consistently categorize evidence by class characteristics, identify subclass marks, and eliminate items using individual characteristics" [10]. Footwear has been singled out by NIST as requiring further research both in terms of feature sets ("[aspects of] the evidence") as well as comparison methods ("aspects of the exam process")[11].

Perhaps the most important aspect of the proposed application, though, is the improvement of the application of justice through wider availability. As mentioned above, forensic laboratories are perennially understaffed and overworked; AI-based tools can improve the speed of analysis without loss of accuracy or reliability. They can also be made available in areas where local conditions and funding preclude the easy availability of forensic laboratories, but not the need of access to justice.

Furthermore, the use and testing of AI systems is likely to improve the democratization of forensic science for technical reasons as well. Systemic bias can be problematic both for human decision-makers and for computational ones. As a simple example, humans have an easier time recognizing faces of people of their own race than of other races [3] [12]. Perhaps surprisingly, this is true for computerized facial recognition software as well; facial recognition performance is generally better on Caucasians [13, 33]. Since the software itself, of course, has no race, this is largely a joint property of the features used to perform the classification and of the classification algorithm itself. If the research on choosing and comparing features or algorithms itself relies on biased data, then the resulting

[9] Forensic Science International, Vol. 208, (2011): 5965. 159OSAC Research Needs Assessment Form. "Study to Assess The Accuracy and Reliability of Firearm and Toolmark." Issued October 2015 (Approved January 2016). Available at: www.nist.gov/forensics/osac/upload/FATM-Research-Needs-Assessment_Blackbox.pdf.

[10] Research Needs Assessment Form. "Assessment of Examiners' Toolmark Categorization Accuracy." Issued October 2015 (Approved January 2016). Available at: www.nist.gov/forensics/osac/upload/FATM-Research-Needs- Assessment_Class-and-individual-marks.pdf

[11] Research Needs Assessment Form. "Examiner Reliability Study: Black/White Box Study on Footwear and Tire Examiners." www.nist.gov/forensics/osac/upload/SAC-Phy-Footwear-Tire-Sub-R-D-001-Examiner-Reliability-Study_Revision_Feb_2016.pdf

[12] Wikipedia suggests (https://en.wikipedia.org/wiki/Cross-race_effect) that there is a same-gender and same-age bias as well.

system is likely to perform worse on test cases outside of its core competency. This effect—that categorization systems tend not to generalize to novel data sets as well as we would like—is generally well-known.

However, this is likely to apply to forensic science as well. There are, for example, well-known genetic markers associated with particular populations; for example, blood type A is most common in Central and Eastern Europe, found in roughly half the population, and almost nonexistent among indigenous populations of Central and South America. This implies that blood type could be a useful feature in DNA analysis to identify individuals, but would not be useful among indigenous tribes in the Andes. In fact, using blood type as a feature would effectively make such tribal populations artificially homogenous, while ignoring other (unknown) features that might serve to distinguish this particular subgroup. Similarly, research has shown that there are effects of race and sex on fingerprints and other associated biometrics [58, 46]. Either a computer program or a human analyst will select features that have proven to be useful *in their past experience*, which is probably not representative either of the world as a whole or of a particular subsection far, far away.

Here again, though, artificial intelligence may have the advantage. It is easier to retrain a computer system than a person, and it is therefore practical for a small research group to develop a suitable data set to look at the features of a particular population of interest in order to better serve that particular population. Again, forensic linguistics provides a case study; while (as expected) most authorship analyses focus on major languages with millions of speakers, Juola [28] showed that it was practical to perform forensic authorship analysis in the Arapaho language (a North American indigenous language) using an existing corpus of language and off-the-shelf open-source software. Similar projects could be undertaken in Quechua, Guarani, Mapudpngan, and other indigenous languages in Argentina and elsewhere, again enhancing access to justice that would otherwise be denied to them due to an absence of appropriately-trained human forensic scientists.

6 Conclusions

The quest for justice is one of the most important social problems. To make the right decision, it is important, perhaps even critical, to have the right information. In this paper, we have presented some examples of how courts can decide which information is "right" (that is, helpful, reliable, and factually well-grounded) and discussed some real-world problems with the information typically offered to courts. We contend that the application of artificial intelligence in forensic science can improve quality and access issues.

