The Uncertain Promise of Predictive Coding

Dana Remus, University of North Carolina at Chapel Hill
Predictive coding encompasses a variety of technology-assisted approaches to discovery that employ machine learning in document review. Proponents contend that these new technologies can solve not only the practical problems of enormous and unmanageable datasets, but also deep-seated problems of excess, abuse, and trust that have long characterized civil discovery practice. Predictive coding can do so, proponents claim, by replacing the unreliable professional discretion of lawyers with the mechanized objectivity of computers. Encouraged by the official endorsement of a handful of trial judges in the last eighteen months, the litigation community is embracing this view of predictive coding as an easy technological fix for the problems of discovery. I argue in this paper that the profession’s unqualified embrace of the new technology is problematic for a number of reasons. First, by ignoring outstanding and contested issues of design and use, the profession is blinding itself to significant variation in the functionality and efficacy of predictive coding technologies. Second, by deferring to the opinions of computer scientists, vendors, and other non-lawyers, and by delegating work to computers, the profession is compromising the scope of its jurisdiction. Finally, by altering ethical rules to facilitate predictive coding’s adoption and use, the profession is threatening the protections and legitimacy of our adversarial system. I argue that the profession has an ethical obligation to tend to the tools of its trade—to explore the costs, as well as the benefits, of predictive coding, and to play a more active role in its design and use.
INTRODUCTION

Computers cannot practice law. Increasingly, however, machine learning is performing legal tasks and displacing lawyers in discovery practice. A growing chorus of judges, lawyers, and software vendors are advancing “predictive coding” as a means of restoring trust in the litigation process. Predictive coding, they claim, can solve long-standing problems of discovery abuse by replacing the subjectivity and discretion of human review with the mechanical objectivity of computer technology.1

Predictive coding encompasses a great variety of underlying technologies and implementing procedures.2 Generally speaking, it refers to technology-assisted approaches to discovery that employ machine learn-

---

1 Cf. THEODORE PORTER, TRUST IN NUMBERS, 90 (1995) (observing that “[m]echanical objectivity serves as an alternative to personal trust,” in contexts where “subjective discretion has become suspect”).

Predictive Coding

Under a typical protocol, one or more senior lawyers code a sample set of documents for responsiveness and privilege. Based on the initial coding, a computer generates a search algorithm for identifying responsive and privileged documents. The computer uses the algorithm to code another sample set of documents for responsiveness and privilege, which the lawyers then review and correct. This iterative machine learning process continues until the lawyers, often in consultation with the software vendor, are satisfied that the computer’s algorithm will adequately identify responsive and privileged documents. The computer then uses the algorithm to code the entire dataset—sometimes containing millions of documents—for responsiveness and privilege. Predictive coding therefore replaces teams of lawyers performing document review with computers.

Proponents of the new technology hope that computers can succeed where lawyers and judges have failed—in resolving the deep-seated problems of excess, abuse, and trust that have long characterized discovery practice. In recent years, reform proposals and scholarly commentary have focused nearly exclusively on e-discovery’s problems of scope and cost. Unquestionably, these problems are central to the crisis facing civil discovery today, but they are layered on top of deeper-seated and longer-standing problems of gamesmanship and adversarial excess. For decades, lawyers have approached discovery strategically, as “the critical battlefield on which the war is waged.” As a result, discovery has long been a principal source of delay in the courts, and a principle cause of public distrust in the judicial system and the legal profession.

---


4 For an overview of the functioning of predictive coding procedures generally, see Predictive Coding and Patented Workflow: A Defensible E-Discovery System, Interview with Howard Sklar, Senior Counsel, Recommind, Inc., 4/1/12 METRO. COMP. COUNS. 16, 2012 WLNR 8077832, p.4 (April 1, 2012) [hereinafter “Predictive Coding and Patented Workflow”].


7 Id.

8 Louis Harris & Assocs., Judges Opinions on Procedural Issues: A Survey of State and Federal Trial Judges Who Spend at Least Half Their Time on General Civil Cases, 69 B.U. L. Rev. 731, 733 (1989) (reporting that of 1000 surveyed federal and state judges, many believed that discovery abuse was “the most important cause of delays in litigation and of excessive costs”); Beisner, supra note 3, at 549.
Against this backdrop, commentators in the early 2000s began advocating predictive coding as a new solution to old and intractable problems of discovery practice. For at least a decade prior, lawyers had been looking to computers for targeted solutions to the problems of scope that computers had created in the first place. They used document management systems to delete duplicates and consolidate email strings. They used key-word searching to cull through millions of documents and produce more manageable volumes for human review. In looking to predictive coding, in contrast, reform-oriented commentators were looking for solutions to problems of excess and abuse rooted in lawyer discretion, as well as problems of volume and scope stemming from computer technology.

Proponents claimed that by eliminating the time and unreliability of human document review, predictive coding could increase accuracy while decreasing costs. By replacing human discretion with mechanized objectivity, it could eliminate abuse and restore trust in the discovery system. An absence of persuasive support for these claims did not detract from their appeal to the litigation community. After a handful of trial courts officially endorsed predictive coding in 2012, judges and lawyers embraced the new technologies as a promising means of meeting discovery obligations. Today, predictive coding is the “hot topic” of discovery reform. Rule-makers are considering amendments to the Federal Rules of Civil Procedure to encourage adoption and use, and countless conferences and CLE classes are extolling its virtues.

Scholars, meanwhile, have largely overlooked predictive coding’s existence and significance. A handful of recent articles trace the new technology’s emergence and growing popularity, but they do not turn a

---

9 See infra notes 56-65 and accompanying text.
10 See infra notes 63-65 and accompanying text.
11 As discussed below, two empirical pilot studies published in 2010 favorably compared predictive coding approaches to traditional manual review. See infra notes 60-62 and accompanying text. But limitations on the methodology and scope of these studies rendered them incapable of supporting proponents’ broad claims of increased accuracy and decreased costs. See infra notes 90-91 and accompanying text.
15 See infra note 82 and accompanying text.
critical eye on the causes and consequences of its adoption and use. They elide significant variation in the underlying technologies and implementing protocols, and they fail to recognize significant ethical trade-offs entailed in this new approach to discovery practice.

In this Article, I argue that although predictive coding technologies have unquestionable potential as a means of addressing unmanageable datasets, the profession’s unqualified embrace threatens unforeseen and dangerous consequences. Among other things, technological ignorance may disable lawyers from providing strong and effective client representation. Excessive deference to non-lawyer experts may subject court processes to the systematic influence of vendors’ commercial values and intellectual property protections. And patent protection may jeopardize access to predictive coding technologies, which, in turn, will entrench existing disparities in resources and power. I argue that the profession has a responsibility to explore these costs, as well as the benefits, of predictive coding. It has an ethical obligation to tend to the tools of its trade, as much as the conduct of its members.

I begin in Part I by describing how proponents successfully advanced predictive coding as a valid replacement for human document review and a promising answer to long-standing problems of discovery practice. Notwithstanding significant variation among predictive coding technologies and unresolved questions regarding use, proponents proceeded—and the larger litigation community is now proceeding—as if all predictive coding technologies are of equivalent and unquestionable benefit in discovery practice.

In Part II, I address three problems with the profession’s current attitude and approach to predictive coding, which give rise to three sets of dangers. First, by ignoring outstanding and contested issues regarding the design and use of these technologies, the profession is blinding itself to significant variation in functionality and efficacy. Second, by deferring to the opinions of computer scientists and vendors, the profession is ceding a part of its jurisdiction to self-interested parties. Third and finally, by altering relevant ethical standards to facilitate adoption and use, the profession is weakening the protections and legitimacy of our adversarial system.

In Part III, I offer a series of suggestions as to how the bar can address these dangers, assert its interests, and take a more proactive role in guiding the development and use of predictive coding technologies. I propose that the bar should take the following steps: raise awareness among lawyers and judges through sustained educational efforts; set minimum functionality standards for predictive coding technologies; establish standardized protocols for adoption and use; and take action to ensure widespread access. Predictive coding is not a magic bullet capable of solving all of discovery’s problems, but when designed carefully and employed wisely, it can be a useful instrument in the profession’s toolkit.

I. A NEW APPROACH TO DISCOVERY

Predictive coding products were first introduced in the discovery context in the early 2000s. At the time, discovery practice was facing many of the same problems it faces today: overwhelming amounts of data, sky-rocketing costs, and pervasive adversarial excess. And yet, over a decade passed before predictive coding entered mainstream practice as a viable and promising discovery tool equipped to address these problems. This Part describes the way in which advocates promoted the new technologies and secured support—first, from a limited number of lawyers and trial judges, and subsequently, from the profession more broadly.

A. Civil Discovery Practice

The drafters of the Federal Rules of Civil Procedure conceived of civil discovery as a means of “securing the just, speedy, and inexpensive determination of every action and proceeding.”17 Allowing parties to file generalized pleadings and to engage in broad discovery would, they believed, ensure that cases were resolved on their merits rather than on pleading technicalities or information asymmetries.18 It would also eliminate unfair surprise from trials and counteract the effects of wealth and power disparities.19

The system they envisioned entrusted pre-trial discovery to lawyers’ discretion, with limited accountability. Pursuant to their vision, lawyers would exercise good judgment in acting as both representatives of clients

---

19 Id.
and as officers of the court. Lawyers for requesting parties would balance duties to their clients to obtain all relevant and helpful information with duties to opponents and the court to refrain from evasion and delay. Lawyers for producing parties would balance duties to their clients to protect harmful and confidential information with duties to their opponents to respond to all legitimate discovery requests with all relevant and non-privileged information and documents. The system would be largely self-policing with the parties resolving disputes and disagreements between themselves and largely out of view. It would be fueled by trust that attorneys, frequently working out of view of their opponents and the courts, would adhere to both the letter and the spirit of the rules.

Despite the drafters’ best intentions, civil discovery quickly became a new arena for gamesmanship. With the goal of gaining a strategic advantage for a client or of exhausting the resources of an opponent, lawyers began to approach discovery as the linchpin of a dispute. Commentators explain this by reference to game theory: litigants inevitably worried that their opponents were using discovery excessively aggressively (for example, by taking extreme positions on relevancy and privilege determinations), or flatly improperly (for example, by taking particular positions solely to impose heightened costs on an opponent or to obstruct discovery of relevant information). The resulting prisoner’s dilemma motivated both sides to abuse discovery, lest they suffer the severe strategic disadvantage of acting in good faith while their opponent did not. Mutual distrust led to discovery excess and misuse, as well as to escalating litigation costs.

Early voices for reform proposed increased judicial involvement in the discovery process. The problem, many commentators believed, was lawyers’ excessive and unchecked discretion. More involved judges, they hoped, could check attorney discretion and ensure that discovery tools were being used as intended—to ensure the just, speedy, and inexpensive resolution of cases—and not as strategic weapons to wear down an opponent and force settlement. This, they hoped, would rein in adversarial excess and restore trust in the system.

---

21 Id.
22 Beckerman, supra note 6, at 522.
23 Id.
After a series of modest reforms through the middle of the century, the Judicial Conference took more significant action in 1983. Through amendments to the Federal Rules, the Conference empowered judges to restrict discovery where proposed requests for document or other information were not reasonable and proportional to the needs of the case. The intent was to allow lawyers to retain discretion in designing discovery requests and making relevancy and privilege determinations, but to check lawyers’ discretion with judges’ discretion in determining the meaning of proportionate discovery.

As implemented, these reforms neither checked adversarial excess and abuse nor restored trust in the system. The proportionality determination itself entailed significant unreviewable discretion, resting on vague and hard to quantify factors and rarely subject to appellate review. Vastly inconsistent rulings and levels of involvement fueled continued perceptions that the system rested on unchecked professional discretion—albeit a combination of judicial and attorney discretion.

B. E-Discovery

The advent of computer technology and the explosion of electronically-stored information layered a new set of problems on top of existing ones. Computers multiplied the volume of documents produced in the daily course of business and enabled virtually indefinite retention, com-

---

26 Beckerman, supra note 6, at 512.
27 See Cavanaugh, supra note 20, at 779-81.
28 FED. R. CIV. P. 26(b)(2)(C) (limiting discovery where: "(i) the discovery sought is unreasonably cumulative or duplicative, or is obtainable from some other source that is more convenient, less burdensome, or less expensive; (ii) the party seeking discovery has had ample opportunity by discovery in the action to obtain the information sought; or (iii) the discovery is unduly burdensome or expensive, taking into account the needs of the case, the amount in controversy, limitation on the parties’ resources, and the importance of the issues at stake in the litigation."). See also ARTHUR R. MILLER, THE 1983 AMENDMENTS TO THE FEDERAL RULES OF CIVIL Procedure: PROMOTING EFFECTIVE CASE MANAGEMENT AND LAWYER RESPONSIBILITY 32-33 (FJC 1984).
29 Singer, supra note 25, at 180 ("Although attorneys had traditionally enjoyed a great deal of freedom to fashion discovery for their individual cases, from 1983 onward, whether discovery was proportional was for a judge to decide.").
30 Id., at 180-81 (describing the rule as “ineffective,” “seldom used,” and “largely ignored.”); John P. Frank, The Rules of Civil Procedure – Agenda for Reform, 137 U. PA. L. REV. 1883 1891 (1989) (“[Rule 26(b)(2)(C) was] an extremely valuable suggestion to the courts, but it has proved too subtle to do the job. The scalpel having been attempted unsuccessfully, it is now time for the axe.”).
31 Id., at 147.
33 See Beisner, supra note 3, at 550 (“The exponential growth in the volume of electronic documents created by modern computer systems has exacerbated the problem of abusive discovery and is jeopardizing the legal system’s ability to handle even routine matters.”); Jacob Tingen, Technologies That Must Not Be Named: Understanding and Implementing Advanced Search Technologies in E-Discovery, 19 RICH J L & TECH 2, 2 (2012) (noting that email alone produces 100 billion new messages daily).
pounding the burdens of discovery for all parties. Producing parties faced new and sometimes crushing costs when asked to produce data from back-up tapes and other hard-to-access sources. Requesting parties faced the seemingly impossible task of carefully reviewing millions of documents to find the needle in the haystack—the key documents on which the case could turn. All parties faced new and costly questions of document retention and preservation.

Computers did not only create problems, however. They offered new solutions as well. Overwhelmed by volume, lawyers and litigants began using computer software to cull through electronically-stored data sets to eliminate duplicates and consolidate email chains. More controversially, they began relying on key word searches to cull through millions of documents and to produce smaller, more manageable datasets for human review.

Even as lawyers began adopting key-word searching, a growing literature observed significant shortcomings. Accustomed to searching in databases like Westlaw and Lexis, in which data is cleaned and primed for Boolean searching, lawyers were unprepared for the complexity of searching in disorganized collections of documents and data. Moreover, given the ambiguity of language, even the most advanced linguists could not design searches capable of capturing all words or phrases used to refer to a particular subject. Commentators observed that keyword searching was both under-inclusive (risking that important documents would be overlooked) and over-inclusive (raising the costs of review by returning large quantities of non-responsive documents). And yet, as manual document review grew increasingly expensive and even impossible, lawyers felt that they had no choice but to cull down datasets in this way.

34 Beisner, supra note 3, at 550.
36 Id.
37 See, e.g., David C. Blair & M.E. Maron, An evaluation of retrieval effectiveness for a full text document retrieval system, COMMUNICATIONS OF THE ACM 289 (1985) (concluding that under a key word searching approach, up to 80% of the responsive documents in a collection may routinely be missed, and over 70% of the documents retrieved may, upon review, routinely be deemed irrelevant). See also Howard Sklar, Match Point with Recommind’s Predictive Coding – It’s “Man with Machine,” “not ”Many vs. Machine,” 8/1/11 METRO. COMP. COUNS. 16, 2011 WLNR 16293653, p. 1 (August 1, 2011) (noting research showing that keyword searching leads to recall of about 50 percent at best, and likely closer to 20 percent of relevant documents).
Uncertainty surrounding key-word searching and e-discovery more broadly exacerbated distrust in the system. Requesting parties accused producing parties of designating inadequate search terms and phrases with the intent of excluding relevant documents and information. Produc-
ing parties accused requesting parties of unnecessarily increasing costs by interfering with the production process. Lawyers on both sides accused judges of failing to understand and effectively manage electronic discovery. Clients, meanwhile, began worrying that lawyers—virtually unsupervised amidst massive datasets—were extending discovery for the sole purpose of increasing billable hours.

In 2003 and 2004, in a series of opinions in Zubalake v. UBS War-
burg, U.S. District Judge Shira Schiendlin took on the multiple challenges of e-discovery. Among other things, she addressed the skyrocketing costs. She concluded that courts should engage in a cost-shifting analysis only with respect to data that is relatively inaccessible, such as data stored on back-up tapes. She also explained that the cost-shifting should only be considered after the documents and data in question have been produced so that actual costs can be evaluated. In the final Zubalake opinion, Judge Schiendlin also set forth new duties for lawyers with respect to their clients’ electronically-stored information, including “taking affirmative steps to monitor compliance so that all sources of discoverable information are identified and searched.”

