The Uncertain Promise of Predictive Coding

Dana A. Remus[*]

> "Technology . . . is a queer thing; it brings you great gifts with one hand, and it stabs you in the back with the other."
>
> – C.P. Snow, 1971[1]

ABSTRACT: Increasingly, machine-learning technologies known as "predictive coding" are automating document review in discovery practice. Recent law school graduates may lament the impact on entry-level law hiring, but the litigation community is embracing the new technologies. Proponents contend that by replacing the unreliable and inconsistent discretion of lawyers with the mechanized objectivity of computers, predictive-coding technologies can solve both the practical problems of e-discovery and the deeper-seated problems of excess, abuse, and trust that have long plagued discovery practice.

In this Article, I advise caution in the adoption of predictive-coding technologies. These technologies hold unquestionable potential as a means of coping with unmanageable datasets, but they entail costs as well as benefits. I argue that if lawyers ignore these costs, they will unwittingly abdicate control to computer scientists and vendors, compromising the profession's jurisdiction and undermining lawyers' ability to serve clients and the judicial system. I conclude that the profession has an ethical obligation to explore the costs as well as the benefits of predictive coding, and to play a more active role in its design and use.

[*] Associate Professor of Law, University of North Carolina School of Law. For helpful comments on earlier drafts, I am grateful to David Ardia, Bernard Burke, John Coyle, William Marshall, Mary Mitchell, Richard Myers, Mary-Rose Papandrea, Richard Saver, Robert Smith, W. Bradley Wendel, and the participants of the 2013 Fordham Legal Ethics Schmooze. For outstanding research assistance, I am grateful to Christina Nasuti, Troy Shelton, and Alex Shiekman.

[1] Anthony Lewis, *Dear Scoop Jackson*, N.Y. TIMES, Mar. 15, 1971, at 37 (quoting C.P. Snow).

INTRODUCTION

Problems of adversarial excess and abuse have long plagued civil discovery. In 1976, a commission convened by Chief Justice Burger concluded that "[w]ild fishing expeditions" had become the norm, along with "[u]nnecessary intrusions into the privacy of the individual, high costs to the litigants, and correspondingly unfair use of the discovery process as a lever toward settlement."[2] Today, nearly forty years later, the ethical and practical problems of discovery practice have only worsened and all too often, lawyers appear to be part of the problem.

When the drafters of the Federal Rules of Civil Procedure instituted civil discovery in 1938, they envisioned a largely self-regulating system, entrusted to the sound professional judgment and discretion of lawyers.[3] Lawyers, they believed, would act as professionals in discovery practice, balancing their tripartite duties to clients, courts, and the public. But lawyers quickly proved themselves to be adversarial advocates first and foremost, prioritizing their clients' interests above all else. Hidden from public view and with limited accountability, they turned discovery practice into a new area of gamesmanship,[4] which, in turn, undermined trust in the court system and the legal profession.[5]

The advent of computer technology and the proliferation of electronically stored information layered a new set of problems on top of existing ones.[6] Companies faced new and expensive questions regarding document retention, preservation, and production. Litigants began "document dumping"—flooding opponents with unmanageable datasets to increase costs and decrease their chances of finding key documents. The unmanageable scope and extent of e-discovery offered new opportunities for abuse and quickly became a principal cause of increasing costs and delays in the court system.

[2] William H. Erickson, *The Pound Conference Recommendations: A Blueprint for the Justice System in the Twenty-First Century*, 76 F.R.D. 277, 288 (1978).

[3] *See* Stephen N. Subrin, *Fishing Expeditions Allowed: The Historical Background of the 1938 Federal Discovery Rules*, 39 B.C. L. REV. 691, 717 (1998) (quoting Edson R. Sunderland, *Improving the Administration of Civil Justice*, ANNALS AM. ACAD. POL. & SOC. SCI., May 1933, at 60, 76).

[4] *See* John S. Beckerman, *Confronting Civil Discovery's Fatal Flaws*, 84 MINN. L. REV. 505, 522 (2000) ("Moreover, to the extent that success in litigation depends on strategic informational advantage, discovery, contrary to its inventors' expectations, is the critical battlefield on which the war is waged.").

[5] *See id.*; John H. Beisner, *Discovering a Better Way: The Need for Effective Civil Litigation Reform*, 60 DUKE L.J. 547, 549 (2010); Louis Harris & Assocs, Inc., *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, 735–36 (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").

[6] *See* Beisner, *supra* note 5, 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.").

Against this backdrop, bar leaders and reformers are now advancing a new solution to the problems of discovery.[7] Rather than advocating trust in the profession, they are advocating trust in computers—more specifically, in machine-learning products referred to as "predictive coding."[8] Although predictive-coding technologies encompass significant variations, they share a common approach: after an initial training period, a computer generates a customized search algorithm for identifying responsive and privileged documents; the computer then uses the algorithm to code an entire document set for responsiveness and privilege, obviating the need for manual human review.[9]

Proponents frame predictive coding as a silver-bullet solution to the problems of discovery practice—not only the practical problems of scope and cost, but also the more vicious problems of excess, abuse, and trust that have long characterized discovery practice. They claim that by eliminating the time and inconsistency of human review, predictive coding can increase the accuracy and decrease the costs of document review; and by replacing human discretion with mechanized objectivity, it can minimize abuse and restore trust in the system.[10] Predictive coding holds far more potential, they argue, than continued efforts to regulate attorney conduct.

Since 2012, a handful of trial courts have accepted these arguments, officially endorsing predictive coding as a valid and promising means of meeting discovery obligations.[11] The litigation community followed this vanguard and embraced the new technology,[12] such that predictive coding is now the "hot topic" of discovery reform.[13] Most large law firms have in-house

[7] *See infra* notes 60–65 and accompanying text.

[8] *See* EDISCOVERY INSTITUTE SURVEY ON PREDICTIVE CODING 2 (2010), [hereinafter EDISCOVERY SURVEY], *available at* http://www.discovia.com/wp-content/uploads/2012/07/2010_EDI_PredictiveCodingSurvey.pdf.

[9] *See infra* notes 49–55 and accompanying text.

[10] *See infra* notes 59–61 and accompanying text.

[11] *See* Da Silva Moore v. Publicis Groupe, 287 F.R.D. 182 (S.D.N.Y. 2012), *adopted sub nom* Da Silva Moore v. Publicis Groupe SA, No. 11 Civ. 1279(ACL)(AJP), 2012 WL 1446534 (S.D.N.Y. Apr. 26, 2012); Case Management Order, *In re* Actos (Pioglitazon—Prods. Liab. Litig.), No. 6-11-md-2299, 2012 WL 3899669 (W.D. La. July 30, 2012); Order Approving the Use of Predictive Coding for Discovery, Global Aerospace, Inc. v. Landow Aviation, L.P., No. CL 61040, 2012 WL 1431215 (Va. Cir. Ct. Apr. 23, 2012); Order Granting Partial Summary Judgment, EORHB, Inc. v. HOA Holdings, LLC, No. 7409-VCL, 2012 WL 4896670 (Del. Ch. Oct. 15, 2012).

[12] Warwick Sharp et al., *Feedback from the Predictive Coding Trenches at LegalTech® 2013: Moving from "Is It Defensible?" to "What Are Best Practices?" in 12 Months*, METROPOLITAN CORP. COUNS., Mar. 2013, at 30, 30, *available at* http://www.metrocorpcounsel.com/pdf/2013/March/30.pdf.

[13] Andrew Peck, *Search, Forward: Will Manual Document Review and Keyword Searches Be Replaced by Computer-Assisted Coding?*, L. TECH. NEWS (Oct. 2011), http://law.duke.edu/sites/default/files/centers/judicialstudies/TAR_conference/Panel_1-Background_Paper.pdf.

predictive-coding specialists, and countless conferences and CLE classes are extolling its virtues.[14]

Perhaps we should not be surprised. In the past decade, advances in artificial learning (machine learning in particular), have infiltrated our lives. Our e-mail applications are peppered with customized advertisements, tailored to the content of our correspondences. Our cell phones, now voice activated, can respond to spontaneous questions that we pose. These developments have undoubted power and potential and, like predictive coding, have been widely embraced. But they are not unqualified goods. Technological advances have both costs and benefits, invariably creating new problems as they solve existing ones.

Accordingly, I argue in this Article that courts, lawyers, and commentators must proceed with deliberate care in the use and adoption of predictive-coding technologies. The legal profession must consider the ethical trade-offs of adoption and take an active role, alongside vendors and computer scientists, in directing the design and development of these evolving technologies.

I begin in Part I by describing how proponents of predictive coding successfully advanced the new technologies in discovery practice. Notwithstanding significant variation among predictive-coding products and unresolved questions regarding their use, proponents proceeded as if all predictive-coding technologies are of equivalent and unquestionable benefit. The litigation community is now doing the same.

In Part II, I contend that the profession's current approach to predictive coding is problematic, giving rise to three sets of dangers: (1) by ignoring outstanding and contested issues regarding the design and use of these technologies, lawyers are blinding themselves to significant variation in functionality and efficacy; (2) by deferring to the opinions of computer scientists and vendors, the bar is ceding jurisdiction to self-interested parties; and (3) by altering relevant ethical standards to facilitate the technologies' use, lawyers are weakening the protections and legitimacy of our adversarial system.

In Part III, I suggest specific ways in which the bar can address these dangers, reassert its interests, and take a more proactive role in guiding the development and use of predictive-coding products. Predictive coding will never be 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 vendors introduced their automated approach to document review in the early 2000s, but over a decade passed before their

[14] *See infra* notes 63–64, 82 and accompanying text.

products entered mainstream practice. In this Part, I describe the ways in which vendors and other advocates promoted the new technologies and eventually 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—the architects of our discovery system—believed that broad discovery and generalized pleadings would "secure the just, speedy, and inexpensive determination of every action and proceeding."[15] Broad discovery, they believed, would eliminate unfair surprises at trial, counteract the effects of wealth and power disparities, and ensure that cases were resolved on their merits and not on information asymmetries or pleading technicalities.[16]

Pursuant to their vision, lawyers as professionals would exercise sound judgment and professional discretion in administering the system.[17] Lawyers for requesting parties would balance duties to clients to obtain 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 clients to protect harmful and confidential information with duties to opponents to respond to legitimate discovery requests with relevant and non-privileged information and documents. The system would be largely self-policing, fueled by trust that lawyers would adhere to both the letter and the spirit of the discovery rules.[18]

Lawyers quickly proved themselves undeserving of that trust, however. Working largely out of view of opponents and the courts, they pushed aside their duties to the judicial system and the public and turned discovery practice into a new area of gamesmanship—an opportunity to exhaust the resources of an opponent and to gain a strategic advantage for a client.[19] Abuse, excess, and exorbitant costs became commonplace.[20]

[15] FED. R. CIV. P. 1; *see also* Subrin, *supra* note 3, at 717.