References

1. Ainsworth, J.: Killer apps for the practice of law: Past, present, and future. Keynote Presentation given at the Symposium on Artificial Intelligence and Appellate Practice, Cumberland School of Law, Samford University (2018)

2. Ainsworth, J., Juola, P.: Who wrote this? modern forensic authorship analysis as a model for valid forensic science. Washington University Law Review 96(5), 1159–1187 (2019)
3. Behrman, B.W., Davey, S.L.: Eyewitness identification in actual criminal cases: An archival analysis. Law and Human Behavior 25(5), 475–491 (2001)
4. Binongo, J.N.G.: Who wrote the 15th book of Oz? an application of multivariate analysis to authorship attribution. Chance 16(2), 9–17 (2003)
5. Burrows, J.F.: 'an ocean where each kind. . . ' : Statistical analysis and some major determinants of literary style. Computers and the Humanities 23(4-5), 309–21 (1989)
6. Chapman, W., Hicklin, R.A.: ACEware Latent Fingerprint Identification: Research and Software Development. Noblis (2017)
7. Chaski, C.E.: Who's at the keyboard: Authorship attribution in digital evidence invesigations. International Journal of Digital Evidence 4(1), n/a (2005), electronic-only journal: http://www.ijde.org, accessed 5.31.2007
8. Coulthard, M.: Author identification, idiolect, and linguistic uniqueness. Applied Linguistics 25(4), 431–447 (2004)
9. Coulthard, M., Johnson, A., Wright, D.: An Introduction to Forensic Linguistics: Language in Evidence. Routledge (2016)
10. Dror, I.E., Charlton, D.: Why experts make errors. Journal of Forensic Identification 56(4), 600 (2006)
11. Dror, I.E., Charlton, D., Péron, A.E.: Contextual information renders experts vulnerable to making erroneous identifications. Forensic Science International 156(1), 74–78 (2006)
12. Eder, M.: Does size matter? authorship attribution, small samples, big problem. Literary and Linguistic Computing (2013), http://llc.oxfordjournals.org/content/early/2013/11/14/llc.fqt066.abstract
13. Garvie, C., Bedoya, A., Frankle, J.: The perpetual line-up: Unregulated police face recognition in america. https://www.perpetual-linup.org (October 18 2016)
14. Gibbons, J.: Forensic Linguistics: An Introduction to Language in the Justice System. Wiley-Blackwell (2003)
15. Grant, T.: Txt 4n6: Describing and measuring consistency and distinctiveness in the analysis of SMS text messages. Journal of Law and Policy XXI(2), 467–494 (2013)
16. Grossman, M.R., Cormack, G.V.: Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review. Richmond Journal of Law and Technology 17, 1 (2010)
17. Hitt, J.: Words on trial. New Yorker 88(21), 24–29 (2012)
18. Holmes, D.I.: The evolution of stylometry in humanities scholarship. Literary and linguistic computing 13(3), 111–117 (1998)
19. Hoover, D.: Collocations, authorship attribution, and authorial style. In: Proceedings of *ACH/ALLC-2003*. Athens, GA (2003)
20. Jakman, A.: Forensic genealogy—the real story. Forensic Magazine (2016)
21. Juola, P.: Ad-hoc authorship attribution competition. In: Proc. 2004 Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities (ALLC/ACH 2004). Göteborg, Sweden (June 2004)
22. Juola, P.: Authorship attribution. Foundations and Trends in Information Retrieval 1(3) (2006)
23. Juola, P.: Large-scale experiments in authorship attribution. English Studies 93(3), 275–283 (May 2012)