The Zubalake opinions filled a void of information and guidance, and spoke directly to countless numbers of lawyers and judges who were eager for clarity and certainty. Although they expressed the views of a single district judge, they were quickly read and received as setting forth definitive standards for e-discovery. They were followed by numerous

---

42 Id.
43 See Singer, supra note 25, at 178 (discussing the need for increased judicial oversight).
44 See id., at 176-77; see also William W. Schwarzer, Slaying the Monsters of Cost and Delay: Would Disclosure be More Effective than Discovery?, 74 JUDICATURE 178, 179 (1991).
46 Zubalake I at 320-23. Judge Schiendlin explained that a key factor in evaluating whether costs were “undue” under Fed. R. Civ. P. 26(b)(2)(iii), warranting cost-shifting, was whether the electronically-stored information was in accessible or inaccessible form—a question that turned on the media it was saved to. Id.
47 Id.
48 Zubalake V, at 432 (explaining that attorneys are obligated to take affirmative steps to ensure that clients retain, identify, and produce relevant electronically-stored information).
federal district courts and state trial and appellate courts, and they were widely cited by secondary sources and practice guides. The Zubalake opinions also set the stage for the December 2006 e-discovery amendments to the Federal Rules of Civil Procedure. Among other things, the new e-discovery rules presumptively authorize discovery of reasonably accessible electronically-stored information, while requiring a showing of good cause for discovery of difficult to access sources. The good cause determination requires judges to evaluate whether the costs of production are proportional to the particulars of the case.

Similar to previous reform efforts, the e-discovery rules seek to check attorney discretion with increased judicial involvement. Lawyers retain discretion to determine, in the first instance, whether particular sources of electronically-stored information are reasonably accessible. Judges exercise discretion in reviewing that decision, and also in determining whether the costs and benefits of discovery from hard-to-access sources are proportional to the case at hand. But also similar to previous reform efforts, the resulting arrangement has failed to restore trust in the discovery system. The system continues to rest on professional discretion that is exercised out of view and with little transparency or accountability. As instances and perceptions of discovery abuse continue, the natural conclusion is that lawyers, judges, or both are abusing their discretion or, at the very least, exercising it poorly.


51 See, e.g., Michael W. Deyo, Deconstructing Pension Committee: the evolving rules of evidence spoliation and sanctions in the electronic discovery era, 75.1 ALBANY L. REV. 305, 306 (Fall 2011) (“Whether one agrees or disagrees with the lines drawn by Judge Scheindlin, her Zubulake opinions indisputably captured widespread attention and left indelible marks on the nation’s judicial system.”).

52 Id.

53 FED. R. CIV. P. 26(b)(2) (“A party need not provide discovery of electronically stored information from sources that the party identifies as not reasonably accessible because of undue burden or cost. On motion to compel discovery or for a protective order, the party from whom discovery is sought must show that the information is not reasonably accessible because of undue burden or cost. If that showing is made, the court may nonetheless order discovery from such sources if the requesting party shows good cause, considering the limitations of Rule 26(b)(2)(C). The court may specify conditions for the discovery.”).

54 Id. (explaining the good cause determination as requiring a court to “balance the costs and potential benefits of discovery”).

DRAFT—DO NOT CITE
C. Predictive Coding

Against this backdrop, a growing chorus from inside and outside the profession began advocating a new type of reform. Rather than seeking to restore trust in the lawyers and judges engaged in discovery—as task that seemed increasingly impossible—they sought to establish trust in technology. They advanced and advocated predictive coding. Predictive coding, they claimed, could provide answers to all of the problems of discovery practice by replacing the subjectivity and discretion of human review with the mechanical objectivity of computer technology.55

Predictive coding encompasses great variety with respect to both underlying technologies and implementing procedures.56 Under a typical protocol, lawyers code a seed set—a sample set of documents from the larger dataset—for responsiveness and privilege.57 Based on the seed set coding, a computer generates a search algorithm for identifying responsive and privileged documents, which is then used to review another sample set of documents for responsiveness and privilege, which the lawyers then review and correct. This iterative process continues until the lawyers, often in consultation with the software vendor, are satisfied that the computer has been adequately trained to identify responsive and privileged documents. The computer then reviews the entire dataset.58 (Predictive coding therefore differs in significant ways from key-word searching, which culls through a document set and produces a smaller set for human review.)

55 Cf. PORTER, TRUST IN NUMBERS, supra note 1, at 90.
56 McCreary, supra note 2, at *1. Some softwares employ Bayesian classifiers, which compute a mathematical thumbprint of each document by placing numerical values on a number of document characteristics relating to the author, custodian, and content. They then employ statistical probability models to translate each document's mathematical thumbprint into a relevancy determination. See Maura R. Grossman & Terry Sweeney, What Lawyers Need to Know about Search Tools, Electronic Discovery: A Special Report, NAT. L. J. 2 (August 23, 2010). Others use latent semantic indexing, which also assigns mathematical values to documents but does so by identifying patterns in the relationships between particular words and usage of words in particular contexts. Id.; see also Jason R. Baron, Law in the Age of Exabytes: Some Further Thoughts on 'Information Inflation' and Current Issues in E-Discovery Search, 17 RICH. J.L. & TECH. 9, 25-26 (2011). Implementing protocols vary substantially as well. Some processes employ a binary system of labeling documents responsive or not, while others rank documents from the most to least relevant. Most often, the documents coded not responsive or with a low relevancy score are never subject to human review, while those coded responsive or with a high relevancy score are subsequently reviewed for confirmation of responsiveness and privilege and for creation of a privilege log. Pursuant to other variations, a statistically significant sample of documents identified as not responsive are also subject to human review. Brendan M. Schulman & Samantha V. Ettari, Federal Court Approves the Use of “Predictive Coding” Technology-Assisted Document Review, 5/1/12 METRO. COMP. COUNS. 18, 2012 WLNR 10398490, p. 1-2 (May 1, 2012).
57 For an overview of the functioning of predictive coding procedures generally, see Schulman & Ettari, supra note 56, at 1; Predictive Coding and Patented Workflow, supra note 4, at p.4.
58 Id.
As noted, the first predictive coding products were introduced in the discovery context in the early 2000s. They attracted little attention, however, until two empirical pilot studies were published in 2010 that favorably compared predictive coding approaches to traditional manual review.\textsuperscript{59} Based on simulated discovery exercises, both studies concluded that the particular predictive coding approaches tested (three in total) achieved higher levels of recall (the fraction of relevant documents that are identified) and precision (the fraction of identified documents that are relevant)\textsuperscript{60} than manual document review.\textsuperscript{61}

These pilot studies enabled vendors and other proponents to argue persuasively that predictive coding provided a solution to the problems plaguing discovery—not only the primary practical problem of increasingly unmanageable datasets, but also deeper-seated and longer-standing problems of excess, abuse, and trust. Pointing to the studies’ findings that human error and inconsistency were unavoidable aspects of manual review, they emphasized the problems of human fallibility and unconstrained attorney discretion that characterized the existing system.\textsuperscript{62}

Pointing to the studies’ findings that predictive coding consistently out-


\textsuperscript{60} See Grossman & Cormack, supra note 59, at 7 (“The fraction of relevant documents identified during a review is known as recall, while the fraction of identified documents that are relevant is known as precision. That is, recall is a measure of completeness, while precision is a measure of accuracy, or correctness.”). But see Roitblat et al., supra note 59, at 76 (“The use of precision and recall implies the availability of a stable ground truth against which to compare the assessments. Given the known variability of human judgments, we do not believe that we have a solid enough foundation to claim that we know which documents are truly relevant and which are not.”).

\textsuperscript{61} The Grossman & Cormack article has proven the most influential. Maura Grossman is a litigator at Wachtell, and Gordon Cormack is a computer scientist at the University of Waterloo. Their study was based on participation in the Text Retrieval Conference (TREC) sponsored by the National Institute of Standards and Technology. See infra note 143. Based on their analysis of data from the 2009 Legal Track Interactive Task, Grossman and Cormack reported that predictive coding had enabled two different teams to achieve results superior to those achieved by teams of human reviewers, where “superior results” were comprised of higher recall and higher precision. Moreover, they claimed that the predictive coding review process was significantly less expensive than that of the teams relying exclusively on human review. Grossman & Cormack, supra note 59, at 11. Significantly, however, Grossman and Cormack acknowledged a diversity of approaches encompassed by the term “predictive coding,” and responsibly limited their findings to the specific protocols employed at TREC’s 2009 Legal Track. But many advocates of predictive coding—lawyers, judges, and vendors alike—ignored these qualifications and began using the study to argue for widespread adoption.

\textsuperscript{62} They note, for example, that traditional manual review regularly entails not only actual error but also divergent judgments. It is not at all unusual for human assessors to disagree on whether a document is relevant or irrelevant, not as a result of error but of legal judgment. Divergent judgments are even more common with respect to privilege determinations, where different lawyers deliberately take more or less aggressive approaches. See Maura Grossman & Gordon Cormack, Inconsistent Responsiveness Determination in Document Review: Difference of Opinion or Human Error? 32 PACE L. REV. 267 (2012).
performed manual review, they argued that the new technologies could solve these problems. Predictive coding, they contended, offered the mechanical objectivity and accuracy of a computer program in place of the fallibility and unpredictable discretion of human review. And it could do so at a fraction of the cost of human review.

As the number of vendors marketing predictive coding technologies multiplied to over fifty, support grew in pockets. A number of corporations brought predictive coding capabilities in-house for data management purposes and, with the help of lawyers, began using it to cull through documents produced by an opponent in litigation. Anecdotally, some lawyers and litigants used predictive coding tools in document productions without disclosing that fact to the court or their opponent. Most lawyers, however, remained either unaware of the new technology or reluctant to use it absent official endorsement. They “watched from the sidelines,” waiting for an official judicial statement that use of predictive coding could fulfill discovery obligations.

In October 2011, Magistrate Judge Andrew Peck, a vocal advocate of predictive coding, published a bar journal article offering that reassurance. Relying on the 2010 empirical pilot studies, he endorsed predictive coding as a vital tool for discovery. He responded to the bar’s uncertainty by stating that “[u]ntil there is a judicial opinion approving (or

---

63 See, e.g., Predictive Coding and TAR after De Silva Moore, Interview with Skip Durocher and Caroline Boudreau Sweeney, 7/1/12 METRO. COMP. COUNS. 39, 2012 WLNR 14913799, p.1 (July 17, 2012) (drawing a sharp contrast between the dangers of “subjective review by humans, with each person making his or her individual judgment call,” and the reliability of predictive coding’s increased accuracy and mechanized objectivity). See also Predictive Coding and Patented Workflow, supra note 4, at p.4 (arguing that the technology would even increase the accuracy and reliability of supervisory lawyers: “[t]he technology makes human beings more accurate by reducing time spent on reviewing irrelevant documents, thereby circumventing the natural human tendency to lose focus when an activity becomes less productive”).

64 See, e.g., Evidence Mounting in the Case for Predictive Coding, Interview with William Tolson, Senior Project Manager, Recommind, 10/1/12 METRO. COMP. COUNS. 29, 2012 WLNR 21693209, p. 2 (October 1, 2012); Howard Sklar & Michael Potters, Getting it Right: Training and Certification in Predictive Coding, 10/1/12 METRO. COMP. COUNS. 32, 2012 WLNR 21693265, *3 (October 1, 2012); McCreary, supra note 2, at *3.


67 Da Silva Moore Panel, supra note 66 (statement of Crowley).

68 Id.

69 Id.

70 Peck, supra note 14, at 29. Judge Peck also expressed support for predictive coding as a speaker at various e-discovery conferences. See Schulman & Ettari, supra note 56, at p.3.

71 Peck, supra note 14, at 29.
even critiquing) the use of predictive coding, counsel will just have to rely on this article as a sign of judicial approval.”

Six months later, Judge Peck issued the first judicial opinion approving use of the new technology as an acceptable and desirable means of meeting discovery obligations. The case, Da Silva Moore v. Publicis Groupe, called upon Judge Peck to resolve a dispute between two parties who had initially agreed to use predictive coding but subsequently disagreed on implementing protocols. Judge Peck took the opportunity to endorse predictive coding broadly as a technology that “appears to be better than the available alternatives and thus should be used in appropriate cases.”

Accepting the claims and conclusions of the 2010 empirical pilot studies uncritically, he interpreted them to support the accuracy of all predictive coding systems. These studies, Peck reasoned, enabled him to set aside the “science of the technology” and to focus on “the ways it was used.” He informed lawyers generally that “[c]omputer-assisted review now can be considered judicially approved for use in appropriate cases.” In the case before him, he accepted and approved the underlying computer technology without reservation and turned his attention towards the most effective way to use it in the discovery process. He believed heightened—even mandatory—cooperation and transparency to be critical. Accordingly, he ordered the producing party to be fully transparent regarding the documents and coding of its seed set used to train the computer.

Judge Peck’s decision, which was affirmed by the district court, was described as a “watershed moment” that “completely mobilized the industry.” Within weeks, it was followed by two additional cases in which judges endorsed the use of predictive coding. Those, in turn,
were followed by an explosion of conferences, CLE classes, and proposed rule changes, all intended to advance its adoption and use.\(^{82}\)

As Judge Peck had hoped, the conversation quickly shifted from whether to use the technology to how to use it.\(^{83}\) In his October 2010 article, Judge Peck had explained that he was “less interested in the science behind the ‘black box’ of the vendor’s software” than in how it could be used to “produce responsive documents with reasonably high recall and high precision.” \(^{84}\) Expressing these preferences in *Da Silva Moore*, he had explicitly directed the parties’ attention away from the underlying technology and towards implementing procedures.\(^{85}\) Taking his cue, other judges, lawyers, and litigants accepted the functionality of the underlying technology and turned their focus exclusively to designing efficient and effective coding protocols. Vendors, eager for widespread acceptance, enthusiastically endorsed this approach through articles and panels.\(^{86}\)

*Da Silva Moore* marked a turning point in the profession’s acceptance of predictive coding. Much like with Judge Schiendlin’s opinions in *Zubalake*, the opinion gained widespread influence. Although it was a single opinion from a magistrate judge, it offered the official support that many stakeholders had been awaiting, at a time when they were primed and ready to receive it.\(^{87}\) When the opinion issued, computer scientists had already offered evidence of the accuracy of certain predictive coding softwares. Vendors were actively marketing countless new products. Proponents within and outside of the profession were framing predictive coding as a solution, not just to the primary practical problem of unmanageable datasets but also to the broader problems of excess, abuse,
and trust that had long plagued civil discovery. The litigation community was ready to accept predictive coding as a silver-bullet answer.

II. THE CONSEQUENCES OF ADOPTION

Notwithstanding Judge Peck’s unequivocal language and the litigation community’s broad support, predictive coding is not an unmitigated good. At the same time that it holds undoubted potential as a useful discovery tool, it also threatens harm to the profession and the public. In this Part, I argue that the profession’s oversimplified view of predictive coding’s promise gives rise to three sets of dangers: It problematically elides significant variation in the definition and use of predictive coding technologies; it threatens the scope of the profession’s jurisdiction; and it undermines the protections and integrity of the adversarial system.

A. Variation and Choice

Predictive coding technologies are currently characterized by significant variation and uncertainty—variation in the products and processes that vendors label “predictive coding,” and uncertainty in their functionality and alleged benefits. Judges and vendors, both key stakeholders with significant power, are working to suppress this variation and uncertainty and to achieve premature and superficial closure. Doing so serves their interests, but not the interests of the profession and the public.

The term “predictive coding” encompasses a countless variety of vendors and technologies with very different types of functionality. By one count, there are currently over 30 different types of classifiers capable of machine learning-based text classification, many of which support countless applications in the context of document review.\(^88\) Given that different predictive coding products employ different classifiers, and that different classifiers are optimally suited for different types of datasets, the choice of product can have significant implications for the quality of results.\(^89\) The 2010 empirical studies reporting increased accuracy—the only of their kind thus far—did not capture this.\(^90\) They looked to three predictive coding technologies, leaving the vast majority untested.\(^91\)

\(^{88}\) Michael McCreary, supra note 2, at *1.
\(^{89}\) Id. at *2 (noting that “when the classifier and data are misaligned, results can be less accurate than even manual review.”).
\(^{90}\) See supra note 59.
\(^{91}\) Herbert Roitblat, Anne Kershaw, and Patrick Oot studied the level of agreement between four teams reviewing the same set of documents. Two of the teams employed manual review and two employed technology-assisted review. They concluded that the teams employing predictive coding achieved about the same recall as the teams employing manual review while achieving somewhat better precision. Unfortunately, the published study fails to describe the protocols em-
Proponents’ promises of reduced discovery abuse and expense are no more certain than claims of increased accuracy. Lawyers who train the computer systems can continue to make aggressive and even abusive relevancy or privilege determinations, which will be extrapolated over the entire document set. Parties can and likely will employ experts to fight over coding protocols. Parties may also feel justified asking for and/or producing exponentially increased numbers of documents and thereby creating a new source of increased costs. Existing data regarding costs savings fail to account for these eventualities. The data stems from pilot projects that calculate savings on document review, but fails to account for the possibility of increased costs elsewhere.

Notwithstanding this variation and uncertainty, the litigation community is embracing predictive coding as if its definition is clear and its efficacy established—as a well settled cure-all. As one lawyer recently explained: “[w]e have moved beyond the technology issue now to discussions of the extent to which the parties should be transparent and how one should focus on issues of defensibility in the process employed.”

Eager for a new kind of solution to long-standing problems and reassured by Judge Peck’s confident judicial endorsement, lawyers are therefore accepting the underlying technology as a fail-safe solution as long as it is implemented properly.

Two implicated parties benefit from this story of technological closure. Vendors, motivated by potential profits, can use it to increase demand for their products and services. Trial judges, motivated by the goal of efficient and effective case management, can rely on it to encourage parties to cooperate in adopting and implementing the new technology.

ployed by any of the four teams. Roitblat et al., supra note 59, at 61. Maura Grossman and Gordon Cormack studied two different computer-assisted approaches employed by teams participating in the TREC 2009 Legal Track. See id. One, the method of the vendor H5, was a sophisticated form of key-word searching that entailed an iterative feedback loop to design complex search strings. See Dan Brassil et al., The Centrality of User Modeling to High Recall with High Precision Search, in 2009 IEEE INTL. CONF. ON SYSTEMS, MAN, AND CYBERNETICS. 91, 91-96. The other was a predictive coding approach relying on machine learning. See Grossman & Cormack, supra note 59, at 40-41. The study concluded that both approaches compared favorably to manual review on the same document set. Id., at 28. However, because the two computer-assisted approaches were used on different document sets, they could not be compared to each other.

92 Looby & Kaplan, supra note 82, at p.3.

93 See, e.g., Evidence Mounting, supra note 64, at p.2 (Calculating cost-savings as follows: “To calculate the numerator, determine the cost of a traditional e-discovery process (from start to finish) and then subtract the new cost of completing that same process using a Predictive Coding system. From this difference, subtract the investment made in implementing a Predictive Coding system. Finally divide the resulting number by the investment made in implementing a Predictive Coding system.”).

94 See id., at p.2 (reporting on an FTI cost-savings study and observing that “[t]he verdict is still out on cost, as well as on potential savings”); id., at p.3 (“In certain circumstances, predictive coding can cost more.”).

95 Da Silva Moore Panel, supra note 66 (statement of Crowley).
The broader profession’s acceptance of this story is dangerous, however. It blinds lawyers to the choices they face and the complexity of the tools they employ. It blinds the bar to a pressing need for studies evaluating and comparing various predictive coding products and minimum functionality standards for products that will be accepted by courts. And, as discussed in the following two sections, it breeds a lack of understanding among lawyers and judges that precludes the profession from employing the new technology in service of clients’ interests and the public interest, rather than the interests of managerial judges or commercial vendors.

B. The Profession’s Jurisdiction

A second but related set of problems stems from the ways in which predictive coding is eroding the profession’s jurisdiction. It is doing so in three principal ways: First and most obviously, it is replacing lawyers’ jobs and raising questions of unauthorized legal practice. Second, it is undermining judges’ and lawyers’ control over court processes. Finally, it is allowing for the patenting of law-practicing algorithms. In all three ways, it is interfering with the profession’s ability to meet its publicly-oriented obligations.

1. The Unauthorized Practice of Law

Predictive coding technologies replace elements of lawyer labor with machine labor. Under a predictive coding approach to discovery, tasks once entrusted exclusively to associates and contract attorneys are now delegated to distributed networks of computers, lawyers, and technology specialists. Predictive coding therefore raises questions about what constitutes the practice of law, but these questions are not readily answerable in the current framework of rules governing unauthorized legal practice. Unlike prior technological developments, predictive coding technologies do not implicate the traditional justifications for these rules. They do, however, raise new and potentially more troubling ethical questions regarding the extent of lawyers’ duties to understand and supervise legal work.

The deskilling of parts of legal practice was also implicated in the 1990s with the introduction of software programs that generated basic legal instruments such as simple wills and contracts—Quicken Family Lawyer, for example. Software companies explicitly marketed these programs as cost-effective alternatives to consultation with lawyers. Ar—

96 Sklar & Potters, supra note 64, at p.3.
guing forcefully that these programs risked harm to consumers by supplan
ting the individualized guidance of a lawyer in determining what 
form to use for what purpose and how to complete it, the bar mobilized 
against them as unauthorized legal practice.98 In a handful of states, high courts agreed and deemed the software programs the unauthorized practice of law.99

The consumer protection concerns that were at play in that context are largely absent in the predictive coding context. Within both firms and large corporate organizations, predictive coding technologies are employed as lawyers’ tools. Clients’ interactions with them are mediated by lawyers who train the computers, monitor protocols, and ensure quality control. As a result, predictive coding implicates far fewer client protection concerns than situations in which clients interact with a technology directly and lawyer advice is eliminated entirely.100

In light of this, a more useful analogy may be the delegation of document review to off-shore document processing firms, where legal work is supervised but not performed by lawyers licensed within United States jurisdictions. The American Bar Association (ABA) and several state commissions have issued formal opinions concluding that outsourcing to these firms does not constitute unauthorized legal practice as long as licensed U.S. attorneys retain strict supervisory roles.101

Like in the offshore outsourcing context, predictive coding technologies are typically employed with lawyer participation and oversight. The

98 See, e.g., William A Scott, Filling in the Blanks?: How Computerized Forms are Affective the Legal Profession, 13 ALB. L. J. SCI. & TECH. 835, 851 (2003) (“The potential harm these forms can cause is great, especially when relied on by a person to draft a will. True harm caused would not be realized until the testator dies and the forms reach the surrogate’s court for probate.”); William H. Brown, Legal Software and the Unauthorized Practice of Law: Protection or Protectionism, 36 CAL. W. L. REV. 157, 162 (1999) (summarizing concerns, which included that “material may give advice that is incorrect or misleading,” and that “a layperson is not subject to malpractice suit or discipline by the state bar”).


100 The principal justification offered for UPL prohibitions is “to protect the public from the consequences of inexpert legal services.” ABA MODEL RULES ANN., R 5.5 cmt at 453 (1996); see also id. (“Other interests including protecting the integrity of the judicial system and providing a means for regulation of the profession.”).

101 ABA Standing Committee on Ethics and Professional Responsibility, Formal Opinion 08-451 (August 2008) (“A lawyer may outsource legal or nonlegal support services provided the lawyer remains ultimately responsible for rendering competent legal services to the client under Model Rule 1.1.”); see also Committee on Professional Ethics of the Association of the Bar of the City of New York Ethics Opinion 2006-3 (Aug. 2006), available at http://www.abcny.org/Ethics/etho6.htm (a lawyer may outsource legal support services to overseas lawyers and nonlawyers if the lawyer supervises the work rigorously); Florida Bar Professional Ethics Committee, Op. 07-2 (Jan. 18, 2008, approved July 25, 2008) (approving of off-shore outsourcing).
analogy is again imperfect, however, with predictive coding raising additional ethical questions. There is a difference between delegating to lawyers who are not licensed in the United States and delegating to non-lawyers—in this case, non-humans. Whereas a supervising attorney would typically be familiar with the legal work product and processes of legal and paralegal staff, she may not understand predictive coding technologies. She may lack the proper analytical tools to assess whether a particular technology is adequate for the task, or whether it is working properly when employed, raising questions about the scope of lawyer supervision.

Indeed, all segments of the legal profession have long been criticized for a lack of technological competence. With regards to predictive coding, judges and lawyers alike lack clear definitions of which computer software programs and processes constitute predictive coding, how predictive coding programs work, and what various accuracy levels mean. This low level of technological literacy is reinforced by the ABA’s Model Rules of Professional Conduct, which prescribe a reduced level of required oversight for automated legal work. A new comment to ABA Model Rule 5.3, Responsibilities Regarding Nonlawyer Assistance, adopted in 2012, addresses a lawyer’s oversight responsibilities with respect to a non-lawyer service provider such as a predictive coding vendor.\textsuperscript{102} Implicitly accepting that lawyers lack the requisite knowledge to supervise such a vendor, the comment uses the word “monitor”—meaning “remain[ing] aware of how nonlawyer services are being performed”\textsuperscript{103}—to indicate a lowered oversight standard. The comment also concedes that in some situations, it will be more appropriate for a client rather than the lawyer to assume vendor oversight responsibilities.

Through these new provisions, the ABA abdicates a portion of the profession’s supervisory responsibilities over discovery practice to clients and other professionals. This, in turn, makes predictive coding ethically risker (at least in some ways), than forms software or offshore outsourcing. Although lawyers continue to be involved—obviating the consumer protection concerns that arose with forms statutes and that generally animate unauthorized practice of law statutes—they fail to understand and take responsibility for the discovery tools they are employing. Clients therefore lack the protections of strict supervision that they have with outsourcing. The ramifications may be severe—lawyer ignorance

\textsuperscript{102} ABA, MODEL RULE OF PROFESSIONAL CONDUCT 5.3, Comment [4].

of discovery tools may compromise the competency of representation and risk an unintentional breach of confidentiality. As discussed next, it will also raise jurisdictional issues in the courtroom.

2. Control over Court Processes

Regardless of whether we characterize predictive coding as unauthorized legal practice, it is clearly encroaching on the profession’s control over court processes. It is transforming litigation procedure—traditionally the exclusive domain of judges and lawyers—into a shared domain with computer scientists, commercial vendors, and others. This would not necessarily be problematic if lawyers and judges were collaborating with IT professionals in this shared domain. But instead, they are deferring to the IT professionals and in doing so, are privileging vendors’ values over the values of the profession and the public.

As just discussed, lawyers and judges frequently lack the technological expertise to understand and oversee the use of predictive coding—a reality that is endorsed by the ABA’s guidance on vendor monitoring. As a result, lawyers are looking to non-lawyer IT experts to determine which products they should use and how they should use them. Increasingly, they are bringing these experts into court to defend the quality and functionality of particular technologies and the efficacy and defensibility of particular protocols. IT experts are opining not only on the quality of the underlying technology, but also on the ways in which it is used in discovery—for example, how lawyers should populate the seed set, train the computer, and check for quality control.

These non-lawyer experts are highly interested parties, however. They likely hold a world view that privileges technological use and development. They have no reason to recognize, much less incorporate, lawyers’ ethical obligations to clients, the court, and the public. Indeed, many are employed by predictive coding vendors and may therefore have internalized their employers’ profit motives. In deferring to them, judges and lawyers are deferring to parties whose values and goals diverge significantly from those of the profession. They are ceding control to them over what has long been considered a core domain of legal work—litigation procedure.

104 See, e.g., Looby & Kaplan, supra note 82, at p.3 (noting the view that “understanding different predictive coding algorithms and classifiers is for the technology geeks to deal with.”).
105 See id.
106 See McCreary, supra note 2, at *2.
107 Da Silva Moore Panel, supra note 66, at p.6 (June 1, 2012) (statement of Foley) (arguing that “predictive coding technology can’t be the exclusive province of technology aficionados…[T]he people who understand the facts and the procedural posture of the case need to be involved n training the system.”)
A handful of judges and commentators have proposed that courts bring these experts back within judicial control through the use of *Daubert* hearings. They contend that requiring parties to establish the reliability of the experts who are designing and defending particular technologies and protocols will ensure that predictive coding delivers on its promised benefits of increased accuracy and efficiency. It could also place a much-needed check on vendor interests and influence.

The applicability of *Daubert* hearings to the opinions and advice of predictive coding experts is not obvious, and would stretch their original purpose. Now codified at Federal Rule of Evidence 705, *Daubert* hearings are intended to ensure the reliability of expert evidence offered at trial. They establish whether or not experts’ testimony should be admitted into evidence by reference to an inquiry into whether the experts meet certain standards of training and reliability in their respective fields. With respect to predictive coding, however, experts are not vouching for the admissibility of scientific evidence but rather for the reliability of the process of identifying documents that are responsive to discovery requests.

Ultimately, repurposing *Daubert* hearings to qualify predictive coding technologies would represent another step in the profession’s abdication of responsibility with regards to the new technologies. In light of the sparse existing empirical data on the functionality and benefits of predictive coding, the immediate result would be to bring vendor-experts before the court to testify about their opinion on the reliability of their own products, which they have a strong interest in promoting. Moreover, incorporating litigation procedure in *Daubert* hearings would reinforce the impression that document review as an aspect of discovery practice is no longer a professional task of lawyers, now falling within the purview of other experts. This, in turn, would reinforce lawyers’ technological ignorance by sending a clear signal that lawyers are not expected to understand predictive coding technologies.

---


110 Fed. R. Evid. 705 (excluding expert testimony unless “(1) the testimony is based upon sufficient facts or data, (2) the testimony is the product of reliable principles and methods, and (3) the witness has applied the principles and methods reliably to the facts of the case.”).
3. Patent Monopolies

Predictive coding is jeopardizing the profession’s jurisdiction in a third way. A number of vendors are applying for and receiving patents on the law-practicing algorithms that lie at the base of their predictive coding technologies. These patents are interfering with the profession’s ability to meet its core public obligation—ensuring widespread access to legal services.

Controversy has long surrounded the patentability of algorithms, which, until the 1970s, were not considered patentable subject matter. They were viewed as either laws of nature or abstract thoughts, and therefore not the product of invention. Then, in 1972, the U.S. Supreme Court held that algorithms could be patented if applied in sufficiently concrete and practical ways. The Court offered little guidance as to how and when that standard was met, but parties began regularly applying for, and the Patent and Trademark Office (PTO) began regularly granting, patents for a range of processes that applied mathematical algorithms. Ever since, courts have struggled to draw a principled line between patentable algorithms and unpatentable abstract ideas. Because of their connection to human cognition, predictive coding technologies offer a salient illustration of the difficulty.

Nevertheless, the PTO has issued several patents on machine learning technologies employed in document review, and many more applications are pending. Some of the issued patents are quite broad, purporting to cover the entire predictive coding landscape.

---

111 Initially, courts concluded that algorithms were not patentable because they were abstract ideas, analogous to mental processes. See Gottschalk v. Benson, 409 U.S. 63, 72-73 (1972). But in the late 1970s, the Federal Circuit held that patents could issue if the algorithm in question “is applied to one or more elements of an otherwise statutory process claim.” In re Freeman, 573 F.2d 1237 (CCPA 1978); In re Walter, 618 F.2d 758 (CCPA 1980); and In re Abele, 684 F.2d 902 (CCPA 1982). Long-standing confusion over how to determine when an algorithm is and is not patentable was not lessened by the Supreme Court’s recent statement on the issue. See Bilski v. Kappos, 130 S. Ct. 3218, 3225 (2010). In the 3 years since Bilski, the courts have struggled to define criteria suitable to determine whether a process based on an algorithm should be patentable.

112 These cannot be patented because they are the building blocks of innovation, properly located in the public domain. See Arrhythmia Research Tech., Inc. v. Corazonix Corp., 958 F.2d 1053, 1057 (Fed. Cir. 1992) (“The law crystallized about the principle that claims directed solely to an abstract mathematical formula or equation, including the mathematical expression of scientific truth or a law of nature, whether directly or indirectly stated, are nonstatutory under section 101; whereas claims to a specific process or apparatus that is implemented in accordance with a mathematical algorithm will generally satisfy section 101.”).


comind, Inc. holds a patent that covers “[a] method for analyzing a plurality of documents . . . via a computing device.”

Others are narrower, claiming particular computer functionalities as applied in document review. For example, Equivio Ltd. holds a patent for a specific method of producing a search algorithm. Thus far, only one predictive coding patent has been challenged in court. The case was dismissed on procedural grounds.

Patents on law-practicing algorithms will interfere with the profession’s ability to ensure widespread access to justice, which includes widespread access to the tools of lawyering. Patents allow their holders to prohibit use, sale, and reverse engineering of an invention for twenty-one years. The patenting of predictive coding technologies therefore enables vendor exclusivity over computerized discovery practice. It could allow a single vendor to drive out all competition and control the predictive coding market. The vendor would then enjoy unlimited power in setting licensing fees or in refusing to license the technology at all. Accordingly, although predictive coding patents cannot encroach upon a firm or lawyer’s ability to conduct discovery manually, they can and will interfere with lawyers’ abilities to provide the best, most competent, and most up-to-date legal services to their clients.

Already, high start-up costs create uneven access to predictive coding technologies. Patents will exacerbate these issues in ways that the profession cannot combat. They will increase the frequency of cases in which one party has access to predictive coding technologies and the other party does not. These cases, in turn, will entail fundamental unfairness and, possibly, increased abuse. Poorer litigants, who will be unfamiliar with use of the technology and unable to hire experts, will have

---

117 See, e.g., U.S. Patent No. 8,346,685 (filed April 22, 2009) (“Computerized System for enhancing expert-based processes and methods useful in conjunction therewith”; see also U.S. Patent App. 20130077857 (filed Apr. 21, 2012) (“System to prepare proactively for discovery requests by storing all correspondence in a predictive-coding like database, that can then be searched for relevant documents when a request for document production is made); U.S. Patent App. 2012078266 (filed April 27, 2012) (claiming a method of populating and coding the seed set to train the computer).
118 U.S. Patent No. 8,346,685 (filed April 22, 2009).
no ability to challenge the discovery approaches and search protocols of wealthier and more sophisticated parties. Wealthier litigants, who will be aware that their opponents have no predictive coding capabilities and limited resources to fund manual review, may hide relevant and damming documents amidst massive document productions.

Predictive coding’s threat to the profession’s jurisdiction is not merely academic, and cannot be labeled a purely protectionist concern. Technological ignorance will disable lawyers from providing strong and effective client representation. Excessive deference to non-lawyer experts will subject court processes to the systematic influence of vendors’ commercial values and intellectual property protections. Finally, patent protection will increase unequal access to predictive coding technologies, which, in turn, will entrench existing disparities in resources and power.

C. Adversarial Values

The profession’s current approach to the adoption and use of predictive coding is causing a third set of problems, which stem from judges’ and commentators’ heightened emphasis on cooperation in the use of predictive coding. Cooperation has long been a core value of the discovery process—a means of ensuring the just and efficient use of discovery in the resolution of cases.120 The drafters of the Federal Rules initially conceived of discovery as a cooperative process between parties,121 and subsequent amendments consistently aimed to increase cooperation.122 Most recently, the rules were amended to establish a system of mandatory initial disclosures between parties.123 Cooperation, however, is only one system value to be balanced among many. In ignoring this and insisting upon new kinds and degrees of cooperation in the use of predictive coding, managerial trial judges are threatening core client protections and the legitimacy of our courts.

---

120 See, e.g., FED. R. CIV. P. 26(f) (requiring parties to discuss an in depth discovery plan in advance “in order to secure the just, speedy and inexpensive determination of the case”); Baron, supra note 56, at 6 (“To meet the challenge of the exploding volume and complexity of potential electronic evidence, lawyers must [among other things… think about new ways of approaching structured cooperation within the bounds of the adversary system.”)).

121 Subrin, supra note 17, at 717.

122 See Fed. R. Civ. P. 26(b)(5) advisory committee’s note to 2006 amendment; Fed. R. Civ. P. 26(a) (1) advisory committee’s note to 2000 amendment; Order Prescribing Amendments to the Federal Rules of Civil Procedure, 507 U.S. 1089, 1118-21, 1125 (1993); see also, Beisner, supra note 3, at 563, 578, 582.

123 FED. R. CIV. P. 26(a)(1) (mandating initial disclosure).
1. New Kinds of Cooperation

Judge Peck and a handful of other judges required new types of cooperation between parties in the predictive coding protocols they approved. The protocol Judge Peck approved in *Da Silva Moore*, for example, required the defendants (the producing party) to grant the plaintiffs full access to the documents and coding of their seed set, as well as an opportunity to provide input on the initial coding decisions and subsequent quality control efforts. Other judges have similarly required or encouraged seed set transparency. Commentators, for their part, are proposing even higher levels of required cooperation. The Sedona Conference suggests that a producing party’s knowledge of its own data may create a duty to disclose defects in proposed predictive coding search methodologies.

The type and amount of cooperation being employed and discussed with regards to predictive coding would be unheard of under a traditional regime of manual review. Once parties make initial mandatory disclosures in a system of manual review, they have no further obligation to direct opponents’ discovery requests towards particular documents of information. Access to the process through which an opponent made relevancy and privilege determinations would typically be denied, likely on the grounds of attorney work product. If challenged, they would be subject to an adversarial proceeding before the judge.

Requiring these new forms of cooperation undermines adversarial protections and threatens client interests. As an initial matter, lawyers may be reluctant to challenge a judge’s order for cooperation, fearing that the risk of retaliation will exceed the harm that a cooperative approach might entail. And yet, the potential harm may be substantial. Non-privileged, nonresponsive documents in a seed set could include

---

124 See *Moore*, 287 F.R.D. at 192.
127 Sedona Conference, *The Case for Cooperation*, 10 SEDONA CONF. J. 339, 344 (2009). In addition, some have argued that it should be considered a violation of Rule 3.4 of the Model Rules of Professional Responsibility to fail to suggest a revised search protocol where a producing party knows that the requesting party’s protocol will not capture documents it knows to be responsive. One commentator claims that “such conduct is tantamount to suppression.” See Jason Baron, Symposium, Ethics and Professionalism in the Digital Age: Ninth Annual Georgia Symposium on Ethics and Professionalism, 60 MERCER L. REV. 863, 877 (2009).
information that reveals unethical or criminal activity by a party, that embarrasses an officer or employee, or that aids the requesting party in an unrelated cause of action. In requiring full disclosure of such information as part of seed set transparency, judges have failed to acknowledge this potential harm. More broadly, they have failed to acknowledge threats posed by predictive coding technologies to core protections for attorney work product, attorney client privilege, and confidentiality.

2. New Degrees of Cooperation

The profession’s embrace of predictive coding is problematically elevating cooperation over other system values in a second way. It is accelerating a pre-existing trend away from the overarching goal of comprehensiveness in discovery practice. If cooperation is elevated and comprehensiveness eliminated, the discovery system will break entirely with the adversary system and calls into question the legitimacy of court processes.

Comprehensiveness—the goal of bringing to light all relevant and non-privileged information so as to construct the strongest possible case for clients and against opponents—has long provided a check on cooperation in discovery. It is always in the requesting and producing parties’ respective interests to interpret the goal differently. Requesting parties emphasize discovering as much information as possible, in order to construct the strongest possible case for their client and against their opponent. Producing parties emphasize withholding non-relevant and privileged information in order to protect their client against confidential disclosures. The parties’ divergent interests ensure that any and all cooperation in discovery practice occurs within the framework of the adversarial system’s strong client protections.

Conceding that the costs of achieving true comprehensiveness were too high for the system to bear, the Judicial Conference qualified and modified the goals of discovery long before the advent of predictive coding. In 1983, in the name of limiting adversarial excess and reducing

---


129 Some commentators have characterized it as creating fundamental and unresolvable tension with the value of cooperation. See, e.g., Beckerman, supra note 6, at 511-12, 516-17. The values of comprehensiveness and cooperation certainly exist in uneasy tension. But they serve as important checks on each other, guarding against adversarial excess, on the one hand, and insufficient client protections, on the other.
costs, the Conference amended the Federal Rules to support the new overarching goal of proportionality. Going forward, discovery requests were to be proportional to the needs of the particular case. In 2006, the Conference again amended the Rules to adapt e-discovery to this overarching goal.  

This trend may be both necessary and wise, but predictive coding is currently taking it to a problematic extreme. Previously, proportionality entailed a substantive inquiry into the appropriateness and necessity of particular discovery requests in light of the case as a whole. Now, it entails agreement on the production of a particular percentage of documents at a particular level of accuracy, with overwhelming if not exclusive reference to projected costs. Sometimes parties reach agreement because of a judge’s order; other times they do so in private without oversight. Regardless, the court then produces a new form of court-produced statistical truth, based on the parties’ agreement as to the number and scope of documents that will be produced.

This new vision of proportionality achieves a high level of cooperation between judges and lawyers, but at a cost to the legitimacy of our adversarial system. If parties to a case agree to save costs by limiting discovery to ten percent of all relevant documents, courts could theoretically resolve the case based on a randomly-selected ten percent of relevant documents. Taken to an extreme, this creates a discovery lottery system. A smoking gun document may or may not be in that ten percent; a group of emails that together establish knowledge may or may not all be in that ten percent.

In arbitration, mediation, and other forms of private dispute resolution, parties are free to gamble for outcomes. But unlike private dispute resolution forums, courts serve public as well as private functions. In addition to resolving disputes, they articulate rules, establish precedents, and serve as visible symbols of the rule of law in society. These public functions require judicial legitimacy and credibility, which, in turn, require courts to have actual and perceived commitments to established procedures for truth-seeking and justice-promotion. These public func-


131 Da Silva Moore Panel, supra note 66, at p.4.


tions are undermined by a system that allows parties to secure the imprimatur of a judge and a court system for their limited view of the truth, based on a statistical sampling of the facts.

Some process-level compromise in the system might be, and likely, is, salutary. But shifts towards increased cooperation and transparency should be made with an awareness of the trade-offs in other system values. They should not be made out of blind faith in predictive coding as a fail-safe method of reducing costs and limiting adversarial excess.

Shifts towards cooperation and transparency should also be effectuated comprehensively rather than on a court-by-court and judge-by-judge basis. Right now, as some judges are uncritically adopting predictive coding, others remain resistant to its use or completely unaware of its existence. Given that discovery generally proceeds within the individual discretion of trial judges and is rarely subject to appellate review, litigants contemplating use of predictive coding face significant uncertainty. Uncertainty in the litigation system is always problematic, but the difficulties are magnified here, where the stakes entail not only the defensibility of using predictive coding to meet discovery obligations, but also potentially altered ethical standards.

III. THE PATH AHEAD

Given that the organized profession is a key stakeholder in the adoption and use of predictive coding technologies—one with professional duties to the country’s court systems and to the public at large—it should take a more systematic approach, with a critical eye towards the trade-offs involved. In this final Part, I propose that the profession take a proactive role in shaping and using predictive coding technologies, rather than deferring to the interests of vendors and trial judges. I suggest that the profession take action along four lines: 1) It should engage in . It should raise awareness and understanding within the legal community; work with IT professionals to establish minimum functionality standards and standardized implementing protocols; and ensure widespread access to the technology.

A. Education

As a preliminary step, the profession should raise awareness and understanding of predictive coding technologies. Although predictive coding has emerged as a core topic of concern in some circles, many judges and lawyers remain unaware of its existence or intimidated by its use.

---

134 See, e.g., Peck, supra note 17; Murphy, supra note 37.
This means that a relatively small sector of the profession exercises disproportionate influence over the technology’s development and use. It also produces significant uncertainty among litigants and lawyers as to whether adoption will be allowed or even required. The organized bar address this situation through proactive educational efforts in law schools, practice communities, and judiciaries.

Courts, meanwhile, should encourage and facilitate use of special masters. Many courts already use special masters to provide judges and parties with technological expertise and guidance, and to facilitate early resolution of e-discovery disputes. Some courts have established protocols for the selection of special masters and many commentators advocate increased reliance on them for e-discovery issues generally. These practices should be continued and expanded to address predictive coding in particular.

B. Quality Control

A second critical step will be for the profession to set minimum functionality standards for products and processes labeled “predictive coding.” Currently, countless vendors market products described as predictive coding. Some are indisputably mislabeled key word searching technologies. The others comprise a dizzying array of document classification technologies and implementing protocols that properly fall within the broad umbrella of predictive coding, but that have as many differences as commonalities. The bar needs to gather data in order to understand and evaluate the extent of this variation. To do so, it should commission a comprehensive third-party efficacy study, perhaps through the ABA or the Sedona Conference, which involves relevant stakeholders from within and outside of the profession.

135 See Daniel B. Garrie, Matrimonial Law Economics: Electronic Discovery and Change in Senior Partner’s Role, 27 J. FAM. L. 1, 3 (2013) (acknowledging that experts will be necessary to educate other legal professionals, but will not be necessary for day-to-day operational work); see also Barry, supra note 16, at 368 (“It will be the legal community’s responsibility to promote and educate the bench about these new technologies.”).


138 See supra note 65 and accompanying text.

139 Predictive Coding and Patented Workflow, supra note 4, at p.2-3.


141 See supra note 126.
The efficacy study should address both processes and results. Because different predictive coding tools are based on different text classifying systems, they are more or less suited to different types of documents and data sets. And because they have different features and functionalities, they are more or less effective at achieving high levels of recall and precision. Understanding both points of comparison—processes and results—will be critical.

The Legal Track Interactive Task of the Text Retrieval Conference, an annual simulation of civil litigation document review, has started important work along these lines. Each year, using a different set of documents in a different case, Legal Track compares the results of manual attorney review with review by various technology-assisted approaches. Legal Track’s project design is limited in a significant way, however. It is based on voluntary participation, overwhelmingly by particular vendors’ representatives. Thus, it represents a useful and important starting point in evaluating various products, but needs to be built upon and expanded.

C. Standardization

The profession should also develop standardized protocols for use. Currently, there is significant variety in implementing procedures, including different methods of populating the seed set, different requirements for statistical confidence in the computer’s search algorithm, different means of checking for quality control, and different approaches to privileged documents. District and magistrate judges enjoy a high degree of autonomy in requiring or influencing particular protocols, and they sometimes do so with insufficient attention to other implicated interests such as client protections. The achieve greater uniformity, the profession should initiate a participatory and deliberation-based process to design standardized protocols. As with a commission to study the efficacy of various products, a group to produce standardized protocols

---

142 Barry, supra note 16, at 368.
should include representatives from all relevant groups of stakeholders, including lawyers, judges, academics, computer scientists, and commercial vendors.

Protocols should address at least four core issues. First, they should address the threshold question of when litigants should be allowed and/or required to use predictive coding in the discovery process. Factors to consider include whether each side has access to predictive coding capabilities and whether the type of data at issue are suitable to predictive coding approaches. Second, standardized protocols should address appropriate procedures for populating and coding the seed set and training the computer. These procedures should seek to strike a desirable balance among the values of cooperation, transparency, and client protection. The third issue in need of standardization is the creation of appropriate and required quality control checks. The fourth concerns proper procedures for handling disclosures of privileged information, including whether, when, and how claw-back agreements are appropriate and should be used. Claw-back agreements stipulate that under certain conditions, inadvertent disclosures of privileged information will not constitute a waiver of privilege. They can be useful in coping with high volumes of privileged information in massive document productions, and will certainly be appropriate in some cases that employ predictive coding.

D. Access

Finally and perhaps most importantly, the profession must stay cognizant of its fundamental obligation to ensure access to effective legal services. Regardless of potential overall savings, predictive coding entails significant up-front costs that are prohibitive for some parties. Additionally and as noted, vendors are obtaining patents on their technologies with increasing frequency. In light of these impediments, the pro-

144 See Baron, supra note 56, at 30 (discussing the necessity for "reasonable forms or measures of quality throughout the e-discovery process, including sampling at different phases of the process"); Barry, supra note 16, at 368-69 (discussing the necessity for selective sampling at multiple points in the e-discovery process).

145 Fed. R. Evid. 502 (providing that the inadvertent disclosure of privileged information does not operate as a waiver if, among other things, the disclosing party took “reasonable steps” to prevent the disclosure); see also Zubulake III, at 290 n.81 ("[I]nadvertent disclosure of a privileged document does not constitute a waiver of privilege, that the privileged document should be returned (or there will be a certification that it has been deleted), and that any notes or copies will be destroyed or deleted. Ideally, an agreement or order should be obtained prior to any production.").

146 Zubulake III, at 289 ("[M]any parties to document-intensive litigation enter into so-called “claw-back” agreements that allow the parties to forego privilege review altogether in favor of an agreement to return inadvertently produced privileged documents.").
profession needs to explore possible means of ensuring widespread access to predictive coding technologies.

One option would be for the bar, perhaps through the ABA, to develop an open-source predictive coding tool. A number of open-source predictive analytics platforms already exist. The bar could build upon one of them to develop a tool suited to the needs and standards of the legal profession. Alternatively or additionally, universities and academic institutions—many of which are already in the process of developing open-source predictive analytics tools—could develop or adapt tools to meet the needs of the legal profession. Developing an open-source tool would be the most effective strategy for granting widespread access, but it would entail a significant drawback, however. Lawyers would have access to the tool, but not necessarily to technical support, which could feed into many of the problems discussed above regarding lawyers’ lack of technological understanding and expertise.

The bar should therefore consider other options as well. The ABA or other bar groups could contract with particular vendors to provide all bar members with a cost-effective predictive coding tool. Universities regularly pursue this approach with products such as word processors, antivirus programs, and photo editing software, obtaining significant discounts for faculty, staff, and students in exchange for guaranteed volumes of sale. This could be a promising path for predictive coding products, where vendors may be willing to offer significant discounts not only in exchange for the market exposure and the official bar support. The profession could also lobby for a compulsory licensing scheme, requiring vendors to license their technologies to lawyers in exchange for a set royalty.

Vendors may object strenuously to such proposals. They would be well-advised, however, to work with the profession to increase access. After all, the profession, and ultimately, the courts, hold a trump card—they can reject the use of one or more predictive coding products as an acceptable means of meeting discovery obligations.

CONCLUSION

Vendors and other proponents of predictive coding successfully advanced the new technology as an answer to the problems of civil discov-

---


148 Id.

They explained that it could address skyrocketing volume and expense of electronically-stored information, while also restoring trust in a system that had long been plagued by problems of excess and abuse.

Like all technologies, however, predictive coding tools are not an unmitigated good. They threaten to create new problems at the same time that they solve existing ones. Many judges and lawyers are problematically ignoring this, failing to recognize the ethical trade-offs of adoption and use. In doing so, they are precluding a role for the profession in the technology’s development. They are also threatening the scope of the profession’s jurisdiction and the integrity of the adversarial system.

The profession should undoubtedly look to predictive coding as a powerful and potentially beneficial tool. But it must do so with a critical eye and a firm commitment to employ this and all technologies in service of its goals and values. In sum, the profession must recognize that it has an ethical obligation to tend to the tools of its trade as much as to the conduct of its members.