[16] *See* Thomas E. Willging et al., *An Empirical Study of Discovery and Disclosure Practice Under the 1993 Federal Rule Amendments*, 39 B.C. L. REV. 525, 527 (1998).

[17] Edward D. Cavanagh, *The August 1, 1983 Amendments to the Federal Rules of Civil Procedure: A Critical Evaluation and a Proposal for More Effective Discovery Through Local Rules*, 30 VILL. L. REV. 767, 775 & n.34 (1985).

[18] *Id.*

[19] Commentators explain these developments by reference to game theory: litigants inevitably worried that their opponents were using discovery too aggressively (for example, by taking extreme positions on relevancy and privilege determinations), or with improper motives (for example, by taking particular positions solely to impose heightened costs on an opponent or to obstruct discovery of relevant information). *See* John K. Setear, *The Barrister and the Bomb: The Dynamics of Cooperation, Nuclear Deterrence, and Discovery Abuse*, 69 B.U. L. REV. 569, 627 (1989); Charles Yablon, *Stupid Lawyer Tricks: An Essay on Discovery Abuse*, 96 COLUM. L. REV. 1618, 1622–23 (1996). 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

Early advocates for reform proposed increased judicial involvement as a means of reining in adversarial excess and restoring trust in the system. After a series of modest reforms to increase judicial involvement through the middle of the twentieth century,[21] rule-makers took more significant action in 1983.[22] They amended the Federal Rules to authorize and empower judges to limit discovery to that which was reasonable and proportional in light of the needs of each case.[23] Under the new rules, lawyers would retain discretion to design discovery requests in the first instance, but judges would acquire a new oversight role in determining whether those requests fit within the meaning of reasonable and proportionate discovery.[24]

The rule-makers had hoped that with a more significant role in discovery practice, judges could check attorney overreaching and ensure that discovery tools were being used to ensure the just and efficient resolution of cases, rather than as strategic weapons to wear down an opponent and force settlement.[25] The reforms did not live up to these hopes, however, and failed to restore trust in the system.[26] The proportionality determination rested on vague and hard-to-quantify factors,

opponent did not. *See* WILLIAM POUNDSTONE, PRISONER'S DILEMMA 116–21 (1992). Mutual distrust led to discovery abuse and escalating costs.

[20] Beckerman, *supra* note 4, at 522–23.

[21] *Id.* at 512.

[22] *See* Cavanagh, *supra* note 17, at 779–81.

[23] FED. R. CIV. P. 26(b)(2)(c) (limiting discovery where: "(i) the discovery sought is unreasonably cumulative or duplicative, or can be obtained from some other source that is more convenient, less burdensome, or less expensive; (ii) the party seeking discovery has had ample opportunity to obtain the information by discovery in the action; or (iii) the burden or expense of the proposed discovery outweighs its likely benefit, considering the needs of the case, the amount in controversy, the parties' resources, the importance of the issues at stake in the action, and the importance of the discovery in resolving the issues"); *see also* ARTHUR R. MILLER, FED. JUDICIAL CTR., THE AUGUST 1983 AMENDMENTS TO THE FEDERAL RULES OF CIVIL PROCEDURE: PROMOTING EFFECTIVE CASE MANAGEMENT AND LAWYER RESPONSIBILITY 32–34 (1984).

[24] Jordan M. Singer, *Proportionality's Cultural Foundation*, 52 SANTA CLARA L. REV. 145, 180 (2012) ("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.").

[25] *Id.* at 176–78; *see also* Wayne D. Brazil, *The Adversary Character of Civil Discovery: A Critique and Proposals for Change*, 31 VAND. L. REV. 1295, 1302–05 (1978); Robert F. Peckham, *A Judicial Response to the Cost of Litigation: Case Management, Two-Stage Discovery Planning and Alternative Dispute Resolution*, 37 RUTGERS L. REV. 253, 256 (1985) (arguing that judicial oversight was necessary to correct lawyers' "attempt[s] to manipulate the discovery rules to frustrate and subvert the opposing party"). *But see* Judith Resnik, *Managerial Judges*, 96 HARV. L. REV. 374, 417–31 (1982) (critiquing the increased managerial role of judges).

[26] John P. Frank, *The Rules of Civil Procedure—Agenda for Reform*, 137 U. PA. L. REV. 1883, 1891 (1989) ("[Rule 26(b)(1)(iii)] 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."); Singer, *supra* note 24, at 180–81 (describing the rule as "ineffective, seldom used, and [largely] ignored" (footnotes omitted) (internal quotation marks omitted)).

itself entailing significant unreviewable discretion.[27] Vastly inconsistent rulings and levels of involvement, rarely subject to appellate review, fueled continued perceptions that the system rested on unchecked professional discretion—albeit a combination of judge and attorney discretion.[28]

B. E-DISCOVERY

The advent of computer technology ushered in a new set of problems.[29] Computers exponentially increased the volume of documents produced in the ordinary course of business and compounded the burdens of discovery. Producing parties faced new and sometimes crushing costs when asked to produce data from difficult-to-access sources. Requesting parties faced the seemingly impossible task of reviewing millions of documents to find the needle in the haystack—the key document or documents on which the case might turn.[30] All parties faced new and costly questions of document retention and preservation.

Computers also offered new discovery tools and strategies, though not without controversy. New software allowed lawyers and litigants to eliminate duplicates and consolidate email chains. Keyword searching programs offered an efficient means of culling through unmanageable datasets.[31] But even as lawyers began adopting keyword searching and related programs, a significant and growing literature criticized these approaches for being 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).[32]

[27] *See* Singer, *supra* note 24, at 147.

[28] *Cf.* Resnik, *supra* note 25, at 424–31 (noting the potential costs of managerial judging).

[29] *See* Beisner, *supra* note 5, 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), *available at* http://jolt.richmond.edu/v19i/article2.pdf (noting that email alone produces 100 billion new messages daily).

[30] Beisner, *supra* note 5, at 550.

[31] *See* Robert C. Manlowe et al., *Paradigm Shifts in E-Discovery Litigation: Cooperate or Continue to Pay Dearly*, 78 DEF. COUNS. J. 170, 171 (2011).

[32] *See* Symposium, *The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery*, 8 SEDONA CONF. J. 189 (2007) [hereinafter *Sedona Conference*]; *see also* David C. Blair & M.E. Maron, *An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System*, 28 COMM. ACM 289, 293 (1985) (concluding that under a keyword-searching approach, up to 80% of the responsive documents in a collection may routinely be missed); Howard Sklar, *Match Point with Recommind's Predictive Coding—It's "Man with Machine," Not "Man vs. Machine,"* METROPOLITAN CORP. COUNS., Aug. 1, 2011, at 16, 16, *available at* http://www.metrocorpcounsel.com/pdf/2011/August/16.pdf (noting research showing that keyword searching leads to recall of about 50% at best, and likely closer 20% of relevant documents). Commentators have observed many reasons for this lack of accuracy. Lawyers are accustomed to searching in databases like Westlaw and Lexis, in which data is cleaned and primed for Boolean searches. They are ill-prepared to design effective keyword

The resulting uncertainty fueled existing distrust of the system. Requesting parties accused producing parties of designating inadequate search terms to exclude relevant documents and information.[33] Producing parties accused requesting parties of intentionally interfering with the production process to increase costs.[34] Lawyers on both sides accused judges of failing to understand and effectively manage electronic discovery.[35] Clients, meanwhile, worried that lawyers—virtually unsupervised amidst massive datasets—were extending discovery for the sole purpose of increasing billable hours.[36]

In 2003 and 2004, in a series of opinions in *Zubulake v. UBS Warburg LLC*, U.S. District Court Judge Shira Scheindlin set out to offer information and guidance regarding the challenges of e-discovery.[37] Among other things, she addressed skyrocketing costs, concluding that courts should only engage in a cost-shifting analysis if the data in question is relatively inaccessible—for example, if it is stored on back-up tapes.[38] She also set forth new duties for lawyers with respect to their clients' electronically stored information, including "tak[ing] affirmative steps to monitor compliance so that all sources of discoverable information are identified and searched."[39]

Judge Scheindlin's opinions in *Zubulake* spoke directly to countless lawyers and judges who were eager for clarity and certainty in a new era of

searches for disorganized collections of documents and data. Barry Murphy, *Is Predictive Coding the Future of Document Review?*, EDISCOVERY J. (Oct. 28, 2010, 11:56 PM), http://old.ediscoveryjournal.com/2010/10/is-predictive-coding-the-future-of-document-review/; Peck, *supra* note 13, at 26–27. Moreover, given the ambiguity of language, even the most advanced linguists cannot design searches that capture all words or phrases used to refer to a particular subject. Manlowe et al., *supra* note 31, at 171; *Sedona Conference*, *supra*, at 201–02; *see also* Victor Stanley, Inc. v. Creative Pipe, Inc., 250 F.R.D. 251, 256–57 (D. Md. 2008); United States v. O'Keefe, 537 F. Supp. 2d 14, 24 (D.D.C. 2008).

[33] Harrison M. Brown, Note, *Searching for an Answer: Defensible E-Discovery Search Techniques in the Absence of Judicial Voice*, 16 CHAP. L. REV. 407, 420–21 (2013).

[34] *See id.*

[35] *See* Singer, *supra* note 24, at 178 (discussing the need for increased judicial oversight).

[36] *See* William W. Schwarzer, *Slaying the Monsters of Cost and Delay: Would Disclosure Be More Effective than Discovery?*, 74 JUDICATURE 178, 178–79 (1991); *see also* Singer, *supra* note 24, at 176–77.

[37] Zubulake v. UBS Warburg LLC (*Zubulake V*), 229 F.R.D. 422 (S.D.N.Y. 2004); Zubulake v. UBS Warburg LLC (*Zubulake IV*), 220 F.R.D. 212 (S.D.N.Y. 2003); Zubulake v. UBS Warburg LLC (*Zubulake III*), 216 F.R.D. 280 (S.D.N.Y. 2003); Zubulake v. UBS Warburg LLC (*Zubulake I*), 217 F.R.D. 309 (S.D.N.Y. 2003).

[38] *Zubulake I*, 217 F.R.D. at 320–23. Judge Scheindlin explained that a key factor in evaluating whether costs were "undue" under Federal Rule of Civil Procedure 26(b)(2)(iii), and therefore warranted 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.* She also explained that the cost-shifting, if appropriate, should occur after the documents and data have been produced so that actual costs can be evaluated. *Id.*

[39] *Zubulake V*, 229 F.R.D. at 432 (explaining that attorneys are obligated to take affirmative steps to ensure that clients retain, identify, and produce relevant electronically stored information).

discovery. Although the opinions expressed the views of a single district court judge, they were read and received as setting forth definitive standards for e-discovery. Numerous federal district courts[40] and state trial and appellate courts[41] followed them, and countless secondary sources and practice guides cited them.[42]

The *Zubulake* opinions also set the stage for the December 2006 e-discovery amendments to the Federal Rules of Civil Procedure.[43] The new rules designate reasonably accessible, electronically stored information as presumptively discoverable, while requiring a showing of good cause for discovery of difficult-to-access sources.[44] To establish good cause, a party must persuade a judge that the burden and costs of production outweigh the likely benefits for the case.[45]

The rule changes proved helpful in clarifying issues particular to e-discovery, but as with prior reform efforts, they failed to restore trust in the system.[46] Discovery continued to rest on largely unchecked professional

[40] *See, e.g.*, Williams v. N.Y.C. Transit Auth., No. 10 CV 0882(ENV), 2011 WL 5024280, at *4 (E.D.N.Y. Oct. 19, 2011); Essenter v. Cumberland Farms, Inc., No. 1:09-CV-0539 (LEK/DRH), 2011 WL 124505, at *3 (N.D.N.Y. Jan. 14, 2011); Passlogix, Inc. v. 2FA Tech., LLC, 708 F. Supp. 2d 378, 409 (S.D.N.Y. 2010).

[41] *See, e.g.*, Howard Reg'l Health Sys. v. Gordon, 952 N.E.2d 182, 189 (Ind. 2011); Voom HD Holdings LLC v. Echostar Satellite LLC, 939 N.Y.S.2d 321, 324 (N.Y. App. Div. 2012) (upholding the trial court's use of the *Zubulake* standard).

[42] *See, e.g.*, Michael W. Deyo, *Deconstructing* Pension Committee*: The Evolving Rules of Evidence Spoliation and Sanctions in the Electronic Discovery Era*, 75 ALB. L. REV. 305, 306 (2011–2012) ("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.").

[43] *Id.* at 305–06.

[44] FED. R. CIV. P. 26(b)(2)(B) ("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."). *See generally* Kenneth J. Withers, *Electronically Stored Information: The December 2006 Amendments to the Federal Rules of Civil Procedure*, 4 NW. J. TECH. & INTELL. PROP. 171, 176–201 (2006); *Materials on E-Discovery*, FED. JUD. CENTER, http://www.fjc.gov/public/home.nsf/autoframe?openform&url_l=/public/home.nsf/inavgeneral?openopen&url_r=/public/home.nsf/pages/196 (last visited Mar. 23, 2014); *Publications*, SEDONA CONF., https://thesedonaconference.org/publications (last visited Mar. 23, 2014) (referencing section on eDiscovery).

[45] *See, e.g.*, Capitol Records, Inc. v. MP3tunes, LLC, 261 F.R.D. 44, 51–52 (S.D.N.Y. 2009) (explaining the good cause standard required under FED. R. CIV. P. 26(b)(2)(B)).

[46] *See, e.g.*, Beisner, *supra* note 5, at 582–84 (citing lack of guidance for parties, rising costs, and continued abuse of the discovery system as significant problems under the amended rules); Robert Hardaway et al., *E-Discovery's Threat to Civil Litigation: Reevaluating Rule 26 for the Digital Age*, 63 RUTGERS L. REV. 521, 522 (2011) (arguing that the amended rules "fail to contain the cost or scope of discovery . . . [and] encourage expensive litigation"); Tonia Hap Murphy, *Mandating Use of Predictive Coding in Electronic Discovery: An Ill-Advised Judicial Intrusion*,

discretion, and instances and perceptions of excess and abuse continued.[47] For many, the natural conclusion was that lawyers, judges, or both were failing to exercise their discretion and judgment wisely.

C. Predictive Coding

In the first decade of this millennium, a growing chorus from inside and outside the profession began advancing a new type of reform. Abandoning their efforts to restore trust in lawyers' and judges' professional discretion—a task that had proven impossible—they advocated trust in automated discovery programs called predictive coding. They claimed that by replacing the subjectivity and discretion of human review with the mechanical objectivity of computer technology, predictive coding could provide answers to all of the problems plaguing discovery practice.[48]

"Predictive coding" encompasses significant variation in underlying technologies and implementing procedures.[49] Generally speaking, the term refers to automated approaches to discovery that employ machine learning

50 AM. BUS. L.J. 609, 639–40 (2013) (warning against increased judicial intervention in e-discovery matters under the revised rules).

[47] *See generally* Beisner, *supra* note 5, at 563–73.

[48] *Cf.* THEODORE M. PORTER, TRUST IN NUMBERS: THE PURSUIT OF OBJECTIVITY IN SCIENCE AND PUBLIC LIFE 90 (1995) (observing that, in contexts where "subjective discretion has become suspect . . . [m]echanical objectivity serves as an alternative to personal trust").

[49] Lauren Sylvester, *What Lawyers and Judges Need to Know About Machine Learning for Complex eDiscovery*, RATIONAL ENTERPRISE (Mar. 20, 2013), http://www.rationalenterprise.com/resources/blog/what-lawyers-and-judges-need-to-know-about-machine-learning-for-complex-ediscovery. 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, *Electronic Discovery: A Special Report: What Lawyers Need to Know About Search Tools*, NAT. L.J., Aug. 23, 2010, *available at* http://www.ned.uscourts.gov/internetDocs/cle/2011-01/National%20Law%20Journal%20(Aug%202010).pdf. 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), *available at* http://jolt.richmond.edu/v17i3/article9.pdf. 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 is also subject to human review. Brendan M. Schulman & Samantha V. Ettari, *Federal Court Approves the Use of "Predictive Coding" Technology-Assisted Document Review*, METROPOLITAN CORP. COUNS., May 2012, *available at* http://www.kramerlevin.com/files/Publication/96389d11-a0d2-4a70-9ac7-545b15e70261/Presentation/PublicationAttachment/746b60aa-a962-4b76-9204-55f188503095/MetCC_May%202012.pdf.

in document review.[50] Under a typical protocol, one or more senior lawyers code a sample set of documents (the seed set) for responsiveness and privilege.[51] Based on the initial coding, a computer generates a search algorithm for identifying responsive and privileged documents.[52] The computer applies the algorithm in coding a second sample set for responsiveness and privilege, which the lawyers then review and correct.[53] This iterative 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.[54] The computer then reviews and codes the entire dataset.[55]

When the first predictive-coding products were introduced in the early years of the new millennium, they attracted little attention. The landscape changed in 2010, however, with the publication of two pilot studies that favorably compared predictive coding to manual document review.[56] Based on simulated discovery exercises, both studies concluded that the 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)[57] than manual review.[58]

[50] *See* EDISCOVERY SURVEY, *supra* note 8; *see also* Charles Yablon & Nick Landsman-Roos, *Predictive Coding: Emerging Questions and Concerns*, 64 S.C. L. REV. 633, 637–48 (2013).

[51] For an overview of the functioning of predictive-coding procedures generally, see Schulman & Ettari, *supra* note 49; *Predictive Coding and Patented Workflow: A Defensible E-Discovery System*, METROPOLITAN CORP. COUNS., Apr. 2012, at 16, 16, *available at* http://www.metrocorpcounsel.com/pdf/2012/April/16.pdf [hereinafter *Predictive Coding and Patented Workflow*] (interview with Howard Sklar, Senior Counsel at Recommind, Inc.).

[52] Yablon & Landsman-Roos, *supra* note 50, at 639.

[53] *Id.* at 639–40.

[54] *Id.* at 640.

[55] *Id.* at 640–41. Accordingly, predictive coding differs in significant ways from keyword searching, which culls through a document set and produces a smaller set for human review. *Id.* at 652.

[56] Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, 17 RICH. J.L. & TECH. 1, 3 (2011), *available at* http://jolt.richmond.edu/v17i3/article11.pdf; Herbert L. Roitblat et al., *Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review*, 61 J. AM. SOC'Y FOR INFO. SCI. & TECH. 70, 70, 74–75 (2010). For a discussion of these studies, see *infra* notes 58, 86. *But see* Roitblat et al., *supra* note 56, 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.").

[57] *See* Grossman & Cormack, *supra* note 56, at 8 ("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." (footnotes omitted)).

[58] The Grossman and Cormack article has proven the most influential. *See id.* Maura Grossman is a litigator at Wachtell, Lipton, Rosen & Katz, 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

These studies provided vendors and other proponents with the empirical support they needed to make a persuasive case for predictive coding. Citing the studies' findings and conclusions, they characterized the existing system of manual review as inefficient, expensive, and rife with problems of human error.[59] They characterized predictive coding, in contrast, as offering the mechanical objectivity and accuracy of a computer program[60] at a fraction of the cost of manual review.[61]

Support for predictive coding grew in pockets.[62] A number of corporations brought predictive-coding products in-house for data management purposes.[63] With the help of their lawyers, they began using

Technology. *See infra* note 144 and accompanying text. 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. Grossman & Cormack, *supra* note 56, at 44. Moreover, they claimed that the predictive coding review process was significantly less expensive than that of the teams relying exclusively on human review. *Id.* 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. *Id.* at 25–29. But many advocates of predictive coding—lawyers, judges, and vendors alike—ignored these qualifications and began using the study to argue for widespread adoption.

[59] Proponents 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, as a result not 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 R. Grossman & Gordon V. Cormack, *Inconsistent Responsiveness Determination in Document Review: Difference of Opinion or Human Error?*, 32 PACE L. REV. 267 (2012).

[60] *See, e.g.*, *Predictive Coding and Technology-Assisted Review After* Da Silva Moore, METROPOLITAN CORP. COUNS., July–Aug. 2012, at 39, 39, *available at* http://www.metrocorpcounsel.com/pdf/2012/July/39.pdf (interview with Skip Durocher and Caroline Boudreau Sweeney of Dorsey & Whitney LLP) (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 51 (arguing that the technology would even increase the accuracy and reliability of supervisory lawyers because "[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").

[61] *See, e.g.*, *Evidence Mounting in the Case for Predictive Coding*, METROPOLITAN CORP. COUNS., Oct. 2012, at 29, 29, *available at* http://www.metrocorpcounsel.com/pdf/2012/October/29.pdf (interview with William Tolson, Senior Product Manager, Recommind, Inc.); Howard Sklar & Michael Potters, *Getting It Right: Training and Certification in Predictive Coding*, METROPOLITAN CORP. COUNS., Oct. 2012, at 32, 32, *available at* http://www.metrocorpcounsel.com/pdf/2012/October/32.pdf; Sylvester, *supra* note 49, at 3.

[62] *See Got Technology-Assisted Review? A Short List of Providers and Terms*, COMPLEX DISCOVERY (July 16, 2013), http://www.complexdiscovery.com/info/2013/01/26/got-technology-assisted-review-a-short-list-of-providers-and-terms/.

[63] *Panel Discussion: Judge Peck,* Da Silva Moore *and the Outlook for Predictive Coding*, METROPOLITAN CORP. COUNS., June 2012, at 8, *available at* http://www.metrocorpcounsel.com/pdf/2012/June/08.pdf [hereinafter *Panel Discussion*]; *Revolutionizing eDiscovery with Predictive Coding*,

the new technologies to cull through documents produced by an opponent.[64] Without informing the court or the opponent, some lawyers and litigants used predictive coding in document productions. Most lawyers, however, remained unaware of the new technologies or reluctant to use them. They "observ[ed] from the sidelines," waiting for judicial endorsement of predictive-coding tools as a valid means of meeting discovery obligations.[65]

In October 2011, they received significant assurance. Magistrate Judge Andrew Peck published a bar journal article recommending predictive coding as a vital tool for effective and efficient discovery.[66] Relying on the 2010 empirical pilot studies,[67] he responded to the bar's uncertainty by writing, "Until there is a judicial opinion approving (or even critiquing) the use of predictive coding, counsel will just have to rely on this article as a sign of judicial approval."[68]

Six months later, Judge Peck had the opportunity to do more. In *Da Silva Moore v. Publicis Groupe*, he was asked to resolve a discovery dispute regarding the use of predictive coding.[69] Uncritically accepting the claims and conclusions of the 2010 empirical studies, he reasoned that the accuracy of the technology had been established,[70] and that "it [was] the process used and the interaction of man and machine that the courts need[] to examine."[71] He therefore focused on implementing protocols in the case before him.[72] Believing that heightened—perhaps mandatory—cooperation was critical, he ordered the producing party to be fully transparent in coding the seed set and training the computer.[73] Then, addressing lawyers generally, he announced that "[c]omputer-assisted review now can be considered judicially-approved for use in appropriate cases."[74]

METROPOLITAN CORP. COUNS., July 2011, at 19, 19, *available at* http://www.metrocorpcounsel.com/pdf/2011/July/19.pdf.

[64] *See Revolutionizing eDiscovery with Predictive Coding*, *supra* note 63.

[65] *Panel Discussion*, *supra* note 63, at 8.

[66] Peck, *supra* note 13. Judge Peck also expressed support for predictive coding as a speaker at various e-discovery conferences. *See* Schulman & Ettari, *supra* note 49.

[67] Peck, *supra* note 13.

[68] *Id.*

[69] Da Silva Moore v. Publicis Groupe, 287 F.R.D. 182, 185 (S.D.N.Y. 2012) (Judge Peck noting "an article [he wrote] in the October Law Technology News called Search Forward, which says predictive coding should be used in the appropriate case"), *adopted sub nom.* Da Silva Moore v. Publicis Groupe SA, No. 11 Civ. 1279(ACL)(AJP), 2012 WL 1446534 (S.D.N.Y. Apr. 26, 2012).

[70] *Id.* ("[W]hile some lawyers still consider manual review to be the 'gold standard,' that is a myth, as statistics clearly show that computerized searches are at least as accurate, if not more so, than manual review.").

[71] *Id.* at 189.

[72] *Id.* at 190–91.

[73] *Id.* at 192.

[74] *Id.* at 193.

Judge Peck's decision, which was affirmed by the district court,[75] was soon thereafter described as a "watershed moment" that "completely mobilized the industry."[76] Within weeks, it was followed by two additional judicial opinions approving the use of predictive coding.[77] Those, in turn, were followed by an explosion of conferences, CLE classes, and proposed rule changes, all extolling the virtues of predictive coding.[78]

As Judge Peck had hoped, the conversation quickly shifted from the question of whether to use the technology to the question of how to use it.[79] 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."[80] Expressing these preferences in *Da Silva Moore,* he had explicitly directed the parties' attention away from the underlying technology and towards implementing procedures.[81] Taking his cue, other judges, lawyers, and litigants accepted the functionality of the underlying technologies and turned their focus to designing efficient and effective coding protocols. Vendors, eager for widespread acceptance, enthusiastically endorsed this approach through articles and panels.[82]

Da Silva Moore marked a turning point in the profession's acceptance of predictive coding. Much like Judge Schiendlin's opinions in *Zubulake,* the opinion gained influence that far outstripped its status as a single opinion

[75] Da Silva Moore v. Publicis Groupe SA, No. 11 Civ. 1279(ALC)(AJP), 2012 WL 1446534 (S.D.N.Y. April 26, 2012).

[76] *Panel Discussion, supra* note 63, at 8 ("[I]t's hard to remember an event in the history of e-discovery that has generated so much excitement and so completely mobilized the industry.").

[77] Case Management Order, *In re* Actos (Pioglitazone—Prods. Liab. Litig.), No. 6-11-md-2299, 2012 WL 3899669 (W.D. La. July 27, 2012); Order Approving the Use of Predictive Coding for Discovery, Global Aerospace, Inc. v. Landow Aviation, L.P., No. CL 61040, 2012 WL 1431215 (Va. Cir. Ct. Apr. 23, 2012); *see also* Nat'l Day Laborer Org. Network v. U.S. Immigration & Customs Enforcement Agency, 877 F. Supp. 2d 87, 109 (S.D.N.Y. 2012) (noting that "beyond the use of keyword search, parties can (and frequently should) rely on latent semantic indexing, statistical probability models, and machine learning tools to find responsive documents."); Transcript of Record at 66, EORHB, Inc. v. HOA Holdings LLC, C.A., No. 7409-VCL, 2013 WL 1960621 (Del. Ch. May 6, 2013) (Judge Travis Laster's statement from the bench: "This seems to me to be an ideal non-expedited case in which the parties would benefit from using predictive coding. I would like you all, if you do not want to use predictive coding, to show cause why this is not a case where predictive coding is the way to go.").

[78] *See, e.g.,* Joe Looby & Ari Kaplan, *Advice from Counsel: Can Predictive Coding Deliver on Its Promise?,* METROPOLITAN CORP. COUNS., Jan. 2013, at 30, 30, *available at* http://www.metrocorpcounsel.com/pdf/2013/January/30.pdf; Sklar & Potters, *supra* note 61, at 32; *Predictive Coding and Patented Workflow, supra* note 51, at 16; *Panel Discussion, supra* note 63.

[79] Sharp et al., *supra* note 12, at 30 ("With the acceptance of predictive coding technology by the courts, the market has transitioned very quickly from the 'whether to' question to the 'how to' question.").

[80] Peck, *supra* note 13.

[81] *See supra* notes 69–76 and accompanying text.

[82] *See Panel Discussion, supra* note 63.

from a magistrate judge. It offered the official support that many parties had been waiting for at a time when they were primed and ready to receive it.[83] When the opinion issued, computer scientists had already offered evidence of the accuracy of certain predictive-coding products, and 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.

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 discovery tool, it also creates risks and costs for the profession and the public. In this Part, I argue that ignoring these risks and costs gives rise to three sets of problems: (1) It elides significant variation in the definition and use of predictive-coding technologies; (2) it threatens the scope of the profession's jurisdiction; and (3) it undermines the protections and integrity of the adversarial system.

A. Variation and Uncertainty

The term "predictive coding" encompasses a countless variety of vendors and technologies with very different types of functionality. By one count, "[t]here are currently over 30 different types of classifiers capable of machine-learning-based text classification," many of which support several predictive-coding products for document review.[84] Given that different classifiers are optimally suited for different types of data, the choice of product can have significant implications for the quality of results.[85] The 2010 empirical pilot studies—at the foundation of the current widespread acceptance of predictive coding—looked at only three products, leaving the vast majority untested.[86] Nevertheless, the litigation community is

[83] *Id.* at 8 (noting that the opinion offered comfort to those who "have been observing from the sidelines, waiting for a federal judicial opinion to provide greater certainty about the defensibility of predictive coding"); *Predictive Coding and Patented Workflow*, *supra* note 51, at 16 ("We've maintained for a long time that a validating case, such as *Moore*, would be a great asset . . . .").

[84] Sylvester, *supra* note 49.

[85] *Id.* ("[W]hen the classifier and data are misaligned, results can be less accurate than even manual review.").

[86] Roitblat et al., *supra* note 56. Herbert Roitblat, Anne Kershaw, and Patrick Oot studied the level of agreement between four teams reviewing the same set of documents. *Id.* Two of the teams employed manual review and two employed technology-assisted review. *Id.* They concluded that the teams employing predictive coding achieved about the same recall as the teams employing manual review while achieving somewhat better precision. *Id.* Unfortunately, the published study fails to describe the protocols employed by any of the four teams. Maura

uncritically embracing predictive coding as if its definition is unitary and clear, its accuracy and efficacy well-established. As one lawyer recently explained, "We 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 . . . ."[87]

Promises of reduced discovery expense and abuse are no more certain than claims of increased accuracy.[88] Claims regarding reduced abuse focus on the elimination of lawyers, and therefore of their sometimes improper motives, from the process. But lawyers who train the computer systems can continue to make aggressive and even abusive relevancy or privilege determinations in coding the seed set, which will then be applied to the entire document set.

As for claims of reduced expense, existing studies calculate the difference between the costs of manual review and computer review.[89] They fail to account for the possibility of increased costs elsewhere, which may be substantial. Parties can (and likely will) employ experts to fight over coding protocols. Parties may also feel justified asking for and producing exponentially increased numbers of documents.

Notwithstanding the uncertainty and variation, vendors and a handful of judges have campaigned aggressively to frame predictive coding as a predictable, unitary, and valuable tool. For vendors, a story of technological closure[90] is beneficial in increasing demand and therefore profitability. For

Grossman and Gordon Cormack studied two different computer-assisted approaches employed by teams participating in the TREC 2009 Legal Track. Grossman & Cormack *supra* note 56, at 2. One, the method of the vendor H5, was a sophisticated form of keyword 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* PROCEEDINGS OF THE 2009 IEEE INTERNATIONAL CONFERENCE ON SYSTEMS, MAN, & CYBERNETICS 91, 91–96 (2009). The other was a predictive coding approach relying on machine learning. *See* Grossman & Cormack, *supra* note 56, at 35–37. The study concluded that both approaches compared favorably to manual review on the same document set. *Id.* at 43–44. However, because the two computer-assisted approaches were used on different document sets, they could not be compared to each other.

[87] *Panel Discussion*, *supra* note 63, at 9 (statement of Crowley).

[88] Looby & Kaplan, *supra* note 78, at 30.

[89] *See, e.g.*, *Evidence Mounting in the Case for Predictive Coding*, *supra* note 61, at 29 (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"); Looby & Kaplan, *supra* note 78, at 30 (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.* ("In certain circumstances, predictive coding can cost more.").

[90] Closure is achieved when a particular technology becomes widely accepted as the solution to a given problem such that all other potential solutions fall away. *See* Trevor J. Pinch & Wiebe E. Bijker, *The Social Construction of Facts and Artefacts: Or How the Sociology of Science and the Sociology of Technology Might Benefit Each Other*, 14 SOC. STUD. SCI. 399, 426–27 (1984).

trial judges interested in efficient case management, it can help limit discovery disputes and encourage cooperation. But this story is incomplete and dangerous. It blinds individual lawyers to the choices they face and the complexity of the tools they employ. It blinds the organized bar to a pressing need for studies evaluating and comparing various predictive-coding products. Finally, it blinds the bench and bar alike to the importance of proactive efforts to ensure that the new technologies are used to serve clients' interests and the public interest, rather than the interests of managerial judges and commercial vendors.

B. The Profession's Jurisdiction

Predictive-coding technologies are also eroding the profession's jurisdiction, giving rise to a second, related set of problems. Most obviously, they are raising questions of unauthorized legal practice by replacing lawyers with machines. In addition, they are undermining control over court processes, and allowing for the patenting of law-practicing algorithms. In all of these ways, predictive-coding technologies are threatening to interfere with the profession's ability to meet its public obligations.

1. The Unauthorized Practice of Law

Under a predictive-coding approach to discovery, tasks that once fell within the exclusive domain of lawyers are now delegated to distributed networks of computers, lawyers, and technology specialists.[91] The shift raises new questions regarding what constitutes the unauthorized practice of law—questions that are not readily answerable in the current framework of unauthorized practice rules.[92] Unlike prior technological developments, predictive-coding technologies do not implicate the traditional justifications for these rules—protecting the public from incompetent and unethical service providers. They do, however, raise new and troubling ethical questions regarding the extent of lawyers' duties to understand and supervise legal work.

The "deskilling"[93] of aspects of legal practice was similarly implicated by the introduction of computer programs that generated basic legal instruments for their users (such as Quicken Family Lawyer in the 1990s). Software companies explicitly marketed these programs as cost-effective alternatives to lawyers. State bars responded forcefully, characterizing the software as threatening significant consumer harm by supplanting lawyers'

[91] *See* Sklar & Potters, *supra* note 61.

[92] *See* Dana A. Remus, *Out of Practice: The Twenty-First Century Legal Profession*, 63 DUKE L.J. (forthcoming 2014), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2344888.

[93] HARRY BRAVERMAN, LABOR AND MONOPOLY CAPITAL: THE DEGRADATION OF WORK IN THE TWENTIETH CENTURY 3–30 (1974) (introducing the term "deskilling" in arguing that capitalism has degraded the workforce).

individualized guidance.[94] In a handful of states, courts agreed and deemed the software programs the unauthorized practice of law.[95]

In contrast to legal-forms software, predictive-coding technologies are used as lawyers' tools (not lawyer substitutes). Clients' interactions with these technologies are mediated by lawyers who train the computers, monitor the protocols, and ensure quality control. As a result, predictive-coding technologies implicate far fewer consumer-protection concerns than technologies that clients interact with directly, which replace a lawyer's advice entirely.[96]

In light of this, a more useful analogy may be the outsourcing of document review to off-shore document processing firms, where legal work is supervised but not performed by lawyers licensed in U.S. jurisdictions. The American Bar Association ("ABA") and several state commissions have concluded that if licensed U.S. attorneys retain strict supervisory roles, this type of outsourcing does not constitute the unauthorized practice of law.[97]

The analogy is again imperfect, however. There is a significant difference between delegating legal work to lawyers who are not licensed in the U.S. and delegating to non-lawyers—in this case, non-humans. Whereas supervising attorneys will typically have the requisite expertise to understand and evaluate the work of legal and paralegal staff, they may lack the requisite

[94] *See, e.g.*, William H. Brown, Comment, *Legal Software and the Unauthorized Practice of Law: Protection or Protectionism*, 36 CAL. W. L. REV. 157, 162 (1999) (summarizing concerns, which include that "materials may give advice that is incorrect or misleading" and that "a layperson is not subject to a malpractice suit or discipline by the state bar"); William A. Scott, Comment, *Filling in the Blanks: How Computerized Forms Are Affecting 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 last will and testament. In these circumstances, the true harm caused by these documents would likely not be realized until the testator dies and the forms reach the surrogate's court for probate.").

[95] *See, e.g.*, Unauthorized Practice of Law Comm. v. Parsons Tech. Inc., No. Civ.A 3:97CV-2859H, 1999 WL 47235, at *11 (N.D. Tex. Jan. 22, 1999), *superseded by statute*, An Act Relating to the Definition of the Practice of Law, 1999 Tex. Sess. Law Serv. 799 (West), *as recognized in* 179 F.3d 956 (5th Cir. 1999); *see also* Catherine J. Lanctot, *Scriveners in Cyberspace: Online Document Preparation and the Unauthorized Practice of Law*, 30 HOFSTRA L. REV. 811, 821 (2002) ("There is ample legal precedent to permit the conclusion that many online document providers are engaged in the unauthorized practice of law.").

[96] The principal justification prohibiting unauthorized practice of law is "to protect the public from the consequences of receiving legal services from unqualified persons." MODEL RULES OF PROF'L CONDUCT, R. 5.5 annot., at 458 (2007); *see also id.* ("The proscriptions also facilitate regulation of the legal profession and protect the integrity of the judicial system.").

[97] ABA Comm. on Ethics & Prof'l Responsibility, Formal Op. 08-451 (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* Prof'l Ethics of the Fla. Bar, Op. 07-2 (2008) (approving of off-shore outsourcing); The Bar of the City of N.Y. Comm. on Prof'l and Judicial Ethics of the Assoc., Formal Op. 2006-3 (2006), *available at* http://www2.nycbar.org/Ethics/eth2006.htm (a lawyer may outsource legal support services to overseas lawyers and non-lawyers if the lawyer supervises the work rigorously).

training and expertise to understand predictive-coding technologies. They may therefore be unable to assess whether a particular technology is adequate for a particular task and whether it is working properly when employed.

Indeed, lawyers have long been criticized for low technological literacy. With regard to predictive coding in particular, judges and lawyers alike lack clear understandings of which computer software programs and processes constitute predictive coding, how those programs work, and what various accuracy levels mean. This low level of technological knowledge and competency 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.[98] 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 perform[ed]"[99]—to indicate a lowered standard of oversight.

Through these new provisions, the ABA abdicates to clients and other professionals a portion of the profession's supervisory responsibilities over discovery practice. This, in turn, makes predictive coding ethically riskier (at least in some ways) than either legal-forms software or offshore outsourcing. Although lawyers will continue to be involved (obviating the consumer protection concerns that arise with forms statutes and that generally animate unauthorized practice of law statutes), they may fail to understand and take responsibility for the discovery tools they employ. As a result, clients will lack the protections of strict supervision that they have with outsourcing. The ramifications may be severe. Lawyers' ignorance of discovery tools may compromise the competency of representation, risk an unintentional breach of confidentiality, and, as discussed next, raise jurisdictional issues in the courtroom.

2. Control over Court Processes

In addition to raising new questions about the unauthorized practice of law, predictive coding is encroaching on the profession's control in the courtroom. It is transforming litigation procedure—traditionally the

[98] MODEL RULES OF PROF'L CONDUCT R. 5.3 cmt. 4 (2013).

[99] John M. Barkett, More on the Ethics of E-Discovery: Predictive Coding and Other Forms of Computer-Assisted Review 30 (2012) (unpublished manuscript) (second alteration in original) (quoting ABA, REPORT TO THE HOUSE OF DELEGATES 8 (2012), *available at* http://www.americanbar.org/content/dam/aba/administrative/ethics_2020/2012_hod_annual_meeting_105c_filed_may_2012.authcheckdam.pdf), *available at* http://law.duke.edu/sites/default/files/centers/judicialstudies/TAR_conference/Panel_5-Original_Paper.pdf.

exclusive domain of judges and lawyers—into a domain that is shared with computer scientists, commercial vendors, and others. If lawyers and judges were collaborating with IT professionals in this shared domain, this would not be problematic. But instead of collaborating, they are entirely deferring to the IT professionals. In doing so, judges and lawyers are privileging the values of commercial vendors over those of the legal profession and the court system.

The lack of technological expertise just discussed is at the root of the problem. To compensate, lawyers and judges are looking to non-lawyer IT experts for guidance on which predictive-coding products to use and how to use them.[100] 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.[101] IT experts are opining, therefore, not only on the quality of particular technologies, but also on the best ways to use them in discovery. They are answering such questions as how to populate the seed set, train the computer, and check for quality control.

In deferring to these non-lawyer experts, judges and lawyers are ceding control over what has long been considered a core domain of legal work—litigation procedure.[102] Moreover, they are ceding control to highly interested parties, whose values and goals diverge significantly from those of the profession. These non-lawyer experts likely hold a worldview that prioritizes technological use and development above all else. They have no reason to recognize, much less incorporate within their opinions, lawyers' ethical obligations to clients, the courts, and the public. Indeed, many of these experts are employed by predictive-coding vendors and may have internalized their employers' profit motive.[103]

A handful of judges and commentators have proposed that courts reassert authority over predictive-coding experts by qualifying them through *Daubert* hearings. Now codified at Federal Rule of Evidence 702, *Daubert* hearings are used to ensure the reliability of expert evidence to be offered at trial.[104] In the predictive-coding context, they would be used to ensure the

[100] *See* Looby & Kaplan, *supra* note 78, at 30 (noting the view that "understanding the different predictive coding algorithms and classifiers is for the technology geeks to deal with" (internal quotation marks omitted)).

[101] *See id.*

[102] *See Panel Discussion*, *supra* note 63, at 6 (statement of Foley) (arguing that "predictive coding technology can't be the exclusive province of technology aficionados. . . . The people who understand the facts and the procedural posture of the case need to be involved in training the system").

[103] *See* Sylvester, *supra* note 49.

[104] FED. R. EVID. 702.

reliability of experts designing and defending particular technologies and protocols.[105]

The applicability of *Daubert* hearings to the opinions and advice of predictive-coding experts is not obvious.[106] In other contexts, their function is to determine whether expert testimony should be admitted into evidence by determining whether the proffered expert meets certain standards of training and reliability.[107] Predictive-coding experts are not offering evidence, however. Rather, they are vouching for the reliability of a particular process of document production in discovery.

Certainly, *Daubert* hearings could be repurposed to qualify predictive-coding experts as a means of establishing consistency. But in light of the sparse empirical data that currently exists comparing predictive-coding products, the immediate result would be to bring vendor–experts before the court to opine on the functionality and reliability of *their own products*. Moreover, using *Daubert* hearings to qualify predictive-coding technologies would reinforce the impression that document review as an aspect of discovery practice is no longer a professional task of lawyers; it now falls within the purview of other experts. The ultimate result would therefore be to cede even greater control over the new technologies to IT professionals and vendors.

3. Patent Monopolies[108]

A number of predictive-coding vendors are applying for and receiving patents on the algorithms that underlie their products. These patents, in

[105] Advocates of this approach argue that it would ensure that predictive coding delivers on its promised benefits of increased accuracy and efficiency while placing a much needed check on vendor interests and influence. *See, e.g.*, United States v. O'Keefe, 537 F. Supp. 2d 14, 24 (D.D.C. 2008) (suggesting *Daubert* hearings for keyword searching); Victor Stanley, Inc. v. Creative Pipe, Inc., 250 F.R.D. 251, 260–61 & n.10 (D. Md. 2008) (suggesting *Daubert* hearings for electronic discovery procedures generally); Craig B. Shaffer, *"Defensible" by What Standard?*, 13 SEDONA CONF. J. 217, 232 (2012); David J. Waxse & Brenda Yoakum-Kriz, *Experts on Computer-Assisted Review: Why Federal Rule of Evidence 702 Should Apply to Their Use*, 52 WASHBURN L.J. 207, 220 (2013).

[106] Judge Peck concluded that Rule 702 and *Daubert* are rules for admissibility of evidence at trial and are therefore not applicable to discovery search methods. Da Silva Moore v. Publicis Groupe, 287 F.R.D. 182 (S.D.N.Y. 2012), *adopted sub nom.* Da Silva Moore v. Publicis Groupe SA, No. 11 Civ. 1279(ALC)(AJP), 2012 WL 1446534 (S.D.N.Y. Apr. 26, 2012).

[107] FED. R. EVID. 702 (excluding expert testimony unless: (1) "the testimony is based on sufficient facts or data"; (2) "the testimony is the product of reliable principles and methods; and" (3) the witness "has reliably applied the principles and methods to the facts of the case").

[108] Many scholars, and even the Supreme Court, have referred to patents as "patent monopolies" even though patents do not necessarily confer a monopoly. *See, e.g.*, Dennis Crouch, *The Paramount Interest in Seeing that Patent Monopolies . . . Are Kept Within Their Legitimate Scope*, PATENTLY-O (Jan. 31, 2014), http://patentlyo.com/patent/2014/01/paramount-monopolies-legitimate.html (noting the Supreme Court and Federal Circuit's increased use of "patent monopoly" in their opinions, which is likely due to courts' recognition of excessive patent power).

turn, are jeopardizing the profession's jurisdiction in a third way—by interfering with the bar's ability to ensure widespread access to legal services.

Controversy has long surrounded the patentability of algorithms.[109] Courts traditionally characterized them as either laws of nature, natural phenomena, or abstract ideas—not the products of invention.[110] But in 1980, the U.S. Supreme Court held that algorithms could be patented if applied in sufficiently concrete and practical ways.[111] The Court offered little guidance as to how and when that standard would be met,[112] but parties began applying for, and the Patent and Trademark Office ("PTO") began granting, patents for a broad range of processes that applied mathematical

[109] In an issue of first impression, the Supreme Court initially rejected patentability of a software algorithm because it resembled an abstract idea, similar to a mental process. Gottschalk v. Benson, 409 U.S. 63, 67–72 (1972) (holding a computer-based algorithm that converted numbers from decimal format to binary format unpatentable). In a 5–4 decision, the Supreme Court's first approval of software patentability occurred in *Diamond v. Diehr*, 450 U.S. 175 (1981). Since then, the Federal Circuit has loosened the standard for software to be eligible for patent protection. *See, e.g.*, Arrhythmia Research Tech., Inc. v. Corazonix Corp., 958 F.2d 1053, 1058 (Fed. Cir. 1992) (explaining the *Freeman–Walter–Abele* test for determining patentability of an algorithm); *see also In re* Abele, 684 F.2d 902 (C.C.P.A. 1982), *abrogated by In re* Bilski, 545 F.3d 943 (Fed. Cir. 2008), *aff'd*, Bilski v. Kappos, 130 S. Ct. 3218 (2010) (rejecting the Federal Circuit's machine-or-transformation test as the sole test for determining patentability of a process); *In re* Walter, 618 F.2d 758 (C.C.P.A. 1980), *abrogated by In re* Bilski, 545 F.3d 943, *aff'd*, Bilski v. Kappos, 130 S. Ct. 3218 (2010) (same); *In re* Freeman, 573 F.2d 1237 (C.C.P.A. 1978), *abrogated by In re* Bilski, 545 F.3d 943, *aff'd*, Bilski v. Kappos, 130 S. Ct. 3218 (2010) (same). Long-standing confusion over software patentability was not lessened by the Supreme Court's recent decision in *Bilski*. *See* 130 S. Ct. at 3221 ("The Court, therefore, need not define further what constitutes a patentable 'process,' beyond pointing to the definition of that term provided in § 100(b) and looking to the guideposts in *Benson*, *Flook*, and *Diehr*."); *see also id.* at 3236 (Stevens, J., concurring) ("The Court . . . never provides a satisfying account of what constitutes an unpatentable abstract idea."). In the three years since *Bilski*, courts have struggled to define criteria suitable to determine whether a process based on an algorithm should be patentable. *See, e.g.*, CLS Bank Int'l v. Alice Corp. Pty. Ltd., 717 F.3d 1269 (en banc) (per curiam) (Fed. Cir. 2013) (containing seven separate opinions, without a majority, on the patent eligibility of software-based processes), *cert. granted*, 134 S. Ct. 734 (2013).

[110] *See, e.g.*, *Diehr*, 450 U.S. 175. These are considered unpatentable subject matter because they are the building blocks of innovation, properly located in the public domain. *See Arrhythmia Research Tech., Inc.*, 958 F.2d at 1057 ("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.").

[111] *Gottschalk*, 409 U.S. at 67–68 (explaining that a patentable "process" cannot be "abstract and sweeping"); *Diehr*, 450 U.S. at 191–92 (holding that patentability requirements are met "when a claim containing a mathematical formula implements or applies that formula in a structure or process which, when considered as a whole, is performing a function which the patent laws were designed to protect [such as] transforming or reducing an article to a different state or thing").

[112] *Diehr*, 450 U.S. at 191–92.

algorithms.[113] Lower courts have been struggling ever since to draw a principled line between patentable processes containing algorithms and unpatentable abstract ideas.[114] Predictive-coding technologies offer a particularly salient illustration of this difficulty because of their connection to human cognition.

Notwithstanding this difficulty, 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.[115] For example, Reccomind, Inc. holds a patent that claims a predictive-coding method using broad functional language.[116] Other issued patents are narrower, claiming particular enhancements of predictive coding.[117] Equivio Ltd., for example, holds a patent for a specific method of producing a search algorithm.[118] Thus far, only one predictive-coding patent has been challenged in court. The case was dismissed on procedural grounds.[119]

Patents on law-practicing algorithms cannot encroach upon a firm or lawyer's ability to conduct manual review, but they can and likely will interfere with the profession's ability to ensure widespread access to the tools of lawyering. Patents allow their holders to prohibit use, sale, and

[113] *See infra* notes 115–18.

[114] *See, e.g.*, Bancorp Servs., L.L.C. v. Sun Life Assurance Co. of Canada (U.S.), 687 F.3d 1266, 1275 (Fed. Cir. 2012) ("[A]n application of a law of nature or mathematical formula to a known structure or process may well be deserving of patent protection." (quoting *Diehr*, 450 U.S. at 187) (internal quotation marks omitted)); Research Corp. Techs., v. Microsoft Corp., 627 F.3d 859, 868 (Fed. Cir. 2010); *see also supra* note 109.

[115] *See, e.g.*, Sys. & Methods for Predictive Coding, U.S. Patent No. 7,933,859 (filed May 25, 2010); Sys. & Method for Assisted Document Review, U.S. Patent No. 8,165,974 (filed June 8, 2009); Method & Sys. for Providing Elec. Discovery on Computer Databases & Archives Using Artificial Intelligence to Recover Legally Relevant Data, U.S. Patent No. 6,738,760 (filed Aug. 9, 2000).

[116] *See* '859 Patent, at col. 18, l. 52.

[117] *See, e.g.*, Computerized Sys. for Enhancing Expert-Based Processes & Methods Useful in Conjunction Therewith, U.S. Patent No. 8,346,685 (filed Apr. 22, 2009); *see also* U.S. Patent Application No. 2013/0077857 (filed Nov. 21, 2012) (claiming a 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 Application No. 2012/0278266 (filed Apr. 27, 2012) (claiming a method of populating and coding the seed set to train the computer).

[118] '685 Patent, at col. 13, l. 9–39.

[119] Adiscov, LLC v. Autonomy Corp., 762 F. Supp. 2d 826, 829 (E.D. Va. 2011). Adiscov, LLC sued Autonomy Corp. and Recommind, Inc. for allegedly infringing on its e-discovery patent. *See also* Adiscov, LLC v. Autonomy Corp, PLC et al., No. 5:11cv2897 (N.D. Cal. filed June 16, 2011); Adiscov, LLC v. Autonomy Corp, PLC et al., No. 1:11cv116 (E.D. Va. filed Feb. 2, 2011); Autonomy, Inc v. Adiscov, LLC, No. 3:11cv420 (N.D. Cal. filed Jan. 27, 2011); Adiscov, LLC v. Autonomy, Inc et al., No. 2:10cv218 (E.D. Va. filed May 17, 2010). In addition, Adiscov, L.L.C. filed suit against Kroll Ontrack, Inc., but the claims were dismissed without explanation. *See* Order of Dismissal as to Defendants Kroll Outrack, Inc. & Engenium Corp., Adiscov, L.L.C. v. Kroll Ontrack, Inc., No. 2:07cv280.2:07-cv-280-DF (E.D. Tex. Jan. 31, 2008).

reverse-engineering of an invention for twenty years.[120] Predictive-coding patents could therefore allow a single vendor to exercise exclusive control over a particular technology, with unlimited power to set licensing fees or to refuse to license the technology at all. Already, high start-up costs are creating uneven access to predictive-coding technologies. Patents will exacerbate this problem by increasing the frequency of cases in which one party has access to the new technologies and the other party does not.

These cases 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 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 limited resources to fund manual review and no access to predictive-coding capabilities, may hide relevant and damning documents amidst massive document productions. Because these problems will be rooted in patent exclusivity, the profession will have little ability to combat them.

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

C. Adversarial Values

The third set of problems with the profession's current approach to predictive coding stems from the extent to which judges and commentators are encouraging and requiring increased cooperation and transparency in the adoption and use of the new technologies. 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.[121] But cooperation is one

[120] 35 U.S.C. § 154(a)(2) (2006).

[121] *See, e.g.*, FED. R. CIV. P. 26(f) (requiring parties to discuss an in-depth discovery plan in advance); FED. R. CIV. P. 1 ("[The Federal Rules of Civil Procedure] should be construed and administered to secure the just, speedy, and inexpensive determination of every action and proceeding."); Baron, *supra* note 49, at 5 ("To meet the challenge of the exploding volume and complexity of potential electronic evidence, lawyers must . . . think about new ways of approaching structured cooperation within the bounds of the adversary system."). The drafters of the Federal Rules initially conceived of discovery as a cooperative process between parties, Subrin, *supra* note 2, at 717, and subsequent amendments consistently aimed to increase cooperation. *See* FED. R. CIV. P. 26(b)(5) advisory comm. n. 2006 amend.; FED. R. CIV. P. 26(a)(1) advisory comm. n. 2000 amend.; Order Prescribing Amendments to the Federal Rules of Civil Procedure, 507 U.S. 1089, 1118–21, 1125 (1993); *see also* Beisner, *supra* note 5, at 563,

value to be balanced among many. In ignoring this and mandating new kinds and degrees of cooperation in connection with predictive coding, managerial trial judges may undermine core client protections.

1. New Kinds of Cooperation

The predictive-coding protocol that Judge Peck approved in *Da Silva Moore* required the defendants (the producing party) to grant the plaintiffs full access to the documents and coding of their seed set and an opportunity to provide input on both initial coding decisions and subsequent quality control efforts.[122] Other judges have similarly required or encouraged seed-set transparency,[123] while commentators, for their part, are proposing even higher levels of required cooperation. The Sedona Conference[124] suggests that a producing party's knowledge of its own data may create a duty to disclose defects in proposed predictive-coding search methodologies.[125]

These forms of cooperation would be unprecedented under a more traditional discovery regime. Parties engaged in manual review and production have an initial obligation to make mandatory disclosures, but no subsequent obligation to direct opponents' discovery requests towards particular documents or information. Access to the process through which an opponent makes relevancy and privilege determinations will typically be denied, likely on the grounds of attorney work product. If challenged, it will be subject to an adversarial proceeding before the judge.

Departing from this approach and requiring seed-set transparency threatens core protections for attorney work product, attorney–client privilege, and confidentiality. Non-privileged, non-responsive documents in a seed set could include information that reveals unethical or criminal activity by a party, embarrasses an officer or employee, or aids the requesting

578, 582. Most recently, the Rules were amended to establish a system of mandatory initial disclosures between parties. FED. R. CIV. P. 26(a)(1).

[122] *See* Da Silva Moore v. Publicis Groupe, 287 F.R.D. 182, 192 (S.D.N.Y. 2012).

[123] Case Management Order: Protocol Relating to the Production of Electronically Stored Information ("ESI"), No. 6:11-md-2299, 2012 WL 6061973 (W.D. La. July 27, 2012) (requiring the parties, among other things, to collaborate in populating the seed set and training the computer); Kleen Prods. LLC v. Packaging Corp. of America, No. 10 C 5711, 2012 WL 4498465, at *4 (N.D. Ill. Sept. 28, 2012) (requiring parties to pursue a cooperative and collaborative approach in designing a search protocol).

[124] The Sedona Conference is a legal think tank that focuses on complex litigation. *See Frequently Asked Questions*, SEDONA CONF., https://thesedonaconference.org/faq (last visited Mar. 23, 2014).

[125] *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* Symposium, *Ethics and Professionalism in the Digital Age: A Symposium of the Mercer Law Review*, 60 MERCER L. REV. 863, 877 (2009) (presenting the comments of Jason Baron).

party in an unrelated cause of action.[126] Requiring disclosure of this information for transparency's sake could impose serious harm on a producing party. But lawyers may be reluctant to challenge a judge's order for cooperation, fearing potential retaliation.

2. New Degrees of Cooperation

In addition to encouraging new *kinds* of cooperation in the use of predictive coding, judges have also been requiring new *degrees* of cooperation. In doing so, they have been accelerating a preexisting trend away from the overarching goal of comprehensiveness in discovery practice—of bringing to light all relevant and non-privileged information so as to construct the strongest possible case on behalf of clients and against opponents. Aspects of this trend are undoubtedly positive, as cooperation can be beneficial for all the parties involved. But if the push toward cooperation is left entirely unchecked by the goal of comprehensiveness, the discovery system will break with the adversary system and call into question the legitimacy of court processes.

Traditionally, the goal of comprehensiveness, which was inevitably interpreted differently by requesting and producing parties, provided an important check on levels of cooperation in discovery practice.[127] Requesting parties sought to discover as much information as possible to construct the strongest possible case for their client and against their opponent. Producing parties sought to withhold non-relevant and privileged information to protect their client against confidential disclosures. The parties' divergent interests ensured that any and all cooperation occurs within the adversarial system's framework of strong client protections.

The costs of achieving true comprehensiveness are too high for the system to bear. Recognizing this, the Judicial Conference qualified and modified the goals of discovery long before the advent of predictive coding.[128] In 1983, the Conference amended the Federal Rules to support the new goal of proportionality between discovery requests and the needs of the case.[129] In 2006, it amended the Rules again to adapt e-discovery to this goal.[130]

[126] *See* Barkett, *supra* note 99, at 32–33.

[127] Some commentators have characterized it as creating fundamental and unresolvable tension with the value of cooperation. *See, e.g.*, Beckerman, *supra* note 4, 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.

[128] Beisner, *supra* note 5, at 561–62.

[129] *Id.* at 562 n.79.

[130] FED. R. CIV. P. 26, comm. n. subdiv. (b)(2), *available at* http://www.uscourts.gov/uscourts/RulesAndPolicies/rules/EDiscovery_w_Notes.pdf; *see* Douglas L. Rogers, *A Search for Balance in the Discovery of ESI Since December 1, 2006*, 14 RICH. J.L. & TECH. 8, 34–35 (2007), *available at* http://jolt.richmond.edu/v14i3/article8.pdf; *see also* Benjamin D. Silbert, *The 2006*

This trend away from comprehensiveness and toward proportionality may be both necessary and wise, but predictive coding will take it a significant step farther. Previously, proportionality entailed a substantive inquiry into the appropriateness and necessity of particular discovery requests in light of the case as a whole. Under a predictive-coding regime, proportionality 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.[131] Sometimes parties reach agreement because of a judge's order; other times they do so without oversight. In either case, the ultimate result will be a new form of court-found 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 it exacts a price from the legitimacy of our adversarial system. If parties to a case agree to save costs by limiting discovery to 5% of all relevant documents, courts could theoretically resolve the case based on a randomly selected 5% of relevant documents. Taken to an extreme, this creates a discovery lottery system. A smoking gun document may or may not be in the selected 5%; a group of emails that together establish knowledge may or may not all be in the 5%.

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.[132] In addition to resolving disputes, courts articulate rules, establish precedents, and serve as visible symbols of the rule of law in society.[133] These public functions require judicial legitimacy and credibility, which, in turn, require established procedures for seeking truth and promoting justice. Courts' public functions would be 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 toward increased cooperation and transparency should be made with an awareness of the implicated trade-offs with other values of our judicial system. They should not be made out of blind faith in predictive

Amendments to the Rules of Civil Procedure: Accessible and Inaccessible Electronic Information Storage Devices, Why Parties Should Store Electronic Information in Accessible Formats, 13 RICH. J.L. & TECH. 4, 14–15 (2007), *available at* http://jolt.richmond.edu/v13i3/article14.pdf.

[131] *See Panel Discussion*, *supra* note 63, at 9.

[132] William M. Landes & Richard A. Posner, *Adjudication as a Private Good*, 8 J. LEGAL STUD. 235, 236 (1979).

[133] Courts therefore differ fundamentally from private forums for dispute resolution. Owen M. Fiss, Commentary, *Against Settlement*, 93 YALE L.J. 1073, 1085–86 (1984); David Luban, *Settlements and the Erosion of the Public Realm*, 83 GEO. L.J. 2619, 2622 (1995).

coding's promise as a fail-safe means of reducing costs and limiting adversarial excess.

Shifts towards cooperation and transparency should also be effectuated comprehensively, not on a court-by-court and judge-by-judge basis. Right now, litigants contemplating use of predictive coding face significant uncertainty. Some judges are adopting predictive coding uncritically, while others remain resistant to its use or unaware of its existence.[134] Uncertainty is always problematic in the litigation system, but the difficulties are magnified here, implicating not only the defensibility of predictive coding as a discovery tool, but also potentially altered ethical standards.

III. THE PATH AHEAD

In light of the risks just discussed, the profession has an obligation to take a more proactive and systematic approach to the use and adoption of predictive-coding technologies. In this final Part, I propose that the profession take action along the following four lines: (1) raising awareness and understanding within the legal community; (2) working with IT professionals to establish minimum functionality standards; (3) standardizing implementing protocols; and (4) ensuring widespread access to the technology. In this way, the profession can participate in the development of the new technologies with a critical eye toward the trade-offs involved and a firm commitment to clients, the court system, and the public at large.

A. EDUCATION

As a preliminary step, the profession should raise awareness and understanding of predictive coding. Although the new technologies have emerged as a core topic of concern in some circles,[135] many judges and lawyers remain unaware of their existence. This means that a relatively small sector of the profession exercises disproportionate influence over the technologies' trajectory. It also produces significant uncertainty among litigants and lawyers. The organized bar should remedy this situation by sponsoring proactive educational efforts in law schools, practice communities, and judiciaries.

Courts, meanwhile, should encourage and facilitate use of special masters.[136] Many courts already use special masters to provide judges and

[134] Their decisions, moreover, are rarely subject to appellate review. Murphy, *supra* note 46, at 639.

[135] *See, e.g.*, Peck, *supra* note 13; Murphy, *supra* note 32.

[136] *See* Daniel B. Garrie, *Matrimonial Law Economics: Electronic Discovery and Change in Senior Partner's Role*, 27 AM. 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* Nicholas Barry, Note, *Man Versus Machine Review: The Showdown Between Hordes of Discovery Lawyers and a Computer-Utilizing Predictive-Coding Technology*, 15 VAND. J. ENT. & TECH. L. 343, 364

parties with technological expertise and guidance and to facilitate early resolution of e-discovery disputes.[137] Some courts have established protocols for the selection of special masters. Many commentators, for their part, advocate increased reliance on them for e-discovery issues generally.[138] These practices should be continued and expanded to address predictive-coding technologies in particular.

B. Quality Control

As a second critical step, the profession should set minimum functionality standards for products and processes labeled "predictive coding." Currently, countless vendors market products described as predictive-coding technologies.[139] Some are mislabeled keyword searching technologies.[140] Others are appropriately called predictive coding but fall within a broad umbrella of technologies that have as many differences as commonalities.[141] The bar should gather data about these products to understand and evaluate the extent of the variation. To do so, it should commission a comprehensive third-party efficacy study, perhaps through the ABA or the Sedona Conference,[142] that solicits the participation of relevant stakeholders from within and outside of the profession.

The efficacy study should address both processes and results.[143] Because different predictive-coding tools are based on different text-classifying systems, their processes are more or less suited to different types of documents and data, and they produce varying degrees of recall and precision when applied to different documents and data sets. Understanding both points of comparison—processes and results—will therefore be critical.

(2013) ("It will be the legal community's responsibility to promote and educate the bench about these new technologies.").

[137] Nora Barry Fischer & Richard N. Lettieri, *Creating the Criteria and the Process for Selection of E-Discovery Special Masters in Federal Court*, FED. LAW., Feb. 2011, at 36, 37.

[138] *Id.* at 38–39; *see also* Shira Scheindlin, *We Need Help: The Increasing Use of Special Masters in Federal Court*, 58 DEPAUL L. REV. 479, 486 (2009) (predicting that the future will see increased use of special masters); Shira A. Scheindlin & Jonathan M. Redgrave, *Special Masters and E-Discovery: The Intersection of Two Recent Revisions to the Federal Rules of Civil Procedure*, 30 CARDOZO L. REV. 347, 347 (2008) (arguing that the use of special masters is "both necessary and desirable" in the field of e-discovery).

[139] *See supra* text accompanying note 82.

[140] *Predictive Coding and Patented Workflow, supra* note 51.

[141] *See* Melissa Whittingham et al., *Predictive Coding: E-Discovery Game Changer?*, EDDE J., Autumn 2011, at 11, 11–12, *available at* http://www.americanbar.org/content/dam/aba/publications/emerging_news/autumn_2011v2i4.authcheckdam.pdf.

[142] *See* Baron, *supra* note 49, at 9 (citing works from the Sedona Conference); Elle Byram, *The Collision of the Courts and Predictive Coding: Defining Best Practices and Guidelines in Predictive Coding for Electronic Discovery*, 29 SANTA CLARA COMPUTER & HIGH TECH. L.J. 675, 697 (2013) (citing works from the Sedona Conference); Barry, *supra* note 136, at 368.

[143] Barry, *supra* note 136, at 368.

The Legal Track Interactive Task of the Text Retrieval Conference ("TREC"), an annual simulation of civil litigation document review systems, has started important work along these lines.[144] 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, while it represents a useful and important starting point in evaluating various products, it needs to be built upon and expanded.

C. *Standardization*

The profession should also develop standardized use protocols. Currently, there is significant variation 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 client protections. To achieve greater uniformity, the profession should initiate a participatory and deliberation-based taskforce to design standardized protocols. As with the proposed commission to study the efficacy of various products, a group to produce standardized protocols should include representatives from all relevant groups of stakeholders, including lawyers, judges, academics, computer scientists, commercial vendors, and potential litigants.

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 data at issue is suitable to predictive-coding approaches.[145] 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

[144] TREC, which was created to "develop search technology that meets the needs of lawyers to engage in effective discovery in digital document collections," hosts an annual Legal Track Interactive Task that simulates a civil litigation document review process. *TREC Tracks*, NAT'L INST. STANDARDS & TECH., http://trec.nist.gov/tracks.html (last updated Feb. 24, 2014); *see also* Douglas W. Oard et al., *Overview of the TREC 2008 Legal Track*, *in* NIST SPECIAL PUBLICATION: SP 500-277, THE SEVENTEENTH TEXT RETRIEVAL CONFERENCE (TREC 2008) PROCEEDINGS (2008), *available at* http://trec.nist.gov/pubs/trec17/papers/LEGAL.OVERVIEW08.pdf; Bruce Hedin et al., *Overview of the TREC 2009 Legal Track*, *in* NIST SPECIAL PUBLICATION: S 500-278, THE EIGHTEENTH TEXT RETRIEVAL CONFERENCE (TREC 2009) PROCEEDINGS (2009), *available at* http://trec.nist.gov/pubs/trec18/papers/LEGAL09.OVERVIEW.pdf.

[145] Byram, *supra* note 142, at 694.

among the values of cooperation, transparency, and client protection. Third, standardized protocols should establish appropriate and required quality-control checks.[146] Fourth and finally, the protocols should address proper procedures for handling disclosures of privileged information, including whether, when, and how to use claw-back agreements.[147] 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 they will be appropriate in some cases that employ predictive coding.[148]

D. ACCESS

Perhaps most importantly, the profession must stay cognizant of its fundamental obligation to ensure access to effective legal services. Regardless of potential long-term savings, predictive coding entails significant up-front costs that will be prohibitive for many parties. The problem will be exacerbated if the technologies are patented, requiring payment of expensive licensing fees. In light of these impediments to use, the profession needs to explore possible means of ensuring widespread access.

One option is 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.[149] The bar could build upon one of these platforms to develop a tool suited to the needs and standards of the legal

[146] *See* Baron, *supra* note 49, 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'" (quoting *The Sedona Conference Committee on Achieving Quality in the E-Discovery Process*, 10 SEDONA CONF. J. 299, 303 (2009))); Byram, *supra* note 142, at 697 (discussing the necessity for quality assurance techniques); Barry, *supra* note 136, at 369 (discussing the necessity for selective sampling at multiple points in the e-discovery process).

[147] 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 v. UBS Warburg LLC (*Zubulake III*), 216 F.R.D. 280, 290–91 n.81 (S.D.N.Y. 2003) ("[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." (quoting THE SEDONA CONFERENCE, THE SEDONA PRINCIPLES: BEST PRACTICES RECOMMENDATIONS & PRINCIPLES FOR ADDRESSING ELECTRONIC DOCUMENT PRODUCTION 33 (2003), *available at* https://thesedonaconference.org/publication/The%20Sedona%20Principles)).

[148] *Zubulake III*, 216 F.R.D. at 290 ("[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.").

[149] *See, e.g.*, *KNIME Open Source Story*, KNIME, http://www.knime.org/knime-open-source-story (last visited Mar. 23, 2014) (discussing KNIME's processing abilities and availability); *Solutions*, RAPIDMINER, http://rapidminer.com/solutions/ (last visited Mar. 23, 2014) (discussing RapidMiner's text-mining and predictive-analytics capacities).

profession. Alternatively or additionally, universities and academic institutions—many of which are already developing open-source predictive-analytics tools—could design or adapt tools to meet the needs of the legal profession. Developing an open-source tool would entail a significant drawback, however: Lawyers would have access to the tool, but not necessarily to technical support. A lack of technical support, in turn, could feed into many of the problems discussed above regarding lawyers' lack of technological understanding and expertise.

Accordingly, the bar should consider other options as well. The ABA or other bar groups could contract with particular vendors to provide all bar members with access to a cost-effective predictive-coding tool. Universities regularly pursue this approach with products such as word-processing programs, antivirus programs, and photo-editing software, obtaining significant discounts for faculty, staff, and students in exchange for guaranteed volumes of sale.[150] This could be a promising path for predictive-coding products, where vendors may be willing to offer significant discounts 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.[151]

Vendors may object to all of these proposals, but they would be well-advised to work with the profession to increase access. 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 these new technologies as an answer to the problems of civil discovery. They explained that predictive coding could address the skyrocketing volume and expense of electronically stored information while restoring trust in a system that had long been plagued by problems of excess and abuse.

Like all technologies, however, predictive coding is not an unmitigated good. It threatens to create new problems while solving existing ones. Many judges and lawyers are ignoring this, failing to recognize that adoption and use entail ethical trade-offs. In doing so, they are precluding a role for the bar in the technology's development, undermining the scope of the

[150] *See, e.g.*, *Academic Volume Licensing*, MICROSOFT, http://www.microsoft.com/education/ww/buy/Pages/volume-licensing.aspx (last visited Mar. 23, 2014).

[151] *Cf.* Peter Maybarduk & Sarah Rimmington, *Compulsory Licenses: A Tool to Improve Global Access to the HPV Vaccine?*, 35 AM. J.L. & MED. 323, 325 (2009).

profession's jurisdiction, and threatening the integrity of the adversarial system.

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. It should undoubtedly look to predictive coding as a powerful and potentially beneficial tool. It must do so, however, with a critical eye and a firm commitment to employ the new technology in service of the goals and values of the legal profession and the judicial system.