24. Juola, P.: An overview of the traditional authorship attribution subtask. In: Proceedings of PAN/CLEF 2012. Rome, Italy (2012)
25. Juola, P.: Rowling and 'Galbraith': An authorial analysis. https://languagelog.ldc.upenn.edu/nll/?p=5315 (July 16 2013)
26. Juola, P.: Stylometry and immigration: A case study. Journal of Law and Policy XXI(2), 287–298 (2013)
27. Juola, P.: The Rowling case: A proposed standard protocol for authorship attribution. DSH (Digital Scholarship in the Humanities) (2015)
28. Juola, P.: Authorship attribution in a Native American language (Arapaho). In: 2018 Annual Meeting of the Linguistic Society of America (2018)
29. Juola, P., Stamatatos, E.: Overview of the authorship identification task. In: Proceedings of PAN/CLEF 2013. Valencia, Spain (2013)
30. Koppel, M., Schler, J., Argamon, S.: Computational methods in authorship attribution. Journal of the American Society for Information Science and Technology 60(1), 9–26 (2009)
31. Lau, T.T., Lee III, E.G.: Technology-Assisted Review for Discovery Requests: A Pocket Guide for Judges. Federal Judicial Center (2017), https://languagelog.ldc.upenn.edu/nll/?p=5315
32. Lee, P., Guan, H., Dienstfrey, A., Theofanos, M., Stanton, B., Schwarz, M.T.: Forensic Latent Fingerprint Preprocessing Assessment. No. 8215 in NISTIR, US Department of Commerce, National Institute of Standards and Technology (2018)
33. Lohr, S.: Facial recognition is accurate, if youre a white guy. The New York Times 9 (2018)
34. Maher, G.: Guarding the gate: Some problems in expert evidence in Scots law. Tech. Rep. 2015/07, Edinburgh School of Law Research Paper (2015)
35. McMenamin, G.: Declaration of Gerald McMenamin. Available online at `http://www.scribd.com/doc/ 67951469/Expert-Report-Gerald-McMenamin` (2011)
36. McMenamin, G.R.: Forensic Linguistics: Advances in Forensic Stylistics. CRC press (2002)
37. Mendenhall, T.C.: The characteristic curves of composition. Science IX, 237–49 (1887)
38. de Morgan, A.: Letter to Rev. Heald 18/08/1851. In Sophia Elizabeth. De Morgan (Ed.) *Memoirs of Augustus de Morgan by his wife Sophia Elizabeth de Morgan with Selections from his Letters.* (1851/ 1882)
39. Mosteller, F., Wallace, D.L.: Inference and Disputed Authorship : The Federalist. Addison-Wesley, Reading, MA (1964)
40. National Research Council, et al.: Strengthening Forensic Science in the United States: A Path Forward. National Academies Press (2009)
41. Noecker Jr., J., Juola, P.: Cosine distance nearest-neighbor classification for authorship attribution. In: Proceedings of Digital Humanities 2009. College Park, MD (June 2009)
42. Office of the Inspector General; Oversight and Review Division: A Review of the FBI's Handling of the Brandon Mavfield Case. United States Government (March 2006)
43. Olsson, J., Luchjenbroers, J.: Forensic Linguistics. A&C Black (2013)
44. Popper, K.R.: The Logic of Scientific Discovery (1934 orig). Routledge, London (2005)
45. Rocha, A., Scheirer, W.J., Forstall, C.W., Cavalcante, T., Theophilo, A., Shen, B., Carvalho, A.R.B., Stamatatos, E.: Authorship attribution for social media forensics. IEEE Transactions on Information Forensics and Security 12(1), 5–33 (2016)

46. Rowe, R.K.: Methods and systems for estimation of personal characteristics from biometric measurements (Aug 28 2007), US Patent 7,263,213
47. Rudman, J.: The state of authorship attribution studies: Some problems and solutions. Computers and the Humanities 31, 351–365 (1998)
48. Saks, M.J., Albright, T., Bohan, T.L., Bierer, B.E., Bowers, C.M., Bush, M.A., Bush, P.J., Casadevall, A., Cole, S.A., Denton, M.B., et al.: Forensic bitemark identification: Weak foundations, exaggerated claims. Journal of Law and the Biosciences 3(3), 538–575 (2016)
49. of Advisors on Science, P.C., (US), T.: Report to the President, Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods. Executive Office of the President of the United States (September 2016)
50. Shuy, R.: DARE's role in linguistic profiling. DARE Newsletter 4(3), 1–5 (2001)
51. Stamatatos, E.: A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology 60(3), 538–56 (2009)
52. Stamatatos, E.: On the robustness of authorship attribution based on character n-gram features. Journal of Law and Policy XXI(2), 420–440 (2013)
53. Stamatatos, E., Stein, B., Daelemans, W., Juola, P., Barrón-Cedeño, A., Verhoeven, B., Sanchez-Perez, M.A.: Overview of the authorship identification task at PAN 2014. In: Proceedings of PAN/CLEF 2014. Sheffield, UK (2014)
54. The Law Commission: Expert Evidence in Criminal Proceedings in England and Wales. The Stationery Office, London (2017), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/229043/0829.pdf
55. Tweedie, F.J., Singh, S., Holmes, D.I.: Neural network applications in stylometry : The Federalist Papers. Computers and the Humanities 30(1), 1–10 (1996)
56. Ulery, B.T., Hicklin, R.A., Roberts, M.A., Buscaglia, J.: Interexaminer variation of minutia markup on latent fingerprints. Forensic Science International 264, 89–99 (2016)
57. Vescovi, D.M.: Best Practices in Authorship Attribution of English Essays. Master's thesis, Duquesne University (2011)
58. Zavala, A., Paley, J.J.: Using fingerprint measures to predict other anthropometric variables. Human Factors 17(6), 591–602 (1975)