UNC SCHOOL OF LAW

NORTH CAROLINA LAW REVIEW

Volume 93 | Number 1 Article 7

12-1-2014

Shaping the Technology of the Future: Predictive Coding in Discovery Case Law and Regulatory Disclosure Requirements

Christina T. Nasuti

Shaping the Technology of the Future: Predictive Coding in Discovery Case Law and Regulatory Disclosure Requirements*

[A]cquiring preemptive knowledge about emerging technologies is the best way to ensure that we have a say in the making of our future.

—Catarina Mota, TEDGlobal Fellow[1]

1. Catarina Mota, *Play with Smart Materials*, TED (July 2012), http://www.ted.com/talks/catarina_mota_play_with_smart_materials/transcript?language=en.

INTRODUCTION

The twentieth and twenty-first centuries brought about a technological revolution for businesses and everyday life. The legal profession was not left unchanged, as innovations from typewriters to computers to the Internet transformed legal practice.[2] The corresponding creation and accumulation of electronic data continues to fundamentally impact legal practice,[3] particularly by accounting for massive increases in digital matter available for legal review.[4] Computer-assisted review technologies, and predictive coding in particular, permit the legal field to foray into the digital mass that now defines the professional world.[5] It is not surprising to see that

2. *See* WORKING GRP. ON ELEC. DOCUMENT RETENTION & PROD., THE SEDONA CONFERENCE, COMMENTARY ON INFORMATION GOVERNANCE 2 (2013) [hereinafter THE SEDONA CONFERENCE], *available at* https://thesedonaconference.org/download-pub/3421 ("We live and work in an information age that is continually—and inexorably—transforming how we communicate and conduct business.").

3. *Cf.* Darla W. Jackson, *Lawyers Can't Be Luddites Anymore: Do Law Librarians Have a Role in Helping Lawyers Adjust to the New Ethics Rules Involving Technology?*, 105 LAW LIBR. J. 395, 397–400 (2013) (discussing the evolving technological demands on lawyers and how law librarians can aid practitioners).

4. *See* John T. Yip, *Addressing the Costs and Concerns of International E-Discovery*, 87 WASH. L. REV. 595, 595–96 (2012).

5. *See, e.g.*, Maura R. Grossman & Gordon V. Cormack, *The Grossman-Cormack Glossary of Technology-Assisted Review with Foreword by John M. Facciola, U.S.*

these technologies provide important future venues for dealing with the electronic realities. What is significant, however, is the rapidity with which these technologies could be incorporated into the current judicial and regulatory regimes. Because the legal field already is feeling the impact of these technologies on its traditional review and conflict-resolution mechanisms, this Comment focuses on how judicial and regulatory regimes should react to the changing landscape.

The framework in which these emerging technologies develop will guide how they evolve and whether they can be sufficiently implemented in the legal context. As a result, the legal profession must advocate for courts and regulatory bodies to promptly develop and update guidelines intended to address predictive coding and other computer-assisted review technologies. While recent cases have provided some transparency into the judicial perspective on predictive coding and its role in litigation, agencies have generally been less transparent in their implementation or acceptance of predictive coding in the regulatory context. This Comment argues that, to effectuate a meaningful evolution and implementation of predictive coding, courts and government agencies should not experiment in vacuums independent of each other. Rather, both courts and agencies should demonstrate the utmost transparency when utilizing and addressing this new technology. Increased transparency will allow each body to learn from the other's experiences and eventually enable both to put predictive coding

---

Magistrate Judge, 7 FED. CTS. L. REV. 1, 27 (2013) ("A 2012 study (Nicholas M. Pace & Laura Zakaras, *Where the Money Goes: Understanding Litigant Expenditures for Producing Electronic Discovery*, RAND Institute for Civil Justice (2012)), indicat[ed] that Document review accounts for 73% of Electronic Discovery costs, and conclude[ed] that '[t]he exponential growth in digital information, which shows no signs of slowing, makes a computer-categorized review strategy, such as predictive coding, not only a cost-effective choice but perhaps the *only* reasonable way to handle many large-scale productions.' ").

technology to its highest and best use in appropriate contexts, while still maintaining a healthy skepticism regarding the different uses and challenges posed in each venue. The need for transparency and consideration of other contexts is especially important when considered in light of normative concerns. Namely, the juxtaposition of these forums raises an important question regarding whether courts and agencies *should* consider each other and work in tandem—otherwise, predictive coding technologies could hypothetically support an agency determination of wrongdoing but remain unacceptable for use in a judicial context.

This Comment addresses the origins of, current status of, and future possibilities for predictive coding. To that end, Part I addresses the rise of electronic data and introduces the reader to predictive coding as well as other terminology and concepts surrounding computer-assisted review. Part II surveys the general status of predictive coding and the opportunities it presents to the legal profession. The next two parts, Parts III and IV, address how predictive coding is applied and addressed in cases by the judiciary and by regulatory agencies. Part V provides a relatively brief overview of miscellaneous issues to consider as judges, regulators, and reformers develop rules, regulations, and recommendations addressing the legal use of predictive coding. Finally, Part VI compares and distinguishes the judicial and regulatory treatments of predictive coding and offers recommendations for each in order to assure best practices.

I. THE RISE OF ELECTRONIC DATA AND AN INTRODUCTION TO PREDICTIVE CODING

A. *The Electronic Revolution*

The recent data revolution resulted in a global shift from hard-copy files and communications to electronic versions.[6] This shift was inspired by the newfound ability to create and maintain electronic records and communications—a change so fundamentally revolutionary that it is sometimes compared to the fifteenth century introduction of the printing press.[7] Companies and individual users are not alone in this electronic shift. The government is a substantial force in the move, as the executive branch has mandated that agencies embrace the digital reality.[8] As a result of the nearly

---

6. *See, e.g.*, Judge Herbert B. Dixon, Jr., *Automating the Search and Review of ESI*, JUDGES' J., Summer 2012, at 36, 36 ("[V]arious estimates [claim] that from 90 percent to 97 percent of today's business and personal records are created and maintained electronically, and that as little as 3 percent of information is printed on paper . . . ."). Grossman and Cormack define electronically stored information ("ESI") as a term

> [u]sed in Federal Rule of Civil Procedure 34(a)(1)(A) to refer to discoverable information "stored in any medium from which the information can be obtained either directly or, if necessary, after translation by the responding party into a reasonably usable form." Although Rule 34(a)(1)(A) references "Documents or Electronically Stored Information," individual units of review and production are commonly referred to as Documents, regardless of the medium.

Grossman & Cormack, *supra* note 5, at 15

7. *See* Working Grp. on Elec. Document Retention & Prod. & Search & Retrieval Scis. Special Project Team, The Sedona Conference, *Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery*, 8 SEDONA CONF. J. 189, 197 (2007) [hereinafter The Sedona Conference Working Group] ("This 'digital realm' . . . resulted in as fundamental a shift in the way information is shared as that which occurred in 1450 when Johannes Guttenberg invented the printing press.").

8. *See* Joseph Marks, *White House Overhauls Electronic Records Requirements*, NEXTGOV (Aug. 24, 2012), http://www.nextgov.com/cio-briefing/2012/08/white-house-overhauls-electronic-records-requirements/57651/?oref=govexec_today_nl ("Federal

universal move towards electronically stored information ("ESI"),[9] the sheer amount of data now available compounds generic concerns regarding information and data production and thus poses greater concerns for litigation in the digital age.[10] To visualize the vast amount of data facing modern litigators and regulators, consider this fact: as of "2011, the digital universe [had] expanded to over 1800 exabytes, enough data to fill 57.5 *billion* 32GB Apple iPads."[11]

As an unsurprising result, electronic discovery ("e-discovery") represents a crucial but overwhelming part of litigation budgets.[12] For numerical orientation, e-discovery costs surrounding discovery and

---

agencies have until the end of 2019 to adopt systems that store and manage all electronic records in formats that will keep them safe and searchable. . . . Agencies have until the end of 2016 to store all email in electronic formats and until Nov. 15, 2012, to appoint a senior official responsible for beefing up their electronic records management programs . . . .").

9. For a definition of electronically stored information ("ESI"), see *supra* note 6.

10. *See, e.g.*, Yip, *supra* note 4, at 595 ("[T]he rapidly increasing volume of ESI has substantially increased the costs of e-discovery for producing parties."). Because of these serious financial concerns, and their impact on litigation, *Zubulake v. UBS Warburg L.L.C.*, 216 F.R.D. 280 (S.D.N.Y. 2003), and its progeny evolved to permit "shift[ing] some of the e-discovery costs from the responding party to the requesting party." Yip, *supra* note 4, at 595. Significantly for the purposes of this Comment, predictive coding and related technologies promise reduced costs for both parties. This holds potential both to free parties from oppressive litigation by decreasing the financial burden even if shifting does not occur and to encourage other parties to engage in frivolous suits, as even if the court shifts the financial burden of their requests, the overall cost of production is far lower than that of litigation historically. For a discussion of the current state of the law regarding the allocation of e-discovery costs in litigation, see generally Jacqueline Hoelting, Note, *Skin in the Game: Litigation Incentives Changing as Courts Embrace a "Loser Pays" Rule for E-Discovery Costs*, 60 CLEV. ST. L. REV. 1103 (2013) and Emily P. Overfield, Note, *Shifting the E-Discovery Solution: Why* Taniguchi *Necessitates a Decline in E-Discovery Court Costs*, 118 PENN ST. L. REV. 217 (2013).

11. Yip, *supra* note 4, at 595 (emphasis added).

12. *See* Hoelting, *supra* note 10, at 1105; *see also* Grossman & Cormack, *supra* note 5, at 15 (defining electronic or e-discovery as "[t]he process of identifying, preserving, collecting, processing, searching, reviewing, and producing Electronically Stored Information that may be [r]elevant to a civil, criminal, or regulatory matter").

document production independently comprised an astounding "$2.8 billion in 2009,"[13] with continued projected increases as "[t]he amount of electronically stored information in the United States doubles every 18–24 months, and 90 percent of U.S. corporations are currently engaged in some kind of litigation."[14] In a case that eventually permitted the use of computer-assisted review, the original electronic records would have required "10 man-years of billable time" simply to adequately locate relevant documents.[15]

These astounding costs represent the irreconcilable chasm between traditional data review and the vast digital prowess of the modern era. In response, human ingenuity and the marketplace developed a solution: computer-assisted review and predictive coding technologies.[16] These advances promise fiscal benefits and increased efficiency.[17] In the aforementioned scenario, utilization of coding technologies would only require "two weeks to cull the relevant

13. *E.g.*, Yip, *supra* note 4, at 595.

14. Overfield, *supra* note 10, at 217; *see also* Dixon, *supra* note 6, at 36 ("[T]he RAND report estimates that human review of documents as part of responding to discovery requests consumes about 73 cents of every dollar spent on the production of ESI.").

15. Dixon, *supra* note 6, at 36–37 (citing Defendants' Memorandum in Support of Motion for Protective Order Approving the Use of Predictive Coding at 5, Global Aerospace Inc. v. Landow Aviation, L.P., Consolidated Case No. CL 61040 (Va. Cir. Ct. Apr. 9, 2012)).

16. For a helpful summary of the evolution of predictive coding beginning in 2008, see Charles Yablon & Nick Landsman-Roos, *Predictive Coding: Emerging Questions and Concerns*, 64 S.C. L. REV. 633, 637–38 (2013).

17. Ronni Solomon, *Are Corporations Ready to Be Transparent and Share Irrelevant Documents with Opposing Counsel to Obtain Substantial Cost Savings Through the Use of Predictive Coding?*, METROPOLITAN CORP. COUNS., Nov. 2012, at 26, 26, *available at* http://www.metrocorpcounsel.com/pdf/2012/November/26.pdf (describing the potential for over a million dollars in savings by adopting predictive coding in *Da Silva Moore v. Publicis Groupe*, 287 F.R.D. 182 (S.D.N.Y. 2012)).

documents" at a fraction of the cost for traditional methods.[18] As a result, humans matched the electronic revolution with potential electronic solutions—computer-assisted review technologies and predictive coding—that hold the potential to revitalize legal review despite massive increases in ESI.

B. Concepts and Terminology Surrounding Predictive Coding

Despite its rapid emergence and strong potential, predictive coding brings with it a host of confusion and concerns for modern attorneys—ranging from gaining the technological expertise necessary to understand its promises to learning how to apply the new technology. Even more simply, however, one must learn to speak the language of these new technologies. This Comment provides a brief overview of some of the essential terms necessary to understand the scholarship and debates surrounding predictive coding. However, parts of this terminology lack uniformity, and new technological advances may change the terminology. As a result, interested readers must dedicate constant self-study to the rapid revisions.[19]

As an initial matter, *technology-assisted review* ("TAR") and *computer-assisted review* ("CAR") are broad terms that encompass a number of technologies, including predictive coding as well as less

18. Dixon, *supra* note 6, at 36–37 ("[B]y use of predictive coding, it would take less than two weeks to cull the relevant documents at roughly 1/100 the cost of using humans to review every document in the database." (citing Protective Order Approving the Use of Predictive Coding for Discovery at 10, Global Aerospace Inc. v. Landow Aviation, L.P., Consolidated Case No. CL 61040 (Va. Cir. Ct. Apr. 23, 2012))).

19. For a comprehensive and helpful description of many terms surrounding computer-assisted review and predictive coding technologies, as well as the technical process and slight distinctions between different technologies at this level, see generally Grossman & Cormack, *supra* note 5.

advanced but similar technologies, such as keyword searching.[20] However, because there is no single lexicon governing computer-assisted review technologies, the terminology distinctions between TAR, CAR, and predictive coding technologies are not followed in much of the literature discussing and comparing predictive coding, technology-assisted review, and computer-assisted review.[21] The common example of a square and rectangle can be applied to clarify the distinction here. While predictive coding—the so-called square—is a type of computer- or technology-assisted review technology, CAR and TAR—the so-called rectangles—are broader terms that include other less automated technologies as well. This Comment seeks to define these terms for the reader as well as to use their most precise forms in order to alleviate confusion.

At its most technical, *predictive coding* is "[a]n industry-specific term generally used to describe a Technology [or Computer]-Assisted Review process involving the use of a Machine Learning Algorithm to distinguish Relevant from Non-Relevant Documents, based on Subject Matter Expert(s)' Coding of a Training Set of Documents."[22] In plain English, predictive coding matches human judgment and hands-on training with computer learning and iterative skill to teach software to quickly and accurately search and categorize documents,

20. *See generally, e.g.*, Sharon D. Nelson & John W. Simek, *Predictive Coding: A Rose by Any Other Name*, LAW PRAC., July–Aug. 2012, at 20 (arguing that all similar forms of analytical technology that have similar features but do not qualify as predictive coding technology are referred to as "TAR [technology-assisted review] or CAR [computer-assisted (or computer-aided) review]").

21. *See, e.g.*, *id.* at 20; *cf.* Grossman & Cormack, *supra* note 5, at 13, 32 (showing the terminology overlap between content-based advanced analytics ("CBAA") and technology-assisted review, which also overlaps with computer-assisted review and predictive coding).

22. Grossman & Cormack, *supra* note 5, at 26.

much like human-only review.[23] Pandora Internet radio provides a helpful analogy for this process: the computer program learns from users' positive or negative feedback—the equivalent of a Pandora "thumbs up" or "thumbs down" of a song—to predict future outputs, such as a desired song to play next, and the users' preferences.[24] For predictive coding, the initial learning process occurs as humans code a primary seed set of documents to teach the program what constitutes relevancy and privilege for the overall document set; the training is then repeated using sets of documents until the machine reaches a pre-determined accuracy in self-categorizing documents.[25]

A number of other terms are especially helpful for understanding how predictive coding is described, how it works, and how to grasp the legal significance and potential for its technological components.[26] First, the repeated interactive process[27] between the machine software

23. *See* Yablon & Landsman-Roos, *supra* note 16, at 638–42.

24. *See About Pandora*, PANDORA, http://www.pandora.com/about (last visited Nov. 19, 2014).

25. *See* Yablon & Landsman-Roos, *supra* note 16, at 638–40; *see also* Nelson & Simek, *supra* note 20, at 22 (defining predictive coding as requiring all of the following: "Integrated, keyword-agnostic analytics to quickly generate accurate seed sets[;] Language and keyword-agnostic machine-learning technology to accurate find relevant documents during the 'training' process[;] A sound and well-documented workflow[;] Integrated sampling to verify results to a statistical certainty before, during and after review[;] A completely integrated, purpose-built system to ensure results are consistent throughout the entire process, every time").

26. For a definition of predictive coding, see *supra* note 22 and accompanying text.

27. This process is known as iterative training. *See* Grossman & Cormack, *supra* note 5, at 20 ("Iterative Training: The process of repeatedly augmenting the Training Set with additional examples of Coded Documents until the effectiveness of the Machine Learning Algorithm reaches an acceptable level. The additional examples may be identified through Judgmental Sampling, Random Sampling, or by the Machine Learning Algorithm, as in Active Learning.").

and the human teachers is known as *active learning*.[28] This process employs a training set of documents coded by humans to teach predictive coding programs how to evaluate the relevance of future data.[29] Predictive coding[30] and similar technologies are able to learn from their human teachers because of their ability to "emulate human judgment"—a characteristic of their artificial intelligence.[31] A machine's ability to use its artificial intelligence to properly engage in the learned coding process, known as *machine learning*,[32] is measured by its accuracy.[33] Accuracy levels are determined by the program's

28. *Id.* at 8 ("Active Learning: An Iterative Training regimen in which the Training Set is repeatedly augmented by additional Documents chosen by the Machine Learning Algorithm, and coded by one or more Subject Matter Expert(s).").

29. *Id.* at 32–33 ("Training Set: A Sample of Documents coded by one or more Subject Matter Expert(s) as Relevant or Non-Relevant, from which a Machine Learning Algorithm then infers how to distinguish between Relevant and Non-Relevant Documents beyond those in the Training Set."). Note also that the first training set is known as the seed set. *See id.* at 29.

30. Coding is a short-hand term that refers to the human and/or "automated" process of "[l]abeling a [d]ocument as Relevant or Non-Relevant." *Id.* at 11.

31. *Id.* at 9 ("Artificial Intelligence: An umbrella term for computer methods that emulate human judgment. These include Machine Learning and Knowledge Engineering, as well as Pattern Matching (e.g., voice, face, and handwriting recognition), robotics, and game playing.").

32. *Id.* at 22 ("Machine Learning: The use of a computer Algorithm to organize or Classify Documents by analyzing their Features. In the context of Technology-Assisted Review, Supervised Learning Algorithms (e.g., Support Vector Machines, Logistic Regression, Nearest Neighbor, and Bayesian Classifiers) are used to infer Relevance or Non-Relevance of Documents based on the Coding of Documents in a Training Set. In Electronic Discovery generally, Unsupervised Learning Algorithms are used for Clustering, Near-Duplicate Detection, and Concept Search.").

33. *Id.* at 8 ("Accuracy: The fraction of Documents that are correctly coded by a search or review effort. Note that Accuracy + Error = 100%, and that Accuracy = 100% - Error. While high Accuracy is commonly advanced as evidence of an effective search or review effort, its use can be misleading because it is heavily influenced by Prevalence. Consider, for example, a Document Population containing one million Documents, of which ten thousand (or 1%) are Relevant. A search or review effort that identified 100%

responsiveness.[34] Responsiveness, a measure of how well the program returns relevant[35] documents, can be determined by the particular informational or legal need at hand and is often based on the proportion[36] of relevant documents returned versus non-relevant documents.[37] In order to measure accuracy, users employ a control set of random documents to assess the sufficiency of the system's coding abilities at that point.[38] Scholars often measure this machine-learning automated process against pre-determined accuracy levels, to ensure *quality control*,[39] as well as against the results that human reviewers would achieve under *manual review*.[40] These comparisons allow

---

of the Documents as Not Relevant, and, therefore, found *none* of the Relevant Documents, would have 99% Accuracy, belying the failure of that search or review effort."); *cf.* Ralph C. Losey, *Predictive Coding and the Proportionality Doctrine: A Marriage Made in Big Data*, 26 REGENT U. L. REV. 7, 21–24 (2013) (describing the machine-learning process, with a focus on predictive coding and the relevancy rankings).

34. Grossman & Cormack, *supra* note 5, at 28 ("Responsiveness: A Document that is Relevant to an Information Need expressed by a particular request for production or subpoena in a civil, criminal, or regulatory matter.").

35. *Id.* ("Relevance / Relevant: In Information Retrieval, a Document is considered Relevant if it meets the Information Need of the search or review effort.").

36. *Id.* at 26 ("Proportion: The fraction of a set of Documents having some particular property (typically Relevance).").

37. *See* Losey, *supra* note 33, at 21–24.

38. Grossman & Cormack, *supra* note 5, at 13 ("Control Set: A Random Sample of Documents coded at the outset of a search or review process that is separate from and independent of the Training Set. Control Sets are used in some Technology-Assisted Review processes. They are typically used to measure the effectiveness of the Machine Learning Algorithm at various stages of training, and to determine when training may cease.").

39. *Id.* at 27 ("Quality Control: Ongoing methods to ensure, during a search or review effort, that reasonable results are being achieved."); *see also id.* ("Quality Assurance: A method to ensure, after the fact, that a search or review effort has achieved reasonable results.").

40. *Id.* at 22 ("Manual Review: The practice of having human reviewers individually read and Code the Documents in a Collection for Responsiveness, particular issues, privilege, and/or confidentiality."); *see, e.g.*, Losey, *supra* note 33, at 13–14 (describing the

scholars not only to assess whether technology is sufficiently advanced to achieve an accuracy level sufficient for the micro-level coding tasks, but also to discuss on a macro policy level whether this technology is sufficiently advanced to complement, or even replace, traditional human review. While a number of more technical aspects and terms surround predictive coding and the mathematical decisions governing acceptable quality control levels, this explanation provides the rudimentary lexicon necessary for a discussion of predictive coding in the judicial and regulatory spheres.

II. PREDICTIVE CODING AND THE LEGAL WORLD

To elaborate on predictive coding's role in the legal world, this Part proceeds with five main themes: (1) predictive coding's legal applications; (2) solutions to problems posed by ESI and discussion by predictive coding's advocates; (3) attorneys' roles in relation to predictive coding; (4) predictive coding's main legal advantages; and (5) predictive coding's main disadvantages. Finally, it provides an intermediate conclusion to resolve these disparate pieces of a complicated technology in the broad legal realm before discussing specific case applications.

Predictive coding has numerous applications beyond the legal world. However, within the legal world, it is often referred to in connection with e-discovery, which is defined as "[t]he process of identifying, preserving, collecting, processing, searching, reviewing, and producing Electronically Stored Information [ESI] that may be [r]elevant to a civil, criminal, or regulatory matter."[41] The Federal Rules of Civil Procedure automatically deem ESI to be available as

traditional human review process and the challenges it faces in the ESI-era in a section that provides a contrast to document review that employs predictive coding).

41. Grossman & Cormack, *supra* note 5, at 15.

potential evidence in lawsuits,[42] consequently rendering almost any medium as fair game for discovery.[43] Because the discovery process was created around traditional review mediums and processes, conventional methods unsatisfactorily address the electronic world.[44]

Predictive coding provides a solution to the problems created when incorporating ESI into the legal arena: "(1) volume and duplicability, (2) persistence, and (3) dispersion."[45] For example, the multitude of emails produced in an investigation is compounded when the emails are produced in duplicates, as a new copy of a document is included for each sender or recipient. This unnecessarily requires evaluating attorneys to classify the same document multiple times. With problems like these, which can also include assessing incoming

42. 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, 346–47 (2013) ("An amendment to the FRCP [Federal Rules of Civil Procedure] in 2006 explicitly made all electronic files discoverable. The amended rules [, however,] did not make ESI discoverable for the first time; courts had long held that electronic files were discoverable even without a specific grant in the rules.").

43. *Id.* at 347 ("[E]-discovery has grown exponentially and now includes, *inter alia*, emails, word-processing files, spreadsheets, databases, video files, MP3 files, and virtually every other file now stored on computers and other electronic devices (such as PDAs, cell phones, flash drives, DVDs, etc.)."). The new role for ESI also creates an enhanced role for "big data" in the legal profession. *See* Jobst Elster, *Big Data for Law Firms: Hype, Reality, Myth, or Legend*, LEGAL MGMT., Oct.–Nov. 2013, at 35, 37. *See generally* THOMSON REUTERS, 50 STATE STATUTORY SURVEYS: CIVIL LAWS: CIVIL PROCEDURE: ELECTRONIC DISCOVERY (2014), *available at* 0020 SURVEYS 4 (Westlaw) (providing citations to forty-nine states' statutes and applicable federal statutes addressing electronic discovery).

44. *See* The Sedona Conference Working Group, *supra* note 7, at 198–99.

45. Barry, *supra* note 42, at 347 (citing WORKING GRP. ON ELEC. DOCUMENT RETENTION & PROD., THE SEDONA CONFERENCE, THE SEDONA PRINCIPLES: BEST PRACTICES RECOMMENDATIONS & PRINCIPLES FOR ADDRESSING ELECTRONIC DOCUMENT PRODUCTION 1 (2d ed. 2007), *available at* https://thesedonaconference.org/download-pub/81).

data, shifting production costs, or searching one's own documents for responsive materials,[46] ESI's sheer financial bulk—document review costs comprise "nearly 75 percent of the eDiscovery budget"[47]—renders predictive coding a necessary advancement in the modern legal profession.[48]

Advocates of predictive coding champion its centrality to revitalizing an efficient modern legal practice. They often concede that that the technology is not fully equivalent to human review,[49] instead arguing that predictive coding works best in mundane contexts, characterized by the easy relevancy or privilege determinations on a large number of documents, whereas humans are better at making close calls.[50] Despite this concession, a number of arguments support predictive coding's efficacy in the legal profession.

---

46. *See generally* Craig D. Ball, *About Predictive Coding: The "Not Me" Factor* (ALI-ABA Continuing Professional Education, July 2013), WL CV001 ALI-ABA 573 (discussing big data's significance for the legal profession in a variety of contexts, ranging from document production to firms' business development).

47. Elster, *supra* note 43, at 38.

48. As the Sedona Conference is a very helpful source to get a sense of what practices and technologies are used and should be used in this area, see generally The Sedona Conference Working Group, *supra* note 7, for further research.

49. *See* Ball, *supra* note 46, at 575.

50. *See id.* ("[A]lthough predictive coding isn't better at dealing with the swath of documents that demand careful judgment, it's every bit as good (and actually much, much better) at dealing with the overwhelming majority of documents that *don't* require careful judgment—the very ones where keyword search and human reviewers fail miserably."). Ball goes on to argue that most documents under review are easy calls—either "obviously relevant" or "obviously irrelevant"—making predictive coding a far more cost-effective and consistently efficient alternative in these contexts, while humans thrive with "the judgment call documents." *See id.* at 575–76 (citations omitted). *But see* The Sedona Conference Working Group, *supra* note 7, at 203 (describing "[r]esistance by [some members of] the [l]egal [p]rofession" who challenge predictive coding based on grounds including superior human capabilities, insufficient foundation for these technologies in the courtroom, and simple ignorance about how to best use automated technologies to their full potential). The decision of when predictive coding *should* be employed, rather than

Numerous factors—such as readability, the area(s) of law at issue in a particular piece of litigation, and the professional judgments made by the attorneys who effectively teach the program a particular issue[51]—impact how well predictive coding works on different documents and implicate its relative success in certain areas of law.[52] The initial "teachers" are often "attorneys with knowledge about the responsiveness of those documents."[53] While the coding program provides the ultimate review for responsiveness, it learns from the attorneys who make the initial structural responsiveness determinations and classifications.[54] However, the mechanical nature of the predictive coding programming still leaves room for human judgment. After the automated processes are designed in compliance with the particular request, "attorneys must then decide which documents to produce."[55] Attorneys' professional judgment is then necessary to determine what to do with the mechanized data and the corresponding documents: (1) manual review of sufficiently

---

when it *can* be employed, consequently remains an important area for discussion and research in the legal community. However, because it provides a meaningful analytical tool in many cases, it is important to understand the overall potential that predictive coding holds for the legal field in general before reviewing its current status with the judiciary and administrative bodies.

51. *See* Yablon & Landsman-Roos, *supra* note 16, at 638 ("This shift [to technology-assisted review] was especially true in securities cases where relevant documents were more readable—and less frequent in antitrust and intellectual property cases in which the document population was more technical and varied.").

52. *See id.* at 637–38 ("The use of technology-assisted review began around 2008 .... The underlying technology, called machine learning, had been available for decades, but only in about the last five years has the legal profession considered its use. ... This shift was especially true in securities cases where relevant documents were more readable—and less frequent in antitrust and intellectual property cases in which the document population was more technical and varied." (citation omitted)).

53. *Id.*

54. *Id.* at 639–40.

55. *Id.* at 641.

responsive documents;[56] (2) a combination of production, culling, and review depending on relevancy benchmarks;[57] or (3) production and culling based solely on relevancy benchmarks.[58] More lawyers will soon face these choices as predictive coding gains increased traction among judges, agencies, and the profession.

For instance, pre-existing legal doctrines may push towards wide-scale implementation of predictive coding protocols due to the fiscal benefits available from implementation. Under the doctrine of proportionality, parties are excused from retrieving and sharing ESI should it not be cost-effective.[59] Because predictive coding significantly reduces the cost of e-discovery,[60] it renders ESI more accessible to litigants. The cost savings provide positive and negative impacts for a variety of parties: (1) it may cause businesses to release unfavorable data in response to disclosure or discovery requests, but that same data may help challengers who would otherwise be stymied by informational inequities; (2) it may allow businesses to comply with stringent regulatory requirements at lower costs to the bottom-line; and (3) it may simply revolutionize all parties' access to the

56. This most closely maintains the traditional approach to ESI: "[h]istorically, the most commonly used discovery search process involves human reviewers looking at each item of ESI to determine whether it is relevant and, if so, whether it is privileged." 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, 208 (2013).

57. *See* Yablon & Landsman-Roos, *supra* note 16, at 641–42.

58. *See id.*

59. *See* Grossman & Cormack, *supra* note 5, at 26–27 ("Proportionality: Pursuant to Federal Rules of Civil Procedure 26(b)(2)(B), 26(b)(2)(C), 26(g)(1)(B)(iii), and other federal and state procedural rules, the legal doctrine that Electronically Stored Information may be withheld from production if the cost and burden of producing it exceeds its potential value to the resolution of the matter. Proportionality has been interpreted in the case law to apply to preservation as well as production."); *see also* Yablon & Landsman-Roos, *supra* note 16, at 663–72 (addressing "predictive coding and proportionality review").

60. *See supra* notes 12–18 and accompanying text.

amount of information available—either by promoting negotiation or by rendering more cases more suitable for trial due to the influx of evidence.[61] Intuitively, predictive coding helps producing parties satisfy the opposing sides' requests at a decreased cost.[62] These savings are particularly helpful when agencies request documents, as companies, and possibly individuals, wish to fully comply with the requests at a minimal fiscal cost.[63] Scholars also recognize, however, that automated learning and coding technologies "can be equally valuable for analyzing incoming document productions" because they allow for an incoming review triage: the protocols "rank documents by degree of responsiveness so attorneys can home in on the most important documents quickly."[64] Consequently, the concept of proportionality does not just change whether predictive coding can be

61. *See* The Sedona Conference Working Group, *supra* note 7, at 198–99 ("Lawyers of all stripes therefore have a vital interest in utilizing automated search and retrieval tools where appropriate. The plaintiff's bar has a particular interest in being able to efficiently extract key information received in mammoth 'document' productions, and in automated tools that facilitate the process. The defense bar has an obvious interest in reducing attendant costs, increasing efficiency, and in better risk-management of litigation (including reducing surprises). All lawyers, clients, and judges have an interest in maximizing the quality of discovery, by means of using automated tools that produce a reliable, reproducible and consistent product."); *cf.* Solomon, *supra* note 17, at 26 (discussing three cases in which the parties were willing to "share irrelevant documents during the predictive coding training process to achieve cost savings").

62. Solomon, *supra* note 17, at 26.

63. *See, e.g.*, Matthew Nelson, *Predictive Coding & the "Risk-Averse" Attorney: Top 3 eDiscovery & Compliance Use Cases (Part 2)*, CORPORATE COMPLIANCE INSIGHTS (July 2, 2013), http://www.corporatecomplianceinsights.com/predictive-coding-the-risk-averse-attorney-top-3-ediscovery-compliance-use-cases-part-2/. Interestingly, while parties may have negotiated alternatives in litigation and companies may be able to use discovery costs as a tool, individuals and companies are at a relative bargaining disadvantage with government agencies, which hold greater leverage.

64. *Id.*

mandated;[65] rather, it also impacts who has access to the fruits born from this technology.

Those convinced by the advantages of predictive coding must face another large hurdle: determining the accuracy and effectiveness levels that predictive coding must achieve before it gains wider acceptance. The legal field promotes cautious but vigorous advocacy for clients.[66] Before fully adopting these relatively new methodologies, the legal field must be certain that predictive coding represents an improvement on or, at the very least, a supplement to current practices. A number of studies support the assertion that predictive coding methodologies are at least equal to, if not more accurate than, traditional human review.[67] Should these studies prove an acceptable foundation for the profession's ethical standards, predictive coding may gain a substantial foothold in the future of the legal profession.

Overall, given the symbolic Mount Everest posed by the ever-growing amount of ESI in all legal spheres, predictive coding provides the metaphorical climbing poles and oxygen allowing attorneys to trek to the top "to assess cases faster and more efficiently[,] mak[ing]

65. Nevertheless, the potential for mandatory predictive coding in the judicial context is a necessary theme in current jurisprudence and this Comment. *See infra* Part III.

66. *See* MODEL CODE OF PROF'L RESPONSIBILITY pmbl. ¶¶ 2, 5 (2012).

67. *See* Maura R. Grossman & Gordon V. Cormack, *Inconsistent Responsiveness Determination in Document Review: Difference of Opinion or Human Error?*, 32 PACE L. REV. 267, 267–68 (2012); 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. 11 ¶ 61 (2011) [hereinafter Grossman & Cormack, *Technology-Assisted Review*]; Herbert L. Roitblat, Anne Kershaw, & Patrick Oot, *Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review*, 61 J. AM. SOC'Y FOR INFO. SCI. & TECH. 70, 79 (2010); Ellen M. Voorhees, *Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness*, 36 INFO. PROCESSING & MGMT. 697, 714–15 (2000).

case preparation easier, more comprehensive, and less expensive."[68] Whether it is employed in the courtroom[69] or when interacting with administrative agencies,[70] predictive coding is a new path for traditional data management and dispute resolution. What is not an option, at this point, is simply ignoring the data mountain in the room, as ESI grows exponentially and automated technologies increase in significance for all attorneys.[71]

Proponents of court-based acceptance of predictive coding point to traditional principles, such as the *Federal Rules of Civil Procedure*'s emphasis on "balanc[ing] costs and completeness" in discovery, in support of their claims.[72] Predictive coding promises a sharp increase in these existing values as it promotes both efficiency and fiscal responsibility.[73] As such, it promises a level of continuity by pursuing

68. Nelson, *supra* note 63.

69. *See infra* Part III.

70. *See infra* Part IV.

71. *See, e.g.*, Nelson & Simek, *supra* note 20, at 20 ("There is a great quote from a *Forbes* blog post by Barry Murphy that indicates why all lawyers need to understand a bit about predictive coding: 'A lawsuit can really knock a company for a loop. Imagine being sued and asked to produce all responsive information, only to find that means sifting through 10 TB of emails. The process is complicated and it can be very costly. After all, the company must somehow determine with confidence whether each and every one of those emails is relevant to the lawsuit and/or subject to attorney-client privilege. This process has become much more manageable using technology to assist the review process.' " (quoting Barry Murphy, *2012: The Year Of Technology-Assisted Review In eDiscovery*, FORBES (Jan. 17, 2012, 2:12 PM), http://www.forbes.com/sites/barrymurphy/2012/01/17/2012-the-year-of-technology-assisted-review-in-ediscovery/)).

72. Barry, *supra* note 42, at 365 ("Predictive coding can meet the FRCP standard if parties can show it is more accurate, efficient, and responsive than manual review." (citing Grossman & Cormack, *Technology-Assisted Review*, *supra* note 67, ¶ 5)).

73. *See supra* notes 17–18 and accompanying text; *see also*, *e.g.*, Barry, *supra* note 42, at 365–66 ("The [TREC Legal Track] study concluded that 'technology-assisted review can achieve at least as high recall as manual review, and higher precision, at a fraction of the review effort, and hence, a fraction of the cost.' " (quoting Grossman & Cormack, *Technology-Assisted Review*, *supra* note 67, ¶ 55).

traditional judicial ideals. Advocates include Judge Andrew Peck,[74] who has asserted that, in his "opinion, computer-assisted coding should be used in those cases where it will help 'secure the just, speedy, and inexpensive' determination of cases in our e-discovery world."[75]

At the same time, however, proponents and critics alike recognize the fundamental choices inherent in the move from manual, human-based review to artificial intelligence-based review through the use of automated predictive coding processes.[76] It is in the necessary re-education and re-orientation process that the link to this Comment's thesis lies: because predictive coding complements existing judicial and litigation values, including efficiency, and because it will require an intellectual overhaul, it is very important to study current case law trends in addition to administrative and agency responses. Studying these trends will determine if the future of predictive coding supports existing legal values and, if it does not, will guide what changes can be made during this re-shaping of the legal consciousness to ensure that the new framework supports healthy development in accordance with those generally accepted principles.

III. PREDICTIVE CODING AND RECENT CASES

Recent cases demonstrate judicial experimentation as courts gain a tentative footing in how to address and incorporate predictive

74. Judge Peck presided over the seminal predictive coding case, *Da Silva Moore v. Publicis Groupe*, 287 F.R.D. 182 (S.D.N.Y. 2012), discussed in *infra* Part III.A.

75. Andrew Peck, *Search, Forward*, LAW TECH. NEWS, Oct. 2011, at 25, 29 (quoting FED. R. CIV. P. 1).

76. Barry, *supra* note 42, at 365 ("The transition to newer, more efficient, and more accurate models of document review will require the legal field, as a whole, to undergo a process of education.").

coding in existing discovery and litigation norms.[77] This experimentation with predictive coding provides more than simply useful data points that demonstrate the significance of ESI. Rather, these cases show that predictive coding can and does work for modern legal issues and that courts are responding to these issues through traditional norms intertwined with an embrace of revolutionary technology.[78] To that end, this Part discusses key recent cases to show not only that predictive coding has been used and discussed by the judiciary, but also to demonstrate with meaningful examples that case law is creating an initial framework through which to evaluate and respond to predictive coding. Finally, this Part uses the review of these cases to get a sense of how that framework compares to similar responses from regulatory agencies in this time of legal and technological flux.

A. Da Silva Moore v. Publicis Groupe

The Southern District of New York's 2012 decision in *Da Silva Moore v. Publicis Groupe*[79] arguably represents predictive coding's most well known foray into case law. The gender discrimination case gave rise to a class action suit,[80] which provided data ripe for coding technologies. The plaintiffs' case rested on allegations that there was

77. Am. Bar Ass'n, *Computerized Review on Trial*, A.B.A. J., Apr. 2013, at 30, 30 (2013) ("[W]e return . . . with a study showing [that] TAR and predictive coding are getting increased attention in America's court system. Kroll Ontrack's 2012 analysis of 70 state or federal judicial opinions affecting electronically stored information . . . found nine percent of the total opinions discuss[ed] either predictive coding . . . or TAR.").

78. *See, e.g.*, *id.* (illustrating the advocacy by proponents, some of whom claim that "many notable e-discovery trends emerged . . . but none promise[] to change the status quo more than the line of opinions approving the use of technology-assisted review" (internal quotation marks omitted)).

79. 287 F.R.D. 182 (S.D.N.Y. 2012).

80. *Id.* at 183.

" 'systemic, company-wide gender discrimination against female PR employees.' "[81] Because this employment complaint was based not on the actions of a select group of individuals, but rather on a comprehensive practice, a number of electronic documents and data—including hiring and promotion practices along gender lines, email communications between employees and supervisors, communications regarding promotions, and any company policies—comprised the relevant discovery material.[82] At the initial stages of this case, the parties sought to address the " 'electronic discovery protocol' . . . [for] approximately three million electronic documents . . . ."[83]

Judge Peck, who served as the magistrate judge in *Da Silva Moore* and whose scholarship addresses emerging litigation technology, recognized the potential that predictive coding and related technologies possess for enhancing the efficiency of review and production processes during discovery.[84] His discussion highlighted that the technologies allow companies to fully respond to the plaintiff's legal requests while engaging in a timely and cost-feasible retrieval process.[85] In this early-stage decision, Judge Peck also acknowledged the numerous legal challenges inherent in introducing a new technology to court practice, including the role

81. *Id.* (quoting Complaint at 3, *Da Silva Moore*, 287 F.R.D. 182 (No. 11-cv-01279), 2011 WL 655226).

82. *See id.* at 183–84.

83. *See id.* at 184 (quoting Transcript of Dec. 2, 2011 Trial Conference at 7, *Da Silva Moore*, 287 F.R.D. 182 (No. 11-cv-01279), *available at* http://www.itlawtoday.com/files/2014/05/DaSilvaMoore_Publicis_TR_12-2-11.pdf).

84. *See id.* at 182–83; *see also* Peck, *supra* note 75, at 26, 29.

85. *See id.* at 183, 189–91, 193 (noting that computer-assisted discovery technologies are especially useful when there is a large amount of data and the parties want to maximize cost and time effectiveness).

Daubert v. Merrell Dow Pharmaceuticals[86] would play in the admissibility of evidence derived from this new, expert-driven technology.[87] Finally, he advocated a uniform manner in addressing *how* parties should signal their intent to use predictive coding in a particular case: "The best approach . . . is to follow the Sedona Cooperation Proclamation model . . . [and] [a]dvise opposing counsel that you plan to use computer-assisted coding and seek agreement; if you cannot, consider whether to abandon predictive coding for that case or go to the court for advance approval."[88]

By layering his academic suggestions into a judicial opinion, Judge Peck provided a logical bridge between a previously hypothetical and purely academic discussion—predictive coding's role, if any, in the courtroom—and the day-to-day efficiency and evidentiary challenges faced by modern judges. He demonstrated more than just the promise of predictive coding, which his academic research had already elucidated. Rather, through this opinion, Judge Peck provided a very real and practical example of coding's applications[89] and moved predictive coding from academic intrigue into the ever-growing array of litigators' tools.

---

86. 509 U.S. 579 (1993). For the remaining prongs of the *Daubert* standard, see Kuhmo Tire Co. v. Carmichael, 526 U.S. 137, 147–53 (1999) and Gen. Elec. v. Joiner, 522 U.S. 136, 142, 146 (1997).

87. *Da Silva Moore*, 287 F.R.D. at 184. For a more thorough analysis surrounding how predictive coding technologies implicate the *Federal Rules of Evidence* and *Daubert*, as well as an argument in favor of folding coding technologies into the existing frameworks for emerging technologies under the evidentiary rules, see Waxse & Yoakum-Kriz, *supra* note 56, at 207 ("This article examines [Federal] Rule [of Evidence] 702 and concludes that it and the *Daubert* standard should be applied to experts who testify or otherwise provide evidence before the court on discovery disputes involving these ESI search methods.").

88. *Da Silva Moore*, 287 F.R.D. at 184 (quoting Peck, *supra* note 75, at 29).

89. In light of this practicality, Judge Peck noted that while the technology is now a viable option in some cases, it is "not going to be perfect. [Rather,] [t]he idea is to make it

Da Silva Moore "recognize[d] that computer-assisted review is an acceptable way to search for relevant ESI in appropriate cases."[90] When accepting predictive coding on the facts of the case, Judge Peck outlined five main factors in support of his decision:

> (1) the parties' agreement, (2) the vast amount of ESI to be reviewed (over three million documents), (3) the superiority of computer-assisted review to the available alternatives (*i.e.*, linear manual review or keyword searches), (4) the need for cost effectiveness and proportionality under [Federal] Rule [of Civil Procedure] 26(b)(2)(C), and (5) the transparent process proposed by [the defendant].[91]

He also recognized *Da Silva Moore*'s significance and the "lessons [it held] for the future."[92] They included the importance of "cooperation among counsel," the need to train the automated programs and to validate the results through quality control processes, the importance of triaging document review based on relevancy in order to minimize costs, and the need for active participation in "court hearings" by "the parties' e[-]discovery vendors."[93] All of these individual "lessons" illustrate the legal implications courts need to address in order to embrace the ready and practical opportunities presented by predictive coding.[94]

---

significantly better than the alternatives without nearly as much cost." *Id.* at 187 (internal quotation marks omitted).

90. *Id.* at 183 (footnote omitted).

91. *Id.* at 192.

92. *Id.*

93. *Id.* at 192–93.

94. *Id.* at 193 ("What the Bar should take away from this Opinion is that computer-assisted review is an available tool and should be seriously considered for use in large-data-volume cases where it may save the producing party (or both parties) significant amounts of legal fees in document review. Counsel no longer have to worry about being the 'first' or 'guinea pig' for judicial acceptance of computer-assisted review.").

Interestingly, even Judge Peck, a strong advocate of incorporating coding technologies, recognized in initial discovery conferences that predictive coding, at least in its current form, is not an automatic answer for electronic discovery in all cases.[95] Judge Peck's approach, in the parties' initial conference, signaled a need to weigh the benefits and costs of technology, particularly in its untested state, and to determine what forms should be mixed and matched to create the most appropriate option in a particular case.[96] The issues raised in this case exemplify areas of the law that are and will continue to be subject to tension when incorporating predictive coding technologies into "typical" litigation. For instance, even though predictive coding makes vast amounts of data relatively more accessible, it does not justify carte blanche judicial access to all document caches. In *Da Silva Moore*, the parties had to agree on what data sources would be searched and subjected to the automated protocol.[97] Additionally, even though these parties shared a general agreement about predictive coding and the necessary confidence levels,[98] they disagreed on next steps and, in particular, *how* the trained system should be used, *where* the cut-off for manual review should lie, and *how* that should be balanced with the potential for

95. *Id.* at 185 (" '[P]redictive coding should be used in the appropriate case. Is this the appropriate case for it? You all [should] talk about it some more. And if you can't figure it out, you are going to get back in front of me [, Judge Peck].' " (quoting Transcript of Dec. 2, 2011 Trial Conference at 20, *Da Silva Moore*, 287 F.R.D. 182 (No. 11-cv-01279), *available at* http://www.itlawtoday.com/files/2014/05/DaSilvaMoore_Publicis_TR_12-2-11.pdf)).

96. *Cf. id.* ("Is this the appropriate case for [predictive coding]? . . . Key words, certainly unless they are well done and tested, are not overly useful. Key words along with predictive coding and other methodology, can be very instructive." (internal quotation marks omitted)).

97. *Id.* at 185–86.

98. The parties established "a 95% confidence level (plus or minus two percent) to create a random sample of the entire email collection." *Id.* at 186.

responsive, but unproduced, documents.[99] This dilemma presented the question of whether "document production is [truly] complete and correct as of the time it was made."[100] In a concrete application of his academic scholarship, Judge Peck also resolved the evidentiary rules and *Daubert* issues presented by these concerns and determined that the rules would not apply to the actual search methodology but would remain relevant if particular pieces of evidence are presented at trial.[101]

This initial acceptance signals an expedition among the judicial community into a new world of predictive coding.[102] Previously, judges were certainly aware of its existence, but had not provided guidance on its admissibility or potential value in the judicial system.[103] For now, the intellectual revolution remains in its infancy,

99. *Da Silva Moore*, 287 F.R.D. at 185 ("Rather, plaintiffs took issue with [defendant's] proposal that after the computer was fully trained and the results generated, [defendant] wanted to only review and produce the top 40,000 documents, which it estimated would cost $200,000 (at $5 per document). The Court rejected [defendant's] 40,000 documents proposal . . . [and] explained that 'where [the] line will be drawn [as to review and production] is going to depend on what the statistics show for the results,' since '[p]roportionality requires consideration of results as well as costs. And if stopping at 40,000 is going to leave a tremendous number of likely highly responsive documents unproduced, [the proposed cutoff] doesn't work.' " (alterations to internally quoted material in original) (citations omitted)).

100. *Id.* at 188 (citation omitted) (internal quotation marks omitted) (addressing Federal Rule of Civil Procedure 26(b)–(c)).

101. *Id.* at 189 ("The admissibility of specific emails [or other pieces of ESI] at trial will depend upon each email [or item] itself . . . , not how it was found during discovery. Rule 702 and *Daubert* simply are not applicable to how documents are searched for and found in discovery."). Judge Peck went on to state that questions regarding relevancy are best decided at a later date. *Id.*

102. *Id.* at 191 ("Computer-assisted review appears to be better than the available alternatives, and thus should be used in appropriate cases.").

103. *Id.* at 182–83 ("To my knowledge, no reported case (federal or state) has ruled on the use of computer-assisted coding. . . . 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

and judges, proponents, and critics of coding alike must recognize the active role they play in shaping the future of automated technologies in the legal system.

B. *Cases Following in* Da Silva Moore*'s Footsteps*

A number of cases followed the predictive coding trail begun in *Da Silva Moore*. One such case, decided mere months later, is *Global Aerospace Inc. v. Landow Aviation, L.P.*[104] Extending *Da Silva Moore*'s initial permit for predictive coding, *Global Aerospace* answered the next logical question: If one party is opposed to the use of predictive coding, can and, arguably, *should* a court mandate its use anyway?[105] As a matter of positive law, *Global Aerospace* mandated "the use of predictive coding for purposes of the processing and production of electronically stored information," while reserving the opposing party's right to later object to the coding methodology itself.[106] However, the normative question—*should* a court force this new and unfamiliar methodology on parties who may, either for personal, strategic, or intellectual reasons, prefer manual review—remains unanswered. Judges, and perhaps legislatures, must address

---

a sign of judicial approval." (quoting Peck, *supra* note 75, at 29) (internal quotation marks omitted)).

104. No. CL 61040, 2012 WL 1431215 (Va. Cir. Ct. Apr. 23, 2012).

105. A case in the northern district of Illinois also touched on these issues. *See Kleen Prods. L.L.C. v. Packaging Corp. of Am.*, No. 1:10-cv-05711, 2012 WL 4498465, at *5–6 (N.D. Ill. Sept. 28, 2012) (noting an attempt by the plaintiffs to compel the use of content-based advanced analytics to find responsive documents).

106. *Global Aerospace*, 2012 WL 1431215, at *1 ("[I]t is hereby ordered Defendants shall be allowed to proceed with the use of predictive coding . . . . This is without prejudice to a receiving party raising with the court an issue as to completeness or the contents of the production or the ongoing use of predictive coding.").

this normative issue because predictive coding is now a potential piece of litigation and discovery.[107]

Only a few months later, *In re Actos (Pioglitazone) Products Liability Litigation*[108] arrived. It too built upon the shift towards predictive coding technologies. First, *Da Silva Moore* established that a party *could* use predictive coding.[109] Then, *Global Aerospace* showed that a court could *require* a dissenting party to engage in predictive coding-based discovery.[110] Finally, *Actos* demonstrated *how* predictive coding worked in discovery.[111] *Actos* presented foundational and technical practices helpful to parties seeking to implement predictive coding in a case management order concerning ESI production.[112] These foundational practices included presenting the sources and the likely custodians that would have been helpful to achieving a comprehensive data foundation as part of the case management process.[113]

Most significantly, however, *Actos* provided a summary of its "search methodology proof of concept to evaluate the potential utility

107. A parallel issue can arise with respect to regulatory agencies—if an agency begins to allow companies to utilize predictive coding with respect to document requests, can or should that agency *require* that all companies do so? A requirement to use predictive coding would provide fiscal benefits and increase overall efficiency. However, it raises similar normative concerns as well as potential delegation issues. *See infra* Part IV. Essentially, the question of whether predictive coding can have an impact on legal matters is already answered—it has arrived and can be tailored to legal concerns. However, just because the potential has been recognized does not mean that predictive coding should be utilized in every situation—rather, policy makers and players in each realm need to evaluate predictive coding in light of the unique considerations of each context to see whether, even though it is a possible answer, it is the right one.

108. MDL No. 6:11-md-2299, 2012 WL 7861249 (W.D. La. July 27, 2012).

109. Da Silva Moore v. Publicis Grp., 287 F.R.D. 182, 191 (S.D.N.Y. 2012).

110. *Global Aerospace*, 2012 WL 1431215, at *1.

111. *Actos*, 2012 WL 7861249, at *1–4.

112. *Id.* at *1.

113. *See id.* at *1–3.

of advanced analytics."[114] This methodology, essentially laying out how these parties would employ predictive coding for electronic discovery, presents both an example of the parties' initial agreement on the details and provides a salient picture of the technicalities that need to be resolved by courts—whether there is, or should be, an official methodology adopted by the judiciary or agencies. Literature spawned in the aftermath of these opinions urged judges to develop unified and defined methodologies for automated technologies.[115] In this way, like the previous cases, *Actos* plays a dual role as an example of and as a spur to future judicial action. Each role occupies a rung in the logical ladder pushing predictive coding from academic hypotheticals to industry possibility to legal reality to, ideally, a fully and adequately integrated part of the discovery and litigation schemes.

As a final point, the line between governmental uses or responses to predictive coding and its status in case law is not as clear as it may seem. For instance, *National Day Laborer Organizing Network v. U.S. Immigration and Customs Enforcement Agency*[116] involved a Freedom of Information Act ("FOIA") request for disclosure of information from the United States Immigration and Customs Enforcement Agency.[117] Under FOIA requests, courts determine whether a search is adequate by evaluating "the methods

114. *Id.* at *3.

115. *See, e.g.*, 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, 693–701 (2013) (arguing that "the cases reveal the disagreement and uncertainty that exists for determining when and how [automated technologies] should be used" and that "[c]larifying standards will assist parties in reaching agreement earlier in the case and more easily, allowing for discovery to proceed more smoothly with less court interference").

116. 877 F. Supp. 2d 87 (S.D.N.Y. 2012).

117. *Id.* at 93.

used to carry out the search."[118] Consequently, the court's willingness to endorse automated technologies represents an important step towards the normalization of predictive coding in FOIA cases.[119] The court's decision presents, like before, the normative question of whether these technologies *should* be endorsed in this situation. Additionally, if the parties meet the normative requirements, the decision requires the secondary evaluation of predictive coding's methodological features to ensure that they are sufficient to meet the FOIA search burden faced by the agency. As such, this judicial observation about machine learning creates a signal for those who interface with government agencies under FOIA, suggesting that the future will likely involve automated systems as a natural feature of agencies' data retrieval.[120]

By showing how agencies may need to use predictive coding on the production-side, rather than receipt-side, of ESI production and discovery, *National Day Laborer* illustrates the overlapping nature of predictive coding for government agencies, which are impacted both

118. *Id.* at 96.

119. *Id.* at 109 ("[B]eyond 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.... [T]hese methods (known as ... 'predictive' coding) allow humans to teach computers what documents are and are not responsive to a particular FOIA or discovery request and they can significantly increase the effectiveness and efficiency of searches." (emphasis added) (footnote omitted)).

120. Maureen E. O'Neill, Obtaining ESI from the Government: Legal and Practical Guidance from *National Day Laborer Organizing Network*, Presentation at the 2013 National Conference on EEO Law 6 (2013), *available at* http://www.americanbar.org/content/dam/aba/events/labor_law/2013/04/nat-conf-equal-empl-opp-law/13_oneill.authcheckdam.pdf ("For parties seeking information from the government, take heed of Judge Scheindlin's directive to work cooperatively with the agency to devise effective, efficient ways to find the documents you seek. Keep in mind that some government agencies are beginning to incorporate predictive coding ... into their routine discovery programs ....").

in the courtroom and in their regulatory duties. The overlap points to a logical necessity: courts and agencies should consider each other when creating the schemes in which to envelop predictive coding. Their uses and concerns are not incompatible, and each will be better served when making the difficult choices necessary to develop this technology properly by learning from the other's mistakes and successes. This would allow for the creation of a framework best poised for the development of a legal system that maximizes predictive coding's potential while avoiding its dangers, including overreliance on technology and frameworks that misunderstand how the technology fits with existing legal obligations.

C. Progressive Casualty Insurance v. Delaney: *Enforcing Judicial Predictive Coding Norms*

In the years since the initial predictive coding cases, courts have issued decisions focusing on the pragmatic concerns that continue to illuminate how they have addressed predictive coding in discovery. One such example is *Progressive Casualty Insurance v. Delaney*,[121] which, similarly to *National Day Laborer*, involved private parties in a suit involving a government agency.[122] In this case involving underlying causes of action based on banks taken over by the FDIC,[123] the parties agreed to use keyword searching, a computer-

121. No. 2:11-cv-00678-LRH-PAL, 2014 WL 3563467 (D. Nev. July 18, 2014).

122. This Comment points out that the respective positions of agencies and courts on predictive coding may interact in cases like this and *National Day Laborer*, thus further promoting the need for collaboration and transparency to create a more efficient overall dynamic. On the other hand, the purpose of this discussion of *Progressive Casualty Insurance* is to show how a court reacted to the behind-the-scenes use of predictive coding without prior approval. However, see *infra* Part IV for a discussion of agencies' similar requirements of up-front approval before parties may employ coding technologies in data retrieval.

123. *Progressive Cas. Ins. Co.*, 2014 WL 3563467, at *1.

assisted review technology, to cull the 1.8 million potentially responsive documents to a supposedly manageable mass of 565,000 documents to review for responsiveness.[124] Progressive, the reviewing party, facing efficiency and fiscal concerns, decided to further review the documents using predictive coding.[125] Significantly, it made this determination "without seeking leave of the court to amend the [parties' previously agreed upon] ESI Order,"[126] or even informing the opposing party of its intention.[127] In addition to making these determinations without consulting the court or opposing party, Progressive also violated the agreed-upon ESI Order protocol by failing to produce the responsive documents "on a rolling basis."[128]

In response to this behavior and corresponding motions by the opposing party, the court discussed predictive coding in the aggregate,[129] describing it as an "accurate means of producing responsive ESI in discovery," particularly as compared to "ineffective tools" like "manual human review[] or keyword searches . . . ."[130] In this discussion, the court highlighted that it is an adherent to the potential for predictive coding in certain ESI cases—even noting that if parties "agree[] at the onset of this case to a predictive coding-based ESI protocol, [it] would not hesitate to approve a transparent, mutually agreed upon ESI protocol."[131] This juxtaposition with the

124. *Id.* at *2, *6.

125. *Id.* at *2.

126. *Id.*

127. *Id.* at *4 ("In this case, Progressive unilaterally developed the predictive coding methodology and implemented it without input or consultation from the FDIC-R's counsel.").

128. *Id.* at *7.

129. *See id.* at *8.

130. *Id.* The court also cited a number of publications to support the assertion that "[s]tudies show [that predictive coding] is far more accurate than human review or keyword searches[,] which have their own limitations." *Id.*

131. *Id.* at *9.

court's hypothetical support for the use of predictive coding highlights the significance of its disapproval of what happened in this case. Here, the parties only agreed to search term or manual review for document retrieval.[132] In emphasizing the problem of unilateral action in this case, the court pointed to literature showing the judicial trend towards a need for "unprecedented . . . transparency and cooperation among counsel" when using computer-assisted review since judges "typically . . . give deference to a producing party's choice of search methodology and procedures in complying with discovery requests."[133] The problem in this case was not that predictive coding had been used; rather, the problem was that Progressive used the technology without adhering to the necessary cooperative and transparency requirements "for a predictive coding protocol to be accepted by the court . . . as a reasonable method to search for and produce responsive ESI."[134] In response to Progressive's violation of these new predictive coding norms, the court required it to turn over the entirety of the 565,000 documents originally located through keyword searching, subject only to privilege restrictions.[135] This harsh result illustrates how seriously the court took the need for above-board behavior when employing coding technologies in discovery.

Significantly, the court showed a remarkable acceptance of predictive coding as something that could be appropriate and even viewed positively, as compared to other review mechanisms, in ESI cases. However, this acceptance came with strict methodological strings attached, emphasizing a focus on cooperation and transparency. Interestingly, although courts and agency responses have developed separately, these same strands appear in regulatory

132. *Id.*
133. *Id.* at *10.
134. *Id.* at *11.
135. *Id.* at *11–12.

agencies' limited publications.[136] Because similar themes emerge in each context, increased cooperation between courts and agencies regarding the specifics—such as how much transparency is needed (*e.g.*, should a party only show its methodology or should it show every document classified as responsive or non-responsive?); how much agreement is required versus if a court or agency could unilaterally order its use; what confidence intervals are appropriate in which contexts; and more—will allow each to develop a strong protocol. These strong protocols will enforce and regulate the use of predictive coding to ensure that it is a positive development and strictly regulated by the enforcing body to avoid any exploitation, such as what happened in *Progressive*.[137]

136. *See infra* Part IV.

137. Interestingly, some cases have begun to address these specifics. For example, in *In re Biomet M2a Magnum Hip Implant Products Liability Litigation*, the court addressed a party's use of predictive coding along with other technologies to see if it sufficiently met its discovery requirements. No. 3:12-MD-2391, 2013 LEXIS 84440 (N.D. Ind. Apr. 18, 2013), at *5–6. There, the court held that cooperation did not reach so far as to "requir[e] counsel from both sides to sit in adjoining seats while rummaging through millions of files that haven't been reviewed for confidentiality or privilege." *Id.* at *6. For individuals researching the metrics other courts facing predictive coding may use when evaluating these concerns, the role Sedona Conference publications play in this opinion is instructive. The court's discussion here is particularly helpful because it signals that the Sedona Conference may play a role in establishing judicial standards measuring how automated review technologies are employed. *Cf. id.* at *6–8 (discussing the parties' citations to the Sedona Conference and measuring the approach in this case against the Sedona Conference reports). While the Sedona Conference could serve as an official vehicle for courts' transparent publication of their standards for predictive coding, it remains to be seen whether there would, or could, be a universal agency vehicle to do the same. In keeping with the modest gains recognized so far for automated technologies in both of these areas, this Comment focuses on the potential transparency and metrics of individual agencies. *See infra* Part IV.

IV. PREDICTIVE CODING AND REGULATORY AGENCIES

While the jurisprudence addressing predictive coding remains in its infancy, its relative youth is offset by increased usefulness for scholars and attorneys due to its fairly broad accessibility. What judges say in these cases is not a secret. Future judges and lawyers, who will argue based on and for these new judicial rules, can see what has and has not been applied and, at the same time, what works and what does not work in the predictive coding jurisprudence. This transparency allows the profession to understand where points of contention will arise, what policy choices must be made, and what solutions are available to resolve these questions.

Unfortunately, government agencies lack full-scale transparency in their interactions with predictive coding—including both their independent use of the technology and their receptiveness to data retrieved with automated methods. This Comment argues that increasing transparency regarding agencies' procedures, uses, and concerns would allow for an information exchange to create comprehensive understandings of the potential for, and challenges surrounding, predictive coding in agencies and regulatory law. More importantly, greater transparency would also allow for a more holistic comparison between regulatory uses for predictive coding and its uses in the judicial system. Significantly, this sort of meaningful comparison would create a feasible environment for both courts and agencies to adapt their policies to best address the issues posed by automated technologies and ESI in the legal world.[138]

138. *See* Tonia Hap Murphy, *Mandating Use of Predictive Coding in Electronic Discovery: An Ill-Advised Judicial Intrusion*, 50 AM. BUS. L.J. 609, 651 (2013) ("The 2012 predictive coding cases suggest reason for concern about cost, delay, and gamesmanship."). *See generally* Andrew Gallo & Sarah Kim, *Predictive Coding: Process and Protocol*, BOS. BUS. J., Fall 2013, at 22, 22 ("This article provides an overview of predictive coding and highlights issues likely to arise when negotiating such a protocol.");

Much like discovery and litigation, ESI's exponential growth creates a mass of documents subject to potential regulatory inquiries.[139] Similarly, predictive coding's promises of lower cost and increased efficiency make it a revolutionary option to facilitate companies' and individuals' responses to regulatory inquiries.[140] The future benefits, which are even greater in regulatory contexts, may create incentives for cooperative methodologies and emphasize a shared end-goal between companies and agencies: the opportunity to manually review fewer documents, resulting in monetary and manpower savings.[141] Moreover, despite a lack of comprehensive and transparent information, evidence indicates that at least some agencies—particularly the Department of Justice, the Securities and Exchange Commission, and the Federal Trade Commission—are

Yablon & Landsman-Roos, *supra* note 16 (providing helpful summaries of what predictive coding entails, current case law and court-ordered responses, and the issues that can arise technically and doctrinally in implementation).

139. *See* THE SEDONA CONFERENCE, *supra* note 2, at 18–19 (discussing "predictive analytics and compliance").

140. *See* Nelson, *supra* note 63 ("Regulatory inquiries by . . . federal agencies might not be cause for celebration, but using predictive coding technology to respond to government inquiries more effectively and with minimal risk may be a silver lining."). However, the converse may also be true, particularly in investigations with criminal undertones—an individual's biases could (un)intentionally taint the training process and thus return inaccurate documents as relevant or exclude relevant documents, both of which would harm the investigation. *Cf.* Tania Mabrey, *Conquering Postindictment Discovery in the Digital Age*, CRIM. JUST., Summer 2012, at 51, 52 (2013) ("A drawback to this technology is that the intelligence draws from the perspective of only one or two human reviewers—so it is important that the subset is reviewed by the person who has the greatest knowledge of the case details and the types of documents that will make the strongest impact as exhibits.").

141. *See* Nelson, *supra* note 63 (highlighting the unique party relationship that is "less adversarial than typical litigation," as many corporations want to maintain good terms with government agencies and those same agencies appreciate the opportunity to decrease the number of documents they must eventually review).

allowing predictive coding to be used to respond to inquiries—although sometimes merely "on a 'case-by-case' basis."[142]

A number of factors support a deliberative approach when considering the use of predictive coding in regulatory contexts. For example, one observer highlighted that "[w]ading through [a] virtual avalanche of data can be intimidating in civil litigation, but effectively sorting through ESI in a government investigation is even more daunting, where one potentially exculpatory document may change the nature of a case."[143] Given that regulatory agencies share many of the same potential benefits and concerns surrounding predictive coding, it is helpful to see how agencies address automated methodologies to provide a meaningful contrast, and possibly comparison, with courts as both entities must reformulate their existing frameworks to adapt to machine-learning technology's growing influence. To that end, this Comment explores how three agencies—the Department of Justice ("DOJ"), the Securities and Exchange Commission ("SEC"), and the Federal Trade Commission ("FTC")—are responding to, regulating, and incorporating predictive coding technologies.

A. The Department of Justice

The Department of Justice, established in 1870, serves under the Attorney General as "the central agency for [the] enforcement of federal laws."[144] Because the Department is quite large, it is

142. *Id.*; *see also* William Kolasky, *Antitrust Litigation: What's Changed in Twenty-Five Years?* ANTITRUST, Fall 2012, at 9, 15, 17 n.85 ("The antitrust agencies have also indicated that they are open to discuss the use of predictive coding in certain cases as well.").

143. Mabrey, *supra* note 140, at 51 (examining "some of the most effective ways to leverage e-discovery technologies in a government investigation and criminal litigation: data culling, concept searching, and predictive coding strategies").

144. *Office of the Attorney General: About the Office*, U.S. DEP'T OF JUSTICE, http://www.justice.gov/ag/about-oag.html (last visited Nov. 19, 2014).

subdivided into smaller divisions that have their own particularized missions.[145] One example is the Antitrust Division, which "promote[s] economic competition through enforcing and providing guidance on antitrust laws and principles."[146] This Comment focuses on the Department's approach to predictive coding technologies through the lens of the actions and perspectives taken by the Antitrust Division.

The Department recognizes that electronic discovery provides new challenges for companies, which are now required to search and produce more documents.[147] In response, it is among the agencies that have already permitted predictive coding to be used to satisfy required document production in some situations.[148] The Antitrust

145. *See Department of Justice Agencies*, U.S. DEP'T OF JUSTICE, http://www.justice.gov/agencies/index-list.html (last visited Nov. 19, 2014).

146. *Antitrust Division: About the Division*, U.S DEP'T OF JUSTICE, http://www.justice.gov/atr/about/index.html (last visited Nov. 19, 2014).

147. TRACEY GREER, ELECTRONIC DISCOVERY AT THE ANTITRUST DIVISION: AN UPDATE 1 (2012), *available at* http://www.justice.gov/atr/public/electronic_discovery/281388.htm ("[E]lectronic productions have grown exponentially.... Today, we have single investigations that involve over 8 terabytes of data [and,] [r]ecently, the Division has completed several investigations with productions [around] ... one million records."). In some of her material published on the DOJ's website, Greer, a senior litigation counsel, states that "[t]he views presented ... are [her] own, [and] do not reflect those of the Department of Justice or the Antitrust Division." *E.g.*, *id.* However, given that the material is published by the DOJ on its website under the heading "Electronic Discovery," *see Electronic Discovery*, U.S. DEP'T OF JUSTICE, http://www.justice.gov/atr/public/electronic_discovery/ (last visited Nov. 19, 2014) (noting that "materials pertain to civil investigations only[,] ... are merely representative[,] ... [and] are subject to change" and instructing the viewer to "consult Division staff for matter-specific guidance before beginning any collection of electronic documents for production to the division" (emphasis omitted)), this Comment will operate under the logical assumption that if the DOJ sees fit to publish this paper on its website under this heading, the views in the paper are actually endorsed by the DOJ.

148. Renata B. Hesse, Deputy Assistant Att'y Gen., Antitrust Div., U.S. Dept. of Justice, IP, Antitrust and Looking Back on the Last Four Years, Presentation at the Global Competition Review 2nd Annual Antitrust Law Leaders Forum 13 (Feb. 8, 2013), *available at* http://www.justice.gov/atr/public/speeches/292573.pdf ("Another innovation

Division is an example of a division that needs access to large amounts of documents and financial information in order to determine whether any criminal wrongdoing has occurred. However, before disclosing relevant information, companies may need to first peruse broad caches of data to gain a sense of what is relevant. Illustratively, the Antitrust Division permitted the use of predictive coding to determine relevance in the "proposed merger of Anheuser-Busch InBev NV and Mexico's Grupo Modelo SAB."[149]

In doing so, the Department recognized the mutual benefits available when properly employing coding technologies. For example, these technologies allow the Department to "reduce the document review and production burden on parties while still providing the [Department] with the documents it needs to fairly and fully analyze transactions and conduct."[150] Predictive coding technologies also allow for a prioritized allocation of resources so that "only the 'really' relevant documents [are] produced."[151] These benefits mirror the fiscal and efficiency gains that the same technology promised in litigation and discovery.

However, predictive coding is not a perfect solution to the imbalances between companies or individuals and government agencies. Many of the same concerns raised in the case law also occur with regulatory inquiries: in order for predictive coding to work within the existing dynamic, parties must employ "a high degree of cooperation and transparency about the implementation and

---

we have been testing over the past several years to help streamline our process is allowing parties to use predictive coding in their document productions. . . . [W]e have allowed parties to use predictive coding in some matters already.").

149. Nelson, *supra* note 63.

150. Hesse, *supra* note 148, at 13.

151. GREER, *supra* note 147, at 4.

structure of the predictive coding process."[152] Notably, these transparency requirements apply only to the producing party, not to the agency. The Department's concern mirrors Judge Peck's discussion with the parties in the discovery conferences in *Da Silva Moore*, where he emphasized collaboration's centrality to the meaningful application of predictive coding.[153] Questions also arise in the regulatory context as to who is qualified to determine a document's relevance and whether intentional or unintentional bias could taint the retrieval process.[154] Because of these concerns and the lack of sufficient data, the Department currently requires "written modification" to the informational request when using the artificial intelligence potential for predictive coding and emphasizes the need for "cooperation, transparency, time, and hard work," along with review of the training set and the qualitative samples and the creation of a mechanism for supplementary document retrieval.[155] These concerns show the likely direction for any predictive coding agreements permitted by the Department.[156]

Significantly, despite the potential issues, the Department does not proscribe potential use of predictive coding in these inquests. Its willingness to at least consider using coding in particular cases bolsters the need for open informational exchanges about predictive

152. Hesse, *supra* note 148, at 13.

153. *See supra* note 93 and accompanying text.

154. GREER, *supra* note 147, at 4–5 ("Relevance . . . is in the eye of the beholder.").

155. *Id.* at 5.

156. *Id.* at 5 ("As an initial framework, . . . a plan [where the Division reviews the training set and the qualitative samples and a mechanism is created for supplementary document retrieval] could serve as the basis for a meaningful negotiation between the Division and the producing party."). However, without consistent publication of these agreements and increased guidance from agencies, parties seeking to create a plan will have to rely more on individual experience and give-and-take with the agency based on these concerns, rather than following clear precedents.

coding so that all agencies can put it to its highest use. Given the Department's emphasis on transparency to be sure that responding parties adequately employ coding technologies,[157] it only makes sense that the Department's similar transparency about its evaluations would allow for the development of a broad-based, efficient framework.

B. The Securities and Exchange Commission

The Securities and Exchange Commission has a comprehensive mandate "to protect investors, maintain fair, orderly, and efficient markets, and facilitate capital formation."[158] In order to effectuate its duties, the SEC requires extensive disclosures from companies to create an informational balance for investors.[159] To produce required information to the SEC, companies may have to sort through extensive caches of information and communications. As such, predictive coding may provide a more efficient retrieval system, allowing these companies to fully comply with the Commission's requirements at the lowest possible cost.

---

157. *See* Allison C. Stanton, *DOJ Director Talks About Investigations and E-Discovery Technology*, METRO. CORP. COUNSEL (Feb. 25, 2013), http://www.metrocorpcounsel.com/articles/22623/doj-director-talks-about-investigations-and-e-discovery-technology ("Transparency goes to the company's and counsel's preparation for discussions with the government. . . . If a company used predictive coding and advanced analytics before producing information, for instance, it can hurt a company's credibility if they don't tell us up front that they are planning to use these technologies."); Template, Dep't of Justice, Request for Additional Information and Documentary Material Issued to Weebyewe Corporation 10 (March 2012), *available at* http://www.justice.gov/atr/public/220239.pdf.

158. *The Investor's Advocate: How the SEC Protects Investors, Maintains Market Integrity, and Facilitates Capital Formation*, U.S. SEC. & EXCH. COMM'N, http://www.sec.gov/about/whatwedo.shtml (last visited Nov. 19, 2014).

159. *Id.* ("[T]he SEC requires public companies to disclose meaningful financial and other information to the public. This provides a common pool of knowledge for all investors to use . . . . The result of this information flow is a far more active, efficient, and transparent capital market . . . .").

Similar to the Department of Justice and Federal Trade Commission, the Securities and Exchange Commission does not have a blanket ban on predictive coding for data production. Rather, it requires specific disclosure by the requesting company and SEC approval before the company can use the technology to satisfy its regulatory guidelines.[160] What sets the Commission apart, however, and represents the next logical progression in regulatory treatment of predictive coding is the fact that it is beginning to use predictive coding software within its own systems and review processes in addition to merely accepting information selected through automated systems.[161] The Commission provides a perfect example of how agencies hold the potential to shape the way private parties use predictive coding: by accepting the fruits of these technologies when the technologies are properly employed per publically accessible guidelines, which still require Commission approval. However, the fact that a body as concerned with data and accurate production as the SEC is signaling acceptance of coding technologies provides an important foundation for private parties, as they now have an external standard by which to measure their interest in using coding in private cases. Additionally, this example could provide a yardstick by which courts could measure their own standards. This potential is strengthened by the fact that regulatory bodies, like the Commission,

160. U.S. SEC. & EXCH. COMM'N, DATA DELIVERY STANDARDS 1 (Rev. Jan. 17, 2013), *available at* http://www.sec.gov/divisions/enforce/datadeliverystandards.pdf ("Any proposed production in a format other than those identified below, the proposed use of *Predictive Coding*, *computer-assisted review* or *technology-assisted review* (TAR), or the use of de-duplication during the processing of documents, must be discussed with and approved by the legal and technical staff of the Division of Enforcement (ENF) and the methodology must be disclosed in the cover letter.").

161. *See* Ari Levy, *Recommind Lands SEC as Software Client*, SFGATE (Feb. 3, 2013), http://www.sfgate.com/default/article/Recommind-lands-SEC-as-software-client-4247554.php.

will learn how to address coding technologies through repeated interaction with it on the other side of the table as well as through its actual use.[162]

Significantly, this potential for insider familiarity with the technology and knowledge about its strengths, weaknesses, and possibilities could allow the Commission, along with other regulatory bodies, to use that familiarity to create effective guidelines for disclosures.[163] The benefit of actual knowledge is without parallel and illustrates the next logical step in regulatory bodies' acceptance of coding technologies.[164] Moreover, this knowledge can be put to even greater use if the employing agencies are willing to be transparent and share their experiences, successes, and trials not only with other agencies, but also with the courts. This transparency would allow effective rules to develop across the spectrum and also permit software providers to adopt more effective methodologies.[165]

---

162. O'NEILL, *supra* note 120, at 6 ("Keep in mind that some government agencies are beginning to incorporate predictive coding and other analytical software tools into their routine discovery programs (for example, the SEC recently acquired a license to use the Recommind predictive coding software).").

163. *Cf.* Murphy, *supra* note 138, at 657 ("Groups such as the Sedona Conference and the Seventh Circuit Electronic Discovery Program . . . might also encourage greater transparency regarding predictive coding and services, development of empirical data regarding its effectiveness, and further improvement in predictive coding technology, which may lead to wider voluntary use of that technology and perhaps more 'just, speedy, and inexpensive discovery.' " (quoting FED. R. CIV. P. 1)).

164. *But see id.* at 654–55 ("A broader question may be whether a judge or indeed any government actor should or can effectively promote public acceptance of new technology. . . . One scholar recommends that rather than mandating new technologies, government 'should identify what are the key goals or problems it is trying to address and then not discriminate against any technologies that can help achieve those stated objectives.' " (quoting Gary E. Marchant, *Sustainable Energy Technologies: Ten Lessons from the History of Technology Regulation*, 18 WIDENDER L.J. 831, 856 (2009))).

165. *Cf.* Murphy, *supra* note 138, at 657 (citation omitted). *But see* Gary E. Marchant, *Sustainable Energy Technologies: Ten Lessons from the History of Technology Regulation*, 18 WIDENDER L.J. 831, 845 (2009) ("If consumers are unwilling to accept or pay for a new

C. *The Federal Trade Commission*

The Federal Trade Commission is tasked with "prevent[ing] business practices that are anticompetitive or deceptive or unfair to consumers; . . . enhance[ing] informed consumer choice and public understanding of the competitive process; and . . . accomplish[ing] this without unduly burdening legitimate business activity."[166] Similar to the Department of Justice's Antitrust Division, the Commission's mandate necessitates perusal of large amounts of data to determine if any inappropriate behavior has occurred.[167] As such, companies targeted by the Commission for disclosure are subject to the same burdens as those faced under obligations to the other agencies and during litigation.

In a move similar to those of the Department of Justice and the Securities and Exchange Commission, the Federal Trade Commission shows preliminary acceptance of predictive coding through its response letters that encourage the use of coding in disclosures procured by informational subpoenas.[168] Interestingly, however, the Commission moved beyond merely permitting predictive coding

---

technology, that technology is unlikely to prosper. Therefore, policies that attempt to 'push' a new technology onto unreceptive or even uninterested consumers are particularly prone to fail.").

166. *About the FTC*, U.S. FED. TRADE COMM'N, http://www.ftc.gov/about-ftc (last visited Nov. 19, 2014).

167. *Cf. A Brief Overview of the Federal Trade Commission's Investigative and Law Enforcement Authority*, U.S. FED. TRADE COMM'N, http://www.ftc.gov/about-ftc/what-we-do/enforcement-authority (last updated July 2008) (laying out the breadth of the FTC's investigative powers, per statutes and regulations, and illustrating the broad powers it has to demand large quantities of information during its investigations).

168. *See, e.g.*, Letter from Donald S. Clark, Secretary, U.S. Fed. Trade Comm'n, to Seth Silber & Douglas H. Meal, Attorneys 8 (Apr. 11, 2012) [hereinafter Letter from Sec'y Clark to Silber & Meal], *available at* http://www.ftc.gov/sites/default/files/documents/petitions-quash/wyndham-hotels-resorts-llc/120411wyndhamletter.pdf.

(which raises the tensions and concerns previously discussed[169]) and instead moved towards requiring companies to, at the very least, consider predictive coding when attempting to decrease their disclosure responsibilities.[170] For example, in an April 11, 2012, response to a motion to quash or limit civil investigative demand, the FTC made a passing reference, when critiquing the petitioner's cost analysis, endorsing the affordability of predictive coding technologies in these contexts.[171] The letter critiqued the cost estimate, stating that it failed to "account for factors that may reduce the cost and time of production . . . [because] Petitioners have not sufficiently addressed the availability of e-discovery technology, such as advanced analytical tools and predictive coding, to enable fast and efficient search, retrieval, and production of electronically stored information . . . ."[172]

Lest this appear to be an isolated incident, in another response letter, the Commission addressed the correlation between ESI and an increased burden, asserting that parties need to include "affirmative suggestions [that] could include . . . predictive coding."[173] Taken in conjunction with each other, responses like these illustrate the Commission's relative acceptance of coding technologies. This step has logical parallels to the progression described in the case law. Automated technologies' mandatory acceptance, however, raises the same needs for transparency and interactive learning with other entities, given predictive coding's growing support in the regulatory community.

169. *See supra* Part IV.A.

170. A similar logical shift occurred in the case law. *See supra* Part III.

171. *See* Letter from Sec'y Clark to Silber & Meal, *supra* note 168, at 8.

172. *Id.*

173. Letter from Donald S. Clark, Secretary, U.S. Fed. Trade Comm'n, to Mark W. Nelson, Attorney 6 (May 23, 2011), *available at* http://www.ftc.gov/sites/default/files/documents/petitions-quash/w.l.gore-associates/110523quashgoreletter.pdf.

Even more importantly, the Commission has begun the process of moving predictive coding decisions from informal determinations into the actual regulatory structure.[174] This shift represents a significant step in the progressive ladder towards full-scale acceptance and implementation of predictive coding. In early 2012, the Commission proposed rule changes to 16 C.F.R. Parts 2, Nonadjudicative Procedures, and 4, Miscellaneous Rules.[175] In explaining these proposed changes, the Commission addressed the "Need for Reform of the Commission's Investigatory Process" and cited three key reasons in support of the use of automated technologies: (1) "information is no longer accurately measured in pages, but instead in megabytes," (2) "ESI[] is widely dispersed throughout organizations," and (3) "because ESI is broadly dispersed and not always consistently organized . . . , searches, identification, and collection all require special skills and, if done properly, may utilize one or more search tools such as . . . predictive coding, and other advanced analytics."[176] Notably, only the explanatory section in the Federal Register specifically mentions predictive coding.[177] However, the specific reference in the Federal Register section still provides helpful insights into how the agency views predictive coding and similar technology. Additionally, and most importantly, it signals that at least one regulatory body is willing to move beyond the intermediate stages of agency responses to meaningful rule-making that seriously considers predictive coding technologies.

174. *See* 16 C.F.R. §§ 2.2, 2.4, 2.6, 2.7, 2.9–2.11, 2.13, 2.14, 4.1 (2014).

175. *See* 77 Fed. Reg. 3191 (proposed Jan. 23, 2013) (codified at 16 C.F.R. §§ 2.2, 2.4, 2.6, 2.7, 2.9–2.11, 2.13, 2.14, 4.1 (2014)).

176. *Id.*

177. *See id.*

V. AN OVERVIEW OF INDUSTRY INSIGHTS AND MISCELLANEOUS FACTORS TO CONSIDER WHEN DEVELOPING THE RULES AND REGULATIONS SURROUNDING PREDICTIVE CODING

Much of the review and analysis in this Comment, and much of the available academic literature,[178] presupposes that private entities, or the attorneys representing them, will be the ones actually employing predictive coding software in response to disclosure inquires, whether in litigation or in regulatory inquests. Interestingