SAVANNAH
LAW REVIEW

VOLUME 5 | NUMBER 1

Legal Jobs in the Age of Artificial Intelligence: Moving from Today's Limited Universe of Data Toward the Great Beyond

Philip Segal[*]

I. Introduction

When telephones were new, many respectable lawyers would not have one on their desks. Senior partners greeted word processors and computers the same way in the 1970s. Those efficiencies were unstoppable, and in their presence more law jobs were produced than ever before.

What we call artificial intelligence is no different. Messengers and typists lost their jobs when legal and non-legal firms adopted word processing, computers, e-mail, voicemail, and other new technology. They were replaced with more people who could add more value. Artificial Intelligence (AI) will do the same thing: out will go the lawyers who do repetitive, uncreative work. Lawyers who survive will need particular skills to manage the machines that will make law more productive and therefore more affordable. The productivity will make room for lawyers in brand new fields—some foreseeable and some not.

This is good news for everyone except the lawyers who will be losing their jobs. But the efficiencies will elude the firm leaders who think buying software is all they need to get lean. Much of the literature about AI and the law presupposes an easy technology-adoption process, but the hardest part about using good software is getting the right people to operate it the right way. Operating sophisticated software can be more art than science. Deciding what to do with the data machines produce is anything but straightforward.

Most of the attention paid to artificial intelligence in law so far has dealt with factual as opposed to legal inquiries: computer programs engaged in pattern

[*] Managing Member, Charles Griffin Intelligence, New York; MSL Yale Law School, 2004; JD, Benjamin N. Cardozo School of Law, 2006. Author, THE ART OF FACT INVESTIGATION (Ignaz Press, 2016).

matching scan large amounts of data (large amounts for a person but trivially small for a machine). Nearly all the remainder of the products on the market today or in development deal in limited data as well: with legal questions concerning precedent and probable outcomes, or knowledge management—making law firms more efficient in their internal operations. If some of the data is not fully structured in these applications, it is more uniform and economical to manage than the raw oceans of pages and video floating free on the internet.

Table 1: Legal/Factual Tools Based on Size of Data Set

Fact Questions – Limited Data	**Legal Questions – Limited Data**
E-discovery (e.g. Catalyst's TAR)	Legal Research engines (e.g. ROSS Intelligence)
Contract analysis (e.g. Kira, LawGeex, LexCheck)	Case outcome predictors (e.g. Lex Machina)
Enterprise Search and Knowledge Management Programs (e.g. iManage-RAVN)	Robot lawyers (DoNotPay for parking tickets)
High-risk emails (e.g. Intraspexion)	
Contract Execution with Blockchain	
Fact Questions – Unlimited Data	**Legal Questions – Unlimited Data**
Current-generation databases such as LexisNexis and Bloomberg, and the largely unexplored area for next-generation AI.	Robot judges (Science Fiction, for now)
Dozens of useful applications for lawyers (The subject of Part IV of this article).	

In the space of unlimited data—at least on the fact gathering side—this Article argues that there are dozens of AI applications currently in development that could help lawyers do their jobs better. The impact on legal employment from these unlimited-fact universe programs does not appear to be negative, for the simple reason that they represent a chance to look at data *previously unexamined* by lawyers. But, to be useful, these programs will demand skills and a job outlook not always in evidence among lawyers.

As for legal questions with unlimited data, this question takes us to automatic judging, in which judges and litigants employ analogical reasoning plucked from a virtually unlimited universe of facts. The computers we use today are not able to

simulate this kind of thinking, much less do it themselves. This article will not speculate about robot judges.[1] It will proceed as follows:

Part II explores the impact of AI on other industries with instructive lessons for the relatively later-adopting legal profession. Other industries instruct that we should not be surprised that even applications that are being most heavily adopted by large law firms leave plenty of room for human input to run the programs.

Part III discusses the major differences between the closed or highly limited universes of data with which today's most popular AI programs deal, beset by algorithmic anomalies and biases, and the vastly larger, unstructured field of data that the AI of tomorrow will attempt to conquer. Many of today's problems encountered by AI are not programming-related but stem from commercial concerns and statutory limitations that may survive advances in computing power. It is not a foregone conclusion that today's issues of databases not communicating will be solved with more data coming in at a faster rate. The less efficient our use of AI continues to be, the more people will be needed to fill gaps.

Part IV discusses some of the applications of new AI that lawyers do not use today, but could easily adopt for fact investigation in the future (the lower left quadrant in Table 1, above). Entire new areas of application will be one way new legal jobs will arise from the wreckage of laid-off document reviewers. Many of these applications are already in the development stage, but are not aimed at lawyers at all. This part also explores the potential developments of AI in law that may be ten years away, but for which the prescient law firm should be preparing today.

Part V argues that adoption of any artificial intelligence (in all of the four quadrants in Table 1) will require all of the same skills as were needed with yesterday's now-routine legal software (skills which were often lacking). Even today, software used in law firms is not employed to its full potential. Lawyers often miss the point—that while computers work purely in a deductive way, higher human functioning is often not deductive at all. What kind of person does a firm want to have running all the sophisticated machinery? It may not be the ideal law student of today. This Article concludes with suggestions about how to train lawyers to use AI wisely. Those expecting to type in a few queries and get a final correct answer will be sorely disappointed.

II. The Lesson of AI Adoption Outside the Law

The definitions of artificial intelligence are many. One definition is: the capacity of a computer to perform operations analogous to learning and decision

[1] This does not mean "never" for robot judging. We should be mindful of the fact that ten years ago many of us were convinced that driverless cars were only a thing of science fiction, while today many of us concede that driverless cars will be on the roads before long. Our grandchildren may laugh at old clips of people who said it couldn't be done, just the way we do at IBM President Thomas Watson's prediction in the 1940s that the world could accommodate five computers at most. Who is to say that in simple cases, litigants would not opt for a robot judge that could save them years of worry and tens of thousands of dollars?

making in humans.[2] Some expand on this and say that AI is designed to give computers "human-like abilities of hearing, seeing, reasoning and learning."[3] Taking different entries of numbers and putting them into a balance sheet or profit and loss statement, as does QuickBooks, would qualify under the former. The latter may be equated to a more advanced form of AI known as machine learning, the general goal of which "is to analyze past data to develop rules that are generalizable going forward."[4] The most advanced kind of AI, Deep Learning, is a subset of machine learning that allows computers to train themselves without being programmed. Deep Learning is the place of doomsday projections, with computers running out of control as they turn on their human inventors.[5]

Some divide AI into "soft" and "hard" varieties. If AI mimics human intelligence, "soft" AI does so in outcomes, not in process.[6] "Hard" AI is the stuff of apocalyptic visions of "The Singularity," when computers are as smart as the people who programmed them. A good example of soft AI is Google Translate, which is able to provide serviceable (if often flawed) translations of simple phrases. Instead of "learning" a language as people do, the program is fed millions of examples of translation between two languages, and merely reproduces what it has already seen. No person learns a language by listening to millions of hours of conversation before speaking, so Google imitates outcome rather than process.[7]

Physicist and geneticist Christoph Adami—who is working on simulating the evolution of millions of years of human knowledge in robots by running thousands of generations of trial and error exercises with machines—divides knowledge into two types: the first is the product of logical inputs.[8] This is also called "narrow" AI.[9] Adami's second kind of intelligence is the kind that classifies sensory information and is the product of long-term stored memory. It is the mind saying to itself, "What is this? Have I seen it before? How should I react?" Programming

---

[2] *Artificial Intelligence*, DICTIONARY.COM, http://www.dictionary.com/browse/artificial-intelligence (last visited Apr. 24, 2018).

[3] *See* Jessica Groopman & Clint Wheelock, *Artificial Intelligence: 10 Key Themes Across Use Cases*, TRACTICA LLC AND THE AI SUMMIT 2 (2017), https://www.tractica.com/resources/white-papers/artificial-intelligence-10-key-themes-across-use-cases/ (last visited Apr. 24, 2018).

[4] Harry Surden, *Machine Learning and Law*, 89 WASH. L. REV. 87, 105 (2014).

[5] While there is the beginning of a discussion among academic lawyers about how to apportion liability when more advanced computer programs cause harm, the vast majority of the literature is about automating routine tasks that today consume large amounts of lawyers' time and clients' budgets.

[6] Daniel Martin Katz, *Quantitative Legal Prediction—or—How I Learned to Stop Worrying and Start Preparing for the Data-Driven Future of the Legal Services Industry*, 62 EMORY L.J. 909, 936 (2013).

[7] On an even simpler level, when was the last time any of us accepted *all* spelling and grammar changes suggested by MS Word in a document of any serious length?

[8] Chris Adami, *A New Path Towards Intelligent Machines*, THE STANFORD COMPLEXITY GROUP, YOUTUBE (May 5, 2015), https://www.youtube.com/watch?v=PHJnJZrUiW0.

[9] *See* comments by Oren Etzioni, CEO of the Allen Institute for Artificial Intelligence in Seattle: Kate Baggaley, *There are Two Kinds of AI, and the difference is important*, POPULAR SCI. (Feb. 23, 2017), https://www.popsci.com/narrow-and-general-ai.

this kind of reasoning is "fiendishly difficult."[10] This second form of intelligence is sometimes called general intelligence. While machines "probably have developed superhuman pattern recognition abilities . . . that's only a small part of what is general intelligence."[11]

With the large range of definitions of AI, it is interesting that almost the only firms that think they have adopted AI are those with 1,000 or more lawyers.[12] But even among this largest group of firms, half have either not yet started implementing AI or have no plans to do so.[13]

The idea that most lawyers think their firms do not use AI when, undoubtedly, their firms are all computerized and using software probably confirms the old quip that something is considered AI right up until we use it, at which time we start calling it software. A variation on this is that if, in discussing AI, we replace it with the word "software," no information is lost. If a law firm uses billing software, a Lexis Nexis database, or Google, it is using AI. If you could return to 1975, and look ahead to today, you would think it miraculous that lawyers can get a list in two seconds of someone's residences, many company affiliations, some litigation drawn from all kinds of routine public records, and credit transactions.

What accounts for this disconnect? Jerry Kaplan of Stanford Law School thinks one problem is AI's name. "Had AI been named something less spooky, it might seem as prosaic as operations research or predictive analytics. . . . We should stop describing these modern marvels as proto-humans and instead talk about them as a new generation of flexible and powerful machines."[14]

What lawyers may think of as AI are new programs that promise to slash the number of billable hours their firms can work, such as the kind of program at work at JP Morgan, which the bank claims has saved 360,000 hours of work by lawyers and loan officers who now let computers interpret (to an extent) commercial loan agreements.[15] That sounds like a lot of jobs and it is, but consider: for a 40-hour per week employee with three weeks' vacation, this many hours eliminated comes to 183 person years of work. In a company the size of JP Morgan, that is less than one-tenth of one percent of the number of employees.

Whatever their attitude toward AI, lawyers should count on the idea that technology will continue to change their profession. It is difficult to think of a single area of modern technology that has not penetrated the law firm. The only question is whether law firms will be early or late adopters. Even if some lawyers fought against phones and computers at first, technology has made society more

---

[10] Adami, *supra* note 8.

[11] Baggaley, *supra* note 9 (quoting Oren Etzioni).

[12] Thomas S. Clay & Eric A. Seeger, *Law Firms in Transition: An Altman Weil Flash Survey*, ALTMAN & WEIL INC. (2017), http://www.altmanweil.com/LFiT2017/.

[13] *Id.*

[14] Jerry Kaplan, *AI's PR Problem*, MIT TECH. REV. (Mar. 3, 2017), https://www.technologyreview.com/s/603761/ais-pr-problem/.

[15] Hugh Son, *JPMorgan Software Does in Seconds What Took Lawyers 360,000 Hours*, BLOOMBERG (Feb. 27, 2017, 7:31 PM), https://www.bloomberg.com/news/articles/2017-02-28/jpmorgan-marshals-an-army-of-developers-to-automate-high-finance.

efficient and did not stand in the way of (and probably helped to cause) a huge increase in legal services as lawyers began drafting their own documents and sending their own letters (email). The idea that a solo practitioner could function without a bookkeeper and secretary would have been unthinkable 50 years ago and is unquestioned today.

Other than the apocalyptic scenarios of robots turning on us,[16] what can be so jarring about technology is that the jobs it creates are often filled by people *different than those whose jobs the technology replaced*. The likelihood has always been and continues to be that the more high-level thinking you do in your job, the less likely it is you will be replaced by a machine.

Bookkeepers lost their jobs to QuickBooks, but sophisticated accountants did not. The (mostly) women in the typing pool were retired by Microsoft Word and laser printers. But the computers do not talk to the client to see what kind of contract the client wants or whether proposed terms from the other side are fair. Those bookkeepers and typists from 1965 may now be selling us billing software or working in QuickBooks technical support.

The difference with this generation of AI is that the computers can now do some of what lawyers can, though lawyers may take comfort from the lesson of automated teller machines (ATM). These were introduced in the 1980s and produced layoffs of human bank tellers, but the invention of the ATM made bank branches much cheaper to open. The result was that banks opened more branches, each of one of which was more efficient than one larger old branch with more human workers clustered together.[17]

The innovations of Uber and Lyft decimated the number of rides New York City yellow taxis have made, but the number of total rides in the city is up.[18] The difference is that Uber and Lyft provide service where it was not provided before, as well as a more efficient dispatch system (where one hardly existed with the yellow cabs previously because of artificial scarcity of supply). The result is the usual outcome of an efficiency upgrade: more of what you used to have for the same price.

Does the fate of bank tellers and taxi drivers tell us anything useful about attorneys? A Capgemini study of 993 non-legal companies surveyed in 2017 reported that 83% of the companies adopting artificial intelligence had created new jobs as a result. Two-thirds of those jobs were at the C-suite, Director or Manager level.[19] Even taking into account net job growth, 63% of companies in this study

---

[16] Even if plausible in twenty-five or thirty years, the debate about out-of-control robots is beyond the scope of this article. However, increasingly autonomous AI does promise brand new areas of legal practice. *See* Matthew U. Scherer, *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*, 29 HARV. J.L. & TECH. 354 (2016) (discussing tort liability if advanced AI escapes human supervision as we foresee today). If AI produces a result that was not foreseeable by a human, are programmers liable? How would liability be apportioned between vendors and users?

[17] Greg Ip, *Workers, Fear Not the Robot Apocalypse*, WALL ST. J. (Sept. 5, 2017), https://www.wsj.com/articles/workers-fear-not-the-robot-apocalypse-1504631505.

[18] *Id.*

[19] Digital Transformation Institute, *Turning AI Into Concrete Value: the Successful Implementer's Toolkit*, CAPGEMINI CONSULTING (Sept. 8, 2017),

said that after implementing AI at scale, no jobs had been destroyed in their organizations.[20]

Today the major focus on AI in law is on the analysis of contracts and eDiscovery.[21] Electronic discovery (eDiscovery) software sorts out email chains, weeds out duplicates, and with some human instruction can do as good a job as people in producing "relevant" documents demanded by the other side. Contract software highlights key clauses, helps to sort thousands of contracts into the ones auto-renewing in the next ninety days or entering into their notification phase. The software will not tell you if the contracts are worth renewing or not. Are they fair given the market's changes? Do they fit in with what a business is trying to achieve? None of the AI in contracting today can help with that.[22]

Still, these programs promise major payroll cuts at large firms, which is why they are so attractive to the largest firms and legal departments that do the most complex eDiscovery and have the most contracts to draft and monitor.

Another area that its proponents argue offers efficiency is in analytics relating to proposed settlements[23] (the outcome of the great majority of litigation today). The efficiency of settlement may save money, but the question remains as to how many lawyers' jobs this threatens. If both sides employ AI in evaluating settlements, there is a possibility of a drastic decrease in the amount of litigation. But are lawyers today, under heavy pressure from clients to reduce costs, really so unreasonable about these potential outcomes? Perhaps a better way to avoid prolonged litigation is to avoid litigation—period. This is the aim of Intraspexion, an AI company that has scanned key phrases in various categories of lawsuits.[24] When language in company emails matches key language found in litigation evidence, company lawyers are alerted to conversations that could increase the chances that the company could face a lawsuit. This sounds promising as far as mitigating damages, but would seem to add tasks for lawyers, such as requiring them to evaluate flagged emails. In the largest companies, this could amount to hundreds or thousands of emails a day.

---

https://www.capgemini.com/wp-content/uploads/2017/09/dti-ai-report_final1.pdf. Sectors surveyed were manufacturing, retail, utilities, telecom, banking, and insurance. The U.S. made up a third of the sample, E.U. and U.K. 49% combined, and India and Australia together 17% of those surveyed.

[20] *Id.*

[21] For eDiscovery, Catalyst's TAR (for Technology Assisted Review) is among the best-known programs for grading for relevancy samples from large universes of documents. Programs that can help analyze contracts are many, including Kira, eBrevia, LawGeex, and LexCheck. Apogee Legal appears to offer AI in due diligence, but this is in fact contract analysis in the mergers and acquisition context combined with an internal web-crawling capability that seeks vendor and supplier information.

[22] Although, the vision for Blockchain-enabled contracting is that trigger prices in contracts, for example, could automatically effectuate previously agreed upon changes in contracts.

[23] SETTLEMENT ANALYTICS, https://settlementanalytics.com/ (last visited Apr. 24, 2018).

[24] ARTIFICIAL LAWYER, https://www.artificiallawyer.com/ (last visited Apr. 24, 2018).

On the societal level (where lawyers get their clients) the transformation of so many industries outside the law by new programs will probably be led by consumer applications. When a market goes from $2.7 billion for consumer-use AI this year to a projected $42 billion by 2025,[25] that kind of change is bound to create legal work in the form of new companies, new-product launches, product liability, mergers, and all of the other areas that drive business law.

Sometimes the precise kinds of new jobs that will stem from more advanced technology are difficult to forecast because they will be in areas relating to technology or applications not yet invented. Just as the internet and computer law were all but non-existent in the 1970s, driverless cars, drones, and other robots will present society with vast new fields of practice. The area of tort liability alone is potentially extremely large and filled with public policy questions.[26]

Current fields such as insurance will be transformed by the new technology. Contract management software is one example that could eliminate some jobs but create others. Such programs "have made it economical for companies to retain lawyers to conduct a review of entire repositories of legacy documents, so those companies can better manage their risks and obligations."[27] Prior to these tools,

> companies let these legacy documents languish in a multitude of electronic file folders or worse—plain old paper files—viewing any attempt to get a handle on all of this information as too cumbersome and costly. Here, technology has created an opportunity for the legal market, quite the opposite of the perceived threat that technology will make people superfluous.[28]

Far larger for law firms than the challenge of further analyzing the information already in their possession are the many oceans of new data being created outside the boundaries of their law firms each day. This, too, will require new AI tools to sort through. And as with today's databases[29] (the AI of yesteryear) there is no escaping the need for human intervention and management of the process. The need will be even greater than today because of the far greater amount of data pouring into firms every second.

---

[25] *Consumer Applications Are the Largest Market Segment for Artificial Intelliegence*, TRACTICA RESEARCH (Sept. 5, 2017), https://www.tractica.com/newsroom/press-releases/consumer-applications-are-the-largest-market-segment-for-artificial-intelligence/.

[26] See for instance, Scherer, *supra* note 16, at 393 for a discussion of robot liability in which he proposes as an alternative to the requirement that AI users be required to register with the government. Instead, he proposes an optional registration process that would confer limited tort liability on registrants, and strict liability on non-registrants.

[27] David Perla & Sanjay Kamlani, *Power to the People: Technology is Here to Help Lawyers, Not to Displace Them*, ABOVE THE LAW (June 20, 2017, 6:13 PM), http://abovethelaw.com/2017/06/power-to-the-people/.

[28] *Id.*

[29] The most common such databases used by lawyers include LexisNexis, Westlaw, Bloomberg Law, and TLO.

III. The Fact Finder's Infinite Universe of Data is AI's Challenge for Tomorrow

The future appears to be bright for lawyers engaged in fact investigation. Unlike the limited sets of data used in even the most complex e-discoveries, fact investigation happens in an open universe of data—nearly all unstructured and virtually unlimited. In searching for a good witness, for attachable assets, for the true background of an individual for the purpose of impeaching credibility, there is no finite universe of documents, websites or other digital records to analyze. The information could be anywhere. The great leap for artificial intelligence is the move from what for a computer is a tiny, trivial universe of data to the great beyond—all the written and spoken data available on the internet. While growing, the adoption of AI that deals in unlimited amounts of data is not yet widespread.

Electronic discovery programs and contract review work in small, reasonably uniform universes of data. If not "structured data" according to the formal definition of being in a relational database, documents scanned with optical character recognition are at least quickly reviewable. [30] Contracts being compared have similar clauses named similar things.

Despite the comparative simplicity of these limited-universe applications, a recent survey of in-house lawyers revealed that just 29% thought they were making effective use of extracted data from contracts to develop business strategy and minimize contract risk.[31] Thinking about truly large universes of data in today's law firm is so unusual that the industry-standard Grossman-Cormack Glossary of Technology-Assisted Review[32] contains no definition of data, whether structured or not.

To see how far many law firms have to go in getting to deep learning, it is helpful to think about AI in terms put forth by Monica Rogati, an equity partner at San Francisco venture capital fund Data Collective. She argues that many companies that want to embrace AI are not ready to do so:

> [U]nder the strong influence of the current AI hype, people try to plug in data that is dirty and full of gaps, that spans years while changing in format and meaning, that's not understood yet, that's structured in ways

---

[30] Traditionally, "structured" data means data that can fit into a structured query language (SQL) database, such as an Excel spreadsheet. An extensive discussion of structure is far beyond the scope of this article, as is the alternative view that open-source program Apache Hadoop and its "data lakes" constitute a new form of structuring. For the purposes of this article, "structured" means data that is easily searchable across one database.

[31] Thomson Reuters, *Ready or Not: Artificial Intelligence and Corporate Legal Departments,* ABOVE THE LAW (Oct. 11, 2017, 10:30 AM), https://abovethelaw.com/?sponsored_content=ready-or-not-artificial-intelligence-and-corporate-legal-departments.

[32] John M. Facciola, *The Grossman-Cormack Glossary of Technology-Assisted Review*, 7 FED. CT. L. REV. 1 (2013).

> that don't make sense, and expect those [data] tools to magically handle it.[33]

The tools won't handle it: "No amount of algorithmic sophistication will overcome a lack of data."[34]

At the bottom of Rogati's imagined pyramid of needs for AI to work, firms need to decide what data they have and what is available. The amount of data available to law firms in electronic format has steadily risen, but firms have had problems handling even this data in any kind of predictable, uniform fashion.[35] The hazards are clear: "An overload of information (particularly if of low-quality) carries the risk of undermining knowledge acquisition possibilities and even access to justice."[36]

Getting more data in raw form is not hard. The world had created five exabytes (five billion gigabytes) of data in all of history up to 2003, when Google's Eric Schmidt was widely quoted as saying the world now creates this much information in two days. Others have taken issue with this estimate, but even if it were to take a month to create this amount of new data, that is still a tidal wave against the drip-drip-pace of yesteryear.

Next comes an analysis of a firm's data flow. Is the data sent automatically to the law firm? Does the firm need lawyers and others to monitor the data and to choose what comes in and what does not? If there is to be automation, a firm needs to decide what kind of electronic decider it will empower as the gatekeeper. Today, firms take in data on a project-by-project basis in e-discovery, and rely on their own databases of contracts for contract analysis. After deciding about how data flow will work, firms must explore and sometimes transform raw data, and then experiment to see if the data set is robust enough to allow reasonable conclusions to be drawn from it.

To see why data structuring can be particularly problematic for lawyers, consider the current state of our traditional AI (yesterday's AI and today's software that we nonchalantly refer to as "LexisNexis" or "Westlaw"). These are the databases most lawyers work with each day and on which they impose demands that are modest in comparison to the sleek machine learning on the drawing boards for tomorrow's AI-friendly firm.

We have no expectation that today's databases will do any "thinking" for us. An untrained person runs the risk of reporting inaccurate and conflicting information if all he does is take the database conclusions as fact. If the databases retrieve information we can then use (or even just stubs of information that tell us where to look in public records), we are usually happy.

---

[33] Monica Rogati, *The AI Hierarchy of Needs*, HACKERNOON.COM (Aug. 1, 2017), https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007.

[34] Kaplan, *supra* note 14.

[35] Susan Nevelow Mart, *The Algorithm as a Human Artifact: Implications for Legal [Re]Search*, 109 L. LIBR. J. 387 (2017).

[36] Mark van Opijnen & Cristiana Santos, *On the concept of relevance in legal information retrieval*, ARTIFICIAL INTELLIGENCE L. (2017), https://link.springer.com/content/pdf/10.1007%2Fs10506-017-9195-8.pdf.

Today's most powerful fact-finding databases are riddled with basic errors about location, assets, and corporate affiliations of individuals. All make different errors that stem not from inadequate computing power but, in part, from legal and financial barriers that will not necessarily disappear as software becomes more sophisticated. An AI program that could evaluate the likelihood of accuracy of competing databases does not exist, and faces several barriers that could survive expected increases in computing power. The more resistance there is to smooth integration of databases, the more people will be needed to sort the various outputs and fit them together to form a single, intelligible picture.

For instance, the databases are providers of restricted information accessible to licensees only under the terms of the Gramm Leach Bliley Act,[37] the Drivers Privacy Protection Act of 1994,[38] as well as varying state laws. The availability of the information depends on a variety of permissible uses under the statutes. In addition, the information is expensive to purchase. Investigators today are unable to use a "Kayak" program that would search all such databases at once and generate five side-by-side windows with comparable results. Each database requires input of payment information that is only accepted after a site visit by the database to ensure that the user is properly licensed. Databases will not sub-contract that job to a central entity that handles its competitors as well.[39]

As mentioned, the output of the databases is error-prone and requires careful human oversight. At a recent conference, I demonstrated the output of one of these databases after entering my own name and address. The database was under the impression that I still used a telephone number that has been disconnected for seven years, and that I worked at an office address vacated in 2011. The probable reason was that the phone number remains on a store discount card I use, which causes the database to conclude that it is an active number. The office address may have come from an outdated entry on Google, which required me to update it manually.[40]

There is also a connections problem inherent in today's databases—not only that they may make too many connections, but also that they may make too few. Databases that are commercial competitors do not share information and thus fail to update stale or incorrect knowledge. Information scientist Don Swanson[41] demonstrated the consistent inability of academic researchers to share and

---

[37] 15 U.S.C. §§ 6821–6827 (1977).

[38] 18 U.S.C. § 2722 (1994).

[39] Nonetheless, databases that accumulate free and unlicensed content from a variety of sites on the internet exist and are being improved daily. Many of these hold much promise for attorneys even though they are not marketed to attorneys at all. This is discussed in Part IV.

[40] For a detailed discussion of the reliability problems of databases, see PHILIP SEGAL, THE ART OF FACT INVESTIGATION (2016), Chapter 4, "Databases: Powerful, Quirky, So Often Wrong." Problems include basics such as confusing several people with common names and incomplete coverage ("Nationwide marriage and divorce records" may cover just a few states due to statutory limitations on such information in most jurisdictions).

[41] Don R. Swanson, *Medical Literature as a Potential Source of New Knowledge*, 78 BULL MED. LIBR. ASS'N 29 (1990) (discussed in SAMUEL ARBESMAN, THE HALF LIFE OF FACTS (2012)).

internalize new information. Even if databases do not have the human biases to favor their own research over the research of others, their programmers do not design them to share for commercial and statutory reasons.

To see how far AI still has to go in sophistication of results, a recent study of current, simple databases dealing in a small universe of data (i.e. U.S. case law) is instructive. The study sought through the use of six databases[42] to gauge the extent of human bias in the different algorithms written for each database. The results ("a remarkable testament to the variability of human problem solving"[43]) were striking. After asking each program for a list of the ten leading cases for the concept, "the right to receive information," an average of forty percent of the cases cited were unique to one database, and only seven percent of all 3,000 cases reviewed appeared in all six databases. They ran the search again three years later with broadly different results but no more uniformity than before.

The study explained that the "black boxes" of programming that dictate the rules computers will use to sort information use four main techniques: prioritization (relevance ranking or other rules to give one fact more importance than another); classification; association (marking relationships between entities); and filtering (which excludes some information according to various human-generated rules and criteria).[44]

The number of variables that go into what seems like a straightforward search is a good way to embrace the difficulty of programming computers to think the way we do. Coding a document "relevant" or "not relevant" misses many nuances that can make or break the success of a search.[45]

If the variability of results AI demonstrates is this large in a restricted search over an extremely small data universe, imagine the variability that exists when asking questions such as:

- "Is this person well-regarded in his profession?"
- "Does this person have substantial assets?"
- "Who would help impeach this witness's credibility?"

Of all the four black box techniques that can stall a computerized investigation with large amounts of data, it is one in particular—association—that very quickly brings us face to face with the limitations of our computing power and injects large amounts of uncertainty into the process. Association has also been called by others the problem of connections: "[W]e need to find connections among our data; they are there, all we need to do is to find them. This is where the real trouble starts . . . . [Y]ou attempt to find connections among these data and so begin to examine

[42] Mart, *supra* note 35. The databases used were Lexis Advance, Fastcase, WestlawNext, and Google Scholar.

[43] *Id.* at 390.

[44] The debate about the imported biases of AI outside the law is wide-ranging. *See, e.g.*, CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION (2016).

[45] "For example, according to the study, if the top ten results from one database have only two relevant cases, but those two are the most relevant in that area of the law, that might be a better result for the research than a search that returns eight relevant documents, but misses those top two most relevant cases." Mart, *supra* note 35, at 410.

various *combinations* of these data."[46] Combinations of even trivial amounts of data rapidly become daunting and immediately require shortcuts.[47]

These shortcuts are programmed into software every day, most commonly on Google, which categorizes every document and attempts to correlate a desired search with "relevant" documents. As most of us know, this can work brilliantly or not at all. If Mr. X owns company Y and company Y is not associated with Mr. X in any document scanned and available on the internet, Google may not make the association. If Mr. X and Company Y share an address, it may associate them and may not, but could think Mr. X is an employee and not the owner. Other owners not associated with the address could be missed. The bias here for those who expect Google to make this association is that all relevant documents are available to Google, whereas this is not so.[48]

Broadly speaking, Google and people using it and other forms of AI are taking advantage of a Bayesian form of analysis. On Google, each document must be categorized, organized, and sorted before the program can decide the degree of its probable interconnectedness to other documents. It attempts to "guess" at the most likely useable answers amid an almost infinite set of choices, this guess based in part on prior knowledge. As a searcher learns more, search terms can be refined and the search narrowed to what subsequently seem to be the most fruitful areas of research.[49]

But Baysianism doesn't mean the computers take over and answer a straightforward question. Human input is required on a consistent basis:

> Bayes' Theorem does not supply (and as a *theorem*, should not be expected to supply) prior probabilities from which to construct prior odds. Real-world forensic applications of Bayes' Theorem, in other words, necessarily rest on subjective human judgements of 'prior' probability. Consequently, any resulting inferences of probative value extracted from Bayes nets can only be as good, or bad, as the initial human inputs.[50]

---

[46] DAVID A. SCHUM, THE EVIDENTIAL FOUNDATIONS OF PROBABILISTIC REASONING 491–92 (Evanston, Ill: Northwestern University Press 1994). With just fifty items of data on record, for example, "[y]ou would then have $1.1259(10)^{15}$ possible combinations of data to search through; for 100 items the number of possible combinations is $1.2677(10)^{30}$. . . . The exponential nature of our search problem demands that we apply some imaginative search strategies." *Id.*

[47] *Id.*

[48] SEGAL, *supra* note 40, at 48–50. At the base of the connections problem: Google depends on documents parties wish to be made public, and it is biased by commercial considerations which are not secret but neglected by many users. A "good" search result may be biased in favor of pages that have paid for placement or which others have paid to look at.

[49] For a good introduction to Bayesianism, see SHARON BERTSCH MCGRAYNE, THE THEORY THAT WOULD NOT DIE: HOW BAYES' RULE CRACKS THE ENIGMA CODE, HUNTED DOWN RUSSIAN SUBMARINES & EMERGED TRIUMPHANT FROM TWO CENTURIES OF CONTROVERSY 244 (2011).

[50] Paul Roberts & Colin Aitken, *The Logic of Forensic Proof: Inferential Reasoning in Criminal Evidence and Forensic Science*, U.K. ROYAL STAT. SOC'Y 104 (2014),

The two most difficult things for new investigators to learn are how to weight prior knowledge in setting up a new inquiry (e.g. prior knowledge of where someone has worked dictates where you may look first) and how to incorporate new knowledge into an ongoing search. These are both problems that Bayesian logic squarely deals with. A broader discussion of investigative techniques and the message they give us for how to work with all AI is the subject of Part V.

IV. Big Data for Fact Gathering

In one sense, the spirit of Big Data—harnessing the oceans of information being generated daily—is incompatible with a precise fact-finding mission that traditionally occupies a lawyer. If Big Data helps to identify an area of the country that could benefit from more information about vaccination programs, more advertising about high-end women's running shoes or help-wanted ads for under-represented groups of computer programmers, what good will it do if we need to find a specific person who witnessed a specific action or a single document that will help advance our case?[51]

The promise of Big Data for lawyers is less that of making accurate predictions about case outcomes and more about being able to get leads on the kinds of information they already look for. In some cases, the Big Data will save them some time, and in others it will open up entire new areas where lawyers never thought to look. From an employment perspective, this looks to be positive for lawyers, on the principle that "as more information becomes available, more research needs to be done."[52]

This part will deal with AI applications that even today are likely to produce value for certain attorneys. Not all applications are on the market yet and not all will be affordable for every size of law firm, but given the history of software, the cost of processing and storage, the likelihood is that access to these will rise as costs fall.[53] If law firms decide they want to get control of their data, some of these applications could provide a new kind of data lawyers are not used to looking at, but which could help gather valuable information.

These applications have an additional benefit that other AI does not: they all search publicly available material and do not draw conclusions, such as credit scoring.[54] As such, they are more benign in that they present lawyers with options on which to follow up, but do not themselves recommend any course of action. That control is left in the hands of the lawyers.

---

http://www.rss.org.uk/Images/PDF/influencing-change/rss-inferential-reasoning-criminal-evidence-forensic-science.pdf.

[51] For a more detailed discussion of Big Data and inductive reasoning, see SEGAL, *supra* note 40, at 38–39.

[52] Mart, *supra* note 35, at 393.

[53] Adopting futurist Amy Webb's probable-plausible-possible spectrum of most to least likely to occur, the idea of more data, cheaper data surely falls under "probable." *See* AMY WEBB, THE SIGNALS ARE TALKING 84 (2016).

[54] Discussed in O'NEIL, *supra* note 44.

(i) News Searches for Corporate and Commercial Links

News and media searches serve a critically important fact-investigation function in litigation today. No search of a person's reputation or corporate and commercial links is complete without seeing what media has written about that person. Media "coverage" means not only what journalists write, but also what companies and individuals *wish* to have written about them (through news releases, website publicity and other means). Since company ownership is not always easily linked to beneficial owners—even in the United States—corporate linkage through media coverage is one way to establish associations. News releases are an excellent source of historical research, both to establish factual connections and to gauge what people wish others to see of them, whether or not this provides an accurate picture.

Today, lawyers in the U.S. use LexisNexis, Factiva, and Bloomberg,[55] as well as free news searching over Google. In addition to providing perspective and background on a subject, these have the advantage of preserving the content of news releases that prove to be inconvenient for their issuers in subsequent years and are therefore removed from company websites.

News searches are being improved all the time by companies such as Diffbot (which counts LexisNexis as a client) and iManage (RAVN). Both these companies allow for customized web crawling, and iManage will allow clients to educate the program through trial and error about what stories are of most use, like the way eDiscovery programs operate. It sells a software suite but can also be used via the cloud for smaller clients. However, experience in crawling social media, video and other non-traditional areas of news research is sparse and has only been done on a few high-cost occasions.[56] Law firms that would want to use Diffbot would need to employ programmers or developers to work with Diffbot's application program interface (API). While Diffbot crawls news sites, it is testing a product that can scan discussions (the comment sections following online news stories),[57] which would be invaluable for witness location and other applications that would possibly require an interview to gauge a person's recent sentiments about a given issue.[58]

To the extent that they need to penetrate pay walls, password requirements or statutory barriers, the limitations on the productive use of any web-crawling AI are common to those that restrict the one-stop shopping in our conventional

[55] While these are excellent starting points, they are hardly comprehensive. Even within the U.S., the most glaring omissions to "full news coverage" come when searching foreign-language newspapers published in the U.S., frequently the best source of information about individuals who were born elsewhere. Our firm has recently hired people to read U.S. papers and websites published in Korean and Greek, for example. In such cases, as well as to search thoroughly foreign news sources, the databases need to be supplemented with other research.

[56] Author interview with company, October 2017.

[57] Author interview with company, July 2017.

[58] Of course, an interview would still be required to figure out who would make the best witnesses among a list of prospects, not to mention recommendations for other people to interview whose names do not appear on social media.

databases.[59] There are even potential barriers to the use of free public data: LinkedIn recently sought to challenge the legality of data scraping,[60] though as long as data is not all hidden behind a paywall, the future of these programs appears promising.

In the meantime, there is still much free material not properly searchable with even today's most expensive databases. While news searches are a cornerstone for any fact-finding enterprise, many of the most useful news sources are not indexed on LexisNexis. We often use these non-indexed sources for evidence of personal connections in community newspapers, not to mention controversies regarding a person's job performance as a schools' superintendent, local judge, or other local position of authority.

(ii) Resume Searches for Due Diligence and Witness Identification

One of the best ways to find witnesses or anyone else who can tell us about a person is to ask former employers, employees, or other colleagues.[61] We can profile opposing parties, impeach credibility of witnesses, gain information about attachable assets, and conduct due diligence in a variety of contexts.

In the early days of the Internet many investigators used Monster.com, paying as if they were searching for possible employees, but in reality, they were seeking contact information. For example, contact information of someone who may have worked on a particular trading desk at a specific bank during a six-month window, four years previously. Later on, LinkedIn provided often superior results because it is not restricted to those who have posted résumés, and more people are likely to list prior jobs than to post resumes for public consumption: A "happy" employee who posts a resume on Monster.com could draw attention from her supervisors, but millions of genuinely happy professionals with no thought of leaving their jobs use LinkedIn.

The current generation of AI offers services that scrape LinkedIn, saving fact investigators time in manually searching pages and running new searches every

[59] Discussed in Part III, *supra*.

[60] hiQ Labs, Inc. v. LinkedIn Corp., No. 17–cv–03301–EMC, 2017 WL 3473663 (N.D. Cal. June 07, 2017). The court has granted a preliminary injunction to prevent LinkedIn from disabling scrapers on LinkedIn data that is available to the public. LinkedIn had alleged that hiQ violated the federal Computer Fraud and Abuse Act of 1986. 18 U.S.C. § 1030 (2008). At the time of writing the case is on appeal to the Ninth Circuit. The District Court held that LinkedIn's terms of service (allegedly violated by hiQ Labs) did not apply because hiQ only accessed publicly available data. LinkedIn could, of course, decide to protect all its data behind a paywall, but then would lose the benefit of appearing in Google search results, presumably an important ingredient in driving traffic to its site and increasing profitability.

[61] Most commonly in the context of litigation, our firm sticks with former colleagues so as not to fall afoul of ABA Model Rule of Professional Conduct 4.2 nomenclature which prohibits contact with represented parties outside the presence of their lawyer. Even non-parties who are senior enough in a company can fall within this rule. Most states have adopted Rule 4.2 and some impose even stronger restrictions on the classes of people who may be contacted.

few days, in the event that someone updates their profile and emerges on a list from which they were previously excluded.

The next generation in this area is typified by a company called Entelo,[62] which accumulates social media the way Kayak.com simultaneously searches travel sites. Entelo searches some 50 sources of public information from LinkedIn, Github, Twitter, and can even find personal emails that are not password-protected. It also employs an algorithm that predicts when employees may be thinking about moving, based on work anniversary, recent résumé updating, or the corporate acquisition of their employer.

Due diligence also presents great opportunities to employ AI's excellent pattern-recognition abilities. If in pre-deal due diligence we are conditioned to sit up and pay attention when a person has business in a disreputable tax haven such as the Cook Islands, we are less attuned to conformity with harder-to-spot patterns: The amount of litigation in which someone appears as a defendant could be a problem, but could be less of a problem if that person is named along with every director. Some suits of merit follow this pattern, but many are nuisance suits that carry little weight in serious due diligence.

(iii) Material Formerly Left Untranscribed

Perhaps the biggest change right now for fact finding in the unlimited universe is the coming capacity to search and analyze spoken data, which is now being churned out daily at ever growing rates. While some people may prefer to write their thoughts down, many prefer the spoken word.[63] How many lawyers today (when doing news searches) conduct searches of podcasts?[64] As of four years ago there were 250,000 unique podcasts in the world in over 100 languages. Such a search can now be done with Spotify, Google, and perhaps soon Apple, which purchased podcast transcriber Pop Up Archive late last year,[65] conferring an ability to search more deeply than would be feasible with manual review for subject matter.[66]

---

[62] ENTELO, www.entelo.com (last visited Apr. 24, 2018), and Author interview with company, Aug. 24, 2017.

[63] AngelList lists 103 speech recognition startup companies. ANGELLIST, https://angel.co/speech-recognition (last visited Apr. 24, 2018).

[64] Lex Friedman, *Apple: One billion iTunes podcast subscriptions and counting*, MACWORLD (July 22, 2013, 2:36 PM), https://www.macworld.com/article/2044958/apple-one-billion-itunes-podcast-subscriptions-and-counting.html.

[65] *See* Brian Heater, *Apple buys podcast search startup Pop Up Archive*, TECHCRUNCH (Dec. 25, 2017), https://beta.techcrunch.com/2017/12/05/apple-buys-podcast-search-startup-pop-up-archive/.

[66] Here, as in most discussions about American artificial intelligence, search capabilities in English are superior to those in most other languages. For languages not widely spoken, AI sorely lags. An easy example is Google Translate, which offers extensive translation and verbal renditions of those translations in major European and Asian languages, but relegates Albanian and even Swahili (spoken by more than 50 million people) to mono-tonal robot transliteration. Basque and Burmese don't even get the robot pronunciation—Google just doesn't bother to program those translations to speak. When it comes to LexisNexis and media searches, a search of "All Non-English Language

The ability to capture and organize speech-based data seems to present greater challenges than organizing written material, in that there is less categorization of this spoken data by those who generate and share it. A podcast about the life of Muhammad Ali would contain Ali's name in the title or sub-title, but a person's discussion of Ali could come in any number of contexts: boxing, the Vietnam war, the Nation of Islam, Parkinson's Disease, Kentucky, or something else. The challenge seems worth meeting: the uses of searching and analyzing the spoken word are nearly limitless, complementing and in some cases exceeding the utility of organizing the written word.

Just as podcasts are searchable, how long will it be before a comprehensive search of all YouTube material will be available—beyond the inadequate Google search of today? At least one billion hours of YouTube material is viewed each day,[67] a stream of data far in excess of what it is reasonable to expect to review manually. Today, the entire transcript of a news conference by a company is often not offered to the public. Journalists choose a particular quotation or two to fit the story they are writing that day, whereas some remark may prove to be prescient or, in retrospect, highly informative even if according to the news cycle of that day it is judged to be of little value.

What would happen if the news conference were on YouTube and searchable automatically? A company in Finland called Valossa already transcribes and analyzes any video a client submits.[68] For now, the company will not ordinarily monitor a particular YouTube channel because YouTube detects web bots and disallows the practice,[69] but this could be subject to negotiation and a fee. Given that YouTube is free to use whereas conventional databases can cost hundreds or thousands of dollars per month, this seems at least of plausible value and a potential source of revenue for YouTube to supplement its advertising.

What Valossa and other applications have in common is that they are organizing data that (until very recently) we never thought was possible to look at in any systematic way. Similar options for lawyers are many and seem set to increase. For example, a company called Clarifai generates keywords from videos and photographs on the internet.[70] Another company, Veritone, detects sentiment, objects and faces in videos. One example it offers is after analyzing days of security camera footage, it can determine when two particular individuals were seen together and can find all recorded phone calls when a certain employee was mentioned in a negative statement. Several companies including Affectiva[71] can

---

News" in fact searches the following short list: Danish, Dutch, Finnish, French, German, Italian, Malay, Norwegian, Polish, Portuguese, Russian, Spanish, Swedish, and Turkish. There is also a listing for "other," but in the author's experience anyone searching Greek or Korean publications, among others, is out of luck at LexisNexis.

[67] *You know what's cool? A billion hours*, YOUTUBE: OFFICIAL BLOG (Feb. 27, 2017), https://youtube.googleblog.com/2017/02/you-know-whats-cool-billion-hours.html.

[68] VALOSSA, www.valossa.com (last visited Apr. 24, 2018). "Valossa" means "in the light" in Finnish.

[69] Author interview with Valossa, Sept. 4, 2017.

[70] CLARIFAI, www.clarifai.com/ (last visited Apr. 24 2018).

[71] AFFECTIVA, www.affectiva.com (last visited Apr. 24, 2018).

use tone of voice and facial express to recognize joy, surprise, and anger in focus groups.[72]

Of course, just because the technology exists does not mean that the applications are always suited to monitoring individuals (as opposed to groups suitable for marketing drives). Affinio,[73] which tracks and organizes consumer sentiment, would (for now) be unable to determine, for example, which individual consumers may be complaining about defective knee replacements (in an effort to gather more plaintiffs for a class action). However, the company could determine which parts of the country contain the most social media discussions about "knee surgery," which could allow better targeting of lawyer ads to find plaintiffs.[74] At $60,000 per year this kind of service would only be of interest to the largest class action firms, especially because the company's data is only kept for 30 days. But if one thing seems eternally true in technology it is that storage and overall costs drop, competitors enter the market, and "AI" turns into "software."

(iv) New Types of Data: Internet of Things (IoT) and Drones

If the new sources of data above stem from traditional sources (news, resumes, and people speaking to one another), the more radical departure comes in the form of data generated by the objects with which we come into contact: the objects we own and the objects others own but use to observe us (drones).

The network of things we own is better known as the Internet of Things (IoT), and is a broad category that encompasses data generated by objects. We already have commercially-available databases that photograph the location of parked, privately-owned vehicles, and Google Earth can tell you that the "vacant lot" in a newspaper story from last year has been filled in with a new house. But this information is not generated automatically by the objects themselves. Google Earth and the databases can have time lags of months or years between updates, and where cars are parked is difficult to chart if they park on private property or in an underground garage.

As with drones, the novelty of the kinds of information that will be available seems to be an obvious new area of specialization for lawyers. Many, and probably most of us, would resent the reselling of photographs of the inside of our refrigerators, not to mention video streaming of activity in our backyards courtesy of a drone patrolling our neighborhoods. Data reporting how often our car is parked behind a bar (for handy use by underwriters) is not something we would all want to be available.[75]

---

[72] Erik Brynjolfsson & Andrew McAfee, *The Business of Artificial Intelligence*, HARV. BUS. REV. (July 2017), https://hbr.org/cover-story/2017/07/the-business-of-artificial-intelligence.

[73] AFFINIO, www.affinio.com (last visited Apr. 24, 2018).

[74] Interview by Author with Affinio, July 27, 2017.

[75] Of course, this data is available today on an expensive, case-by-case basis. Anyone can hire a licensed investigator to follow anyone else, but this is not seen to be a major social problem mostly because of the expense of doing so and also because investigators usually require a license. At $80 to $100, few have the means to follow anyone else for very long. GPS trackers already require a warrant for the government to use them (United

However, the IoT and drones open the door for data interpretation using an emerging mathematical technique known as differential privacy. This refers to the ability to learn a lot about a group while keeping the identities of the individuals within the group anonymous, by inserting "mathematical noise" into a user's pattern of behavior, according to Apple's iOS 10 literature.[76] Differential privacy is designed to improve on previous data sets that anonymized users but still proved to be vulnerable to finding out individual user identities.

This does not mean that differential privacy creates privacy where none existed before, but may offer enough comfort for people to free their objects' data to participate in surveys and happily use iOS10, both of which will generate more data. Will consumers be able to opt in to surveys generated on the IoT? Will they be required to opt out? All of this presents more opportunities for lawyers no longer required to perform document review. For fact investigators, differentiated privacy could allow much larger and more varied sets of data with which to conduct fact-finding research.

(v) Quantum Computing

Very much in the category of the possible but not the near-term probable, this advancement in technology is probably more than ten years away, according to Gartner Inc.[77] Nonetheless, law firms thinking longer term should not assume it will never arrive. Unlike today's computers that encode information in bits (with each bit holding a value of one or zero), quantum computing's qubits can hold the values one and zero at the same time. The result could be a million-fold increase in computing speed.

All of the processes described so far in this article would happen this much faster with quantum computing, and limits of computing time would eventually fall away. The change would be as momentous as the first move to computers from the world of typewriters and calculators.

Yet if all that quantum computing did was to make computing faster, this development would not bring human lawyering to an end even if every word spoken was recorded and structured into searchable data. Not all legal reasoning is deductive, and logical deduction is the basis for today's computational logic. Michael Genesereth of Codex, Stanford's Center for Legal Informatics wrote

States v. Jones, 132 S. Ct. 945, 949 (2012)), and state legislators are rapidly imposing various restrictions on private placement of such devices.

[76] Quoted in Andy Greenberg, *Apple's 'Differential Privacy' is About Collecting Your Data—But Not Your Data*, WIRED (June 13, 2016, 7:02 PM), https://www.wired.com/2016/06/apples-differential-privacy-collecting-data/. For a fuller academic explanation of differential privacy, see Cynthia Dwork & Aaron Roth, *The Algorithmic Foundations of Differential Privacy*, THE ESSENCE OF KNOWLEDGE (2014), https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf.

[77] *2017 Gartner Hype Cycle for Emerging Technologies: AI, AR/VR, Digital Platforms*, WHAT'S THE BIG DATA? (Aug. 16, 2017), https://whatsthebigdata.com/2017/08/16/2017-gartner-hype-cycle-for-emerging-technologies-ai-arvr-digital-platforms/.

recently that computational law "simply cannot be applied in cases requiring analogical . . . reasoning." [78]

We can imagine a world in which computers think abductively like fact finders, using defeasible reasoning as they pick their way toward the most likely satisfactory conclusion. The problem is that for now there is no evidence that computers are anywhere close to being able to do this. Faster computers do not equate to smarter computers. To quote the Gartner Hype Cycle[79] we may be at a "Peak of Inflated Expectations" as to what artificial intelligence will do to the practice of law. Next in most cycles comes the "Trough of Disillusionment," when the most far-reaching projections prove difficult to implement, followed by a "Slope of Enlightenment" when people figure out how to use the technology correctly.

The final part of this Article is aimed at easing users of AI for legal purposes into a much less extreme period of disillusionment, and to a faster path toward enlightenment later on. The main technique is the one currently in use by anyone who gets the most out of today's technology: suppressing the urge to take data outputs as the final product, and the grooming of creative thinkers to work with the software.

It seems certain that they will need to continue working with the software. In the words of MIT's Erik Brynjolfsson, "AI won't be able to replace most jobs anytime soon. But in almost every industry, people using AI are starting to replace people who don't use AI, and that trend will only accelerate."[80]

V. The Person to Manage it All

What kind of person is best for using AI? My belief for years based on experience has been that being good at practicing law does not necessarily mean a person will be good at investigation in the unlimited universe of facts. Law schools in the U.S. are nearly uniform in their omission of fact investigation as a subject in their curricula. They teach the basics of discovery, but not fact gathering prior to filing a case or during discovery when the lawyer presumes she is being lied to by her opponent's client.[81]

Management professor Ed Hess argues that in the age of artificial intelligence, being smart won't mean the same thing as it does today. Because smart machines can process, store, and recall information faster than any person,

---

[78] MICHAEL GENESERETH, COMPUTATIONAL LAW: THE COP IN THE BACKSEAT, WHITE PAPER, CODEX: THE STAN. CTR. FOR LEGAL INFORMATICS (2015), http://complaw.stanford.edu/.

[79] *2017 Gartner Hype Cycle for Emerging Technologies: AI, AR/VR, Digital Platforms*, *supra* note 77.

[80] Brynjolfsson, *supra* note 72.

[81] Some law schools such as Vanderbilt, Northwestern, and Harvard are now offering concentrations or programs in law and innovation, and while I have long held that creativity and innovation are the keys to good investigative practice, these programs appear to be more geared toward being good lawyers for representing innovators. Any extra help in getting lawyers to think about facts in a manner other than logical deduction is welcome.

the skills of memorizing and recall (widely rewarded on exams) are not as important as they once were. The new smart "will be determined not by what or how you know but by the quality of your thinking, listening, relating, collaborating, and learning."[82]

Among the many things this will mean for lawyers are two long-true aspects of fact investigation that will have to inform the working day of more attorneys:

- Open-mindedness is indispensable. You can never see all the data, and sometimes there is no data to be had.
- Logical deduction is out, logical inference is in.

(i) Dealing with Absent Data

A good deal of any fact investigator's time today is sorting through mountains of conflicting or inaccurate database output in order to decide which leads from public records to confirm and act upon, and which ones to scratch. But even if the databases were flawless, lawyers would still miss some of the most relevant pieces of information about people because much of the good information is not only unavailable on the internet, it is *not written down at all*.

As an exercise, I often ask people to Google themselves. At most we come up with two or three percent of what we know about ourselves. All our friends? Viewpoints? Former colleagues, employers, spouses and significant others? What we like to read? If Google can only do a three percent job on you, why should it do any better on anyone else you are examining? Along with other sources, Google provides a very useful starting point, but not the end of the inquiry.

The oftentimes insufficient amount of *useable* data means that a lot of AI will be of little use *by itself* once it generates (still valuable) leads on which way to proceed. The director of data science at Microsoft says that many organizations can overestimate the capabilities of machine learning because they have insufficient data with which to build patterns needed to make good predictions.[83] Just as Google Translate needs lots of examples of old translations before trying a new translation, AI in other areas cannot predict patterns without similar past patterns fed into a computer's memory.

This kind of learning is known as induction or inductive generalization, and carries drawbacks for fact investigators: "not every law firm will have a stream of cases that are sufficiently similar to one another such that past case data that has

[82] Ed Hess, *In the AI Age, "Being Smart" Will Mean Something Completely Different*, HARV. BUS. REV. (June 19, 2017), https://hbr.org/2017/06/in-the-ai-age-being-smart-will-mean-something-completely-different. For full treatment, see EDWARD D HESS & KATHERINE LUDWIG, HUMILITY IS THE NEW SMART: RETHINKING HUMAN EXCELLENCE IN THE SMART MACHINE AGE (Berrett-Koehler 2017).

[83] Sam Ransbotham, et al., *Reshaping Business with Artificial Intelligence*, MIT SLOAN MGMT. REV. (Sept. 6, 2017), http://sloanreview.mit.edu/projects/reshaping-business-with-artificial-intelligence/.

been catalogued contain elements that will be useful to predicting future outcomes."[84]

This is unlikely to change. Imagine that in five years from now, a lawyer needs a witness who talks negatively about a company her firm is suing. She will not need people to listen to endless videos on YouTube because computers will search and transcribe those videos for her. But she will require people to analyze the results of possible witnesses these programs flag. Are the facts complained about in the videos analogous to those in her matter?

Then, she will need sensitive people to interview the short list to see if they would in fact make persuasive witnesses. Part of the job will be cross-examining them as to their stories and conducting the same kind of due diligence that opposing counsel will. Is the information on social media picked up by the AI accurate? In context? Emotionally compelling in the right way for the matter at hand? One day, in the era of quantum computing, a machine may be able to conduct an entire conversation with a human being and make all of the judgments about tone, motive, truthfulness and other factors we use to make a decision, but in the meantime, this is still the job of a person.

(ii) The Need to Grasp Logical Inference

The kind of imagination that investigators use when they try to form an explanatory hypothesis for something they observe is called abductive reasoning, in which conclusions *may*—but need not always—follow from a premise.[85] Without on-site surveillance, it is not logically certain that Person A lives in House X. Databases may (and often do) say that Person A along with Persons B, C, D, and E live in House X, even though Persons B through E belong to the same family, two of whose members bought the house from Person A. Person A may still live there as a tenant, but that is unlikely. Deductive logic does not enter into the matter.

Instead, a fact finder needs to weigh this database's output with the output of other sources of information. If they all suggest that Person A still lives there and have no other address for him, the chances of tenancy increase. But if they note that Person A bought House Y on the other side of the country the same day he sold House X, and that he is supposed to be living in Houses X and Y simultaneously, most of us would proceed with the working assumption that he

---

[84] Surden, *supra* note 4, who also notes that "[t]here are other well-known problems with induction. Induction relies upon analyzing examples from the past to generalize about the future. However, under the so-called "Black Swan" problem, there may be never-before-seen, but salient scenarios that may arise in the future. In such an instance, a model trained upon past data may be insufficiently robust to handle rare or unforeseen future scenarios." *See, e.g.*, NASSIM NICHOLAS TALEB, THE BLACK SWAN: THE IMPACT OF THE HIGHLY IMPROBABLE 9–11 (2d ed. 2010).

[85] The credit for identifying and naming the process of abduction goes to an American philosopher, Charles Sanders Peirce (1839–1914). What is most striking about this is that abduction appears to have been largely overlooked and under-analyzed by almost 2,400 years of formal logic and philosophy. As Peirce wrote, abduction is the only logical operation which introduces any new idea. While deduction proves that something *must* be, abduction merely suggests that something *may be*.

lives in House Y. The database would have information that indicates residence at House X, but it could be nothing more than an unchanged magazine subscription or remnants from an outdated mailing list.

The most common, easily understood example of abductive reasoning in recent years related to the accounting firm used by convicted fraudster Bernard Madoff.[86] Among the many signs that Madoff was deceiving investors was his use of a tiny strip-mall accounting firm in the suburbs of New York. Firms with the total investments of the size Madoff claimed to be running would have used a "Big Four" accounting firm, or perhaps one of the larger second-tier firms known across the country. There is no logical deduction to tell you this, however. It requires a person to remark, "If I had a firm that big, how could I get business done with a tiny accounting firm? Either I've got the name of Madoff's firm wrong, or something is not right here."

It sounds like child's play in retrospect, but many sophisticated investors lost billions of dollars because they neglected to put themselves in Madoff's shoes. Even if Madoff had used a larger but still-too-small firm, the unusual accounting decision should have put investors on alert to conduct more inquiries into other aspects of his firm (such as, in the final two years before he was caught, the paucity of securities filings that would accompany holdings worth tens of billions of dollars).

The opposite problem of absent data is this: so much data that data scientists don't know where to look first. Again, AI can provide indications, but no firm answers. As discussed in Part III, a Bayesian approach to investigation is a critical element of any investigation when the possibilities are, in effect, unlimited. Madoff presented any number of possibilities, from the extreme of being a Ponzi-schemer, to a legitimate investor who embezzled, to a real investor with separate legal problems reflecting on his character. Only investigation would tell you which, if any, was relevant.

Searches across the country for witnesses or asset identification worldwide encounter what feels like an ocean of possible places to begin looking. When air-crash investigators look over tens of thousands of square miles of real ocean for an aircraft's "black box," it is not a matter of "if this, then that" deductive logic to find the most likely spot. Investigators use the laws of physics, known weather at the time of the crash, subsequent wind and current information to decide on the *likeliest* areas on which to concentrate their search. If they assign a ten percent probability that the box is in one portion of the ocean, the box could still turn up there. If it does, they probably will discover it later rather than sooner, after using subsequently-derived information to alter their probabilities and shift their search to what was originally the least likely spot.[87]

---

[86] *In re* Bernard L. Madoff Inv. Sec. LLC, No. 08-99000, 2015 WL 4734749, at *9 (Bankr. S.D.N.Y. Aug. 11, 2015).

[87] This example is not such a distant relative to what lawyers face all the time. Suppose you are asked to profile a person who has lived and worked in fifteen different jurisdictions over the past twenty years. If you cannot afford the time and money to search courthouses for litigation in all of these, you need a hypothesis about the ones most likely to yield information. So, you search in the five that account for twelve of the

So too, in legal investigation. A search for a defendant's assets so that a client may collect on an $80 million judgment could potentially involve every bank in the world and every luxury real estate development. But when beginning such a search, we make assumptions about where we should start, and this is where AI can help: in quickly letting us know where such a person may have left a "footprint." Social media can fill in blanks left by incomplete or obscure public records, but deductive logic does *not* tell us to stop looking in the most likely spot and to move on. We are left only with revised probabilities of where to look. These probabilities can be generated with the help of AI, but analogical reasoning only gets us so far. Pattern matching of prior similar behavior results in a list of probabilities, but at what point can a computer say, "This is the answer. Investigation terminated"? For now, there is no way for a computer to make that kind of judgment. [88]

(iii) The Backbone of an AI Training Program

If the trend continues that users of AI replace non-users, what should a training program for those AI cover? Such a program should probably be considered for anyone who will use AI and will be called upon to make dozens of decisions about how to operate the system and how to ask it follow-up questions. Just because someone who arrives at a firm with a law degree knows what Google is does not mean that person is an expert searcher. Those just getting by with a knowledge of Westlaw and social media are in danger of being swamped by the waves of data already becoming available to anyone with a computer or smartphone.

The person who should *not* be leading this portion of the training is the sales representative of the software vendor. While useful contributors, these people enter with pre-set searches designed to show their products in the best possible light. And while sales staff may have subject-matter expertise in law (it helps if they are lawyers themselves) they (as in *any* lawyer) cannot possibly specialize in all facets of the law. Their job is to sell and they should not be criticized for it. But at the very least, buyers of AI products ought to design their own test jobs, both before purchase and during the after-purchase training of lawyers who will be using the product.

---

person's fifteen residences and workplaces in the past twenty years. If he was arrested for assault while on vacation or in a county adjacent to the one in which he worked, you may miss it. If you happen to learn about a vacation arrest late in the investigation through an interview, you need to go to a courthouse that was not on your original list. If you find out that your man changed his name years ago, you need to go back and search again under his older name.

[88] This question is at the heart of what computer scientist Alan Turing (1912–1954), leader of the team that cracked Germany's Enigma code during the Second World War, called the "halting problem." *See also* Jeffrey M. Lipshaw, *Halting, Intuition, Heuristics, and Action: Alan Turing and the Theoretical Constraints on AI-Lawyering*, 5 SAVANNAH L. REV. 133 (2018).

(a) Control Group and Search Training

There is no substitute for usage to measure the effectiveness of any software product. The "Google yourself" exercise above is a good start, but lawyers should conduct a variation on that test with every database they use. A person searching his own data is his own best control experiment, but if dealing with a database in which a person may not appear (such as a database of business records), then a different control group is required. Subjects could be the lawyer's own firm, the client in the matter, or a company well known to the lawyer. How much "easy" or "obvious" information can the database retrieve quickly?

For databases requiring the entry of a search string, novices often fail to see the delicate trial and error involved in teasing out the best results. Word search order and time of day are relevant factors in a Google search and other databases all have their quirks and shortcuts as well. All databases respond to altered lengths of keywords, or the failure to anticipate the coding decisions of those who set up the database. Will civil fraud show up in a search for criminal fraud? If not, will it be referenced if a fraud is the subject of both civil and criminal litigation? Is the search too broad? Too narrow? A control group needs to be monitored closely against the urge to enter search terms that should only be known after the fact.

Keyword searching is a cornerstone skill for any investigator. In a case handled by our firm, a client's paralegal spent two days attempting to link two individuals suspected of being in the same ethnic group and possibly related. The conclusion of the paralegal was that they were not related and of different national backgrounds. Our firm, with nothing more than a news search, established that they were brothers. The difference between our news searching and that of our client was that entering the two men's names into news databases produced too many results (since they both had common names). Entering one name with the man's company produced no result since the company had not been written about in national media. Entering one man's name with his city of residence produced no hits. But entering his name with his former city of residence produced a four-year old obituary of his father, and linked him to his father's other son—his brother. The searching took hours of patient trial and error.

(b) Relative AI Product Strengths and Weaknesses

AI is a tool, and, like all tools, it is more useful for some things than others. In the current generation of software, the different databases excel at different jobs, but lawyers do not learn this from sales representatives. The knowledge comes from usage.

Westlaw is particularly good at linking individuals with private companies—not even close to infallibly, but better than its competition. However, Westlaw lags others in current real estate ownership or criminal litigation. Transunion's TLOxp® database is probably the best for a (far from comprehensive but still better than the rest) look at criminal litigation. TLO has the most current telephone numbers but still makes mistakes and can be stumped fairly often. This is all subject to the occasional exception and to change in software capability and

new competitors. But it serves as a useful reminder that AI's user experience is something to be discovered before and after purchase.

Again, no sales or training representative can be counted on to mention the great weaknesses of any AI product. That is an internal training task.

(c) Database Bias

Databases of today illustrate all the foibles of the databases of tomorrow's AI, for the simple reason that both today's "software" and tomorrow's AI have one crucial factor in common: they are designed by humans, and humans code and populate the content. The literature is growing quickly on the topic of human biases[89] and their influence on artificial intelligence, and mostly this has to do with biases about race, sex, income, and other "hot button" issues.

There is another kind of bias that will always be with us when we use machines to translate human language into a question and then ask that machine to turn its answers back into a language spoken by a person: what is commonly known as the "garbage-in-garbage-out" problem. In some cases, it may be less a question of bias and more one of human error.

- Incorrectly spelled or ordered names (for example, transposing first and last names) yield incorrect database outputs as Wang Li becomes, after data entry, Li Wang. Good investigators anticipate such errors and search on the assumption that the database is wrong, not correct.
- "Harmless" coding decisions that require more knowledge of the user than if the coding worked with great foresight.

A simple example to illustrate the need to work *with* AI as opposed to having AI *work for us* was an exercise I used to give my fact investigation students: "who owns the building we are standing in?" Some progressed enough to be able to find the public records repository on line (many relied on web-based applications with no obvious measure of reliability but which scored high with Google's search algorithm). However, very few got the public record that could answer the question, because the building was a condominium with separately-owned floors. Without knowledge of the lot number of our floor, there was no *logical* way to look up the publicly-recorded deed. The database could easily have defaulted to a message to users that in such a case, "Please look for the condominium offering document for a full list of lot numbers." The programmers did not bother, and it was up to the user to do the thinking himself.[90]

---

[89] *See, e.g.*, Mart, *supra* note 35.

[90] The puzzle was not hard to crack. In the offering plan for the condominium was, in fact, a list of all lot numbers, but students needed to think the problem through and did not. Failing that, low-cost experimentation would have helped. Entering the number 1 for lot number brought up a document with all of the lot numbers, but no student thought to try this costless experiment. One conclusion I drew from this was that many people who go into law are not by nature experimenters. *See infra* section (iv).

(iv) Personality Types for Fact Finders

Not everyone has a temperament suited to dealing with great amounts of variability. Even those who think the sciences are a place to run lots of repeated, well thought out experiments to come up with "the" inductively-produced answer should be prepared for unpredictability. Mathematician John W. Tukey famously remarked: "economists are not expected to give identical advice in congressional committees. Engineers are not expected to design identical bridges – or aircraft. Why should statisticians be expected to reach identical results from examinations of the same set of data?"[91]

What kind of person is valuable not by increasing how much they know, but in dealing with what they don't know? A still widely quoted study of attorney personality types from the ABA's Journal more than 20 years ago[92] indicates that not just any lawyer is a "natural" to be a fact-finder, especially because of the kinds of people who choose law over other careers. The article examines findings of lawyers who took the Myers-Briggs test on personality type.

One of the four measures of personality type in this methodology is Extravert vs Introvert. The former are interested in people, places, and events, and "prefer to focus their awareness and obtain their mental stimulation primarily from the world around them." This would seem to be a good attribute for someone tasked with fact-finding, versus the introverts who prefer to obtain their mental stimulation primarily from within. Three-quarters of the general population in the U.S. is primarily extravert, but just 43% of lawyers in the study were extraverts.[93]

More significantly for fact finders in the unlimited universe is the division between Judgers and Perceivers: Judging has nothing to do with being judgmental but refers to how one comes to a conclusion. This pair of traits describes how we deal with people and information. Judgers are planned, decisive, and orderly. Perceivers are flexible, open, and spontaneous. Fifty-five percent of the general population are judgers, but 63% of lawyers are. That means that the critical trait in an investigator to keep an open mind as we pursue multiple possible routes in inquiry is something that is disfavored by nearly two-thirds of lawyers.

Does this mean that if fact-finders are not perceivers that they will do a bad job? Not necessarily. It only means that if one recognizes the need to improve perceiving skills and to second-guess that need to come to a conclusion too early, a lawyer may pick up on a clue or a stray fact that he would have ignored before. As a first-year investigator fresh out of law school, my job review was overall very good, but the one critique was that I was sometimes too keen to run with an initial theory. In the end I could be moved off the theory, but my supervisors wanted to see more flexibility—in other words, judge a little less and perceive a little more.

Most lawyers could probably benefit from that sort of feedback, and lots of careful training as well, as they confront the software of the future. In the words

[91] *See* MCGRAYNE, *supra* note 49, at 169.
[92] Larry Richard, *The Lawyer Types*, A.B.A. J. (July 1993).
[93] *Id.* at 75.

of Seyfarth Shaw Chair Emeritus J. Stephen Poor, in the age of AI "people are more important than ever."[94]

[94] Remarks at conference, *Knowledge Management in the Legal Profession*, New York, Oct. 18, 2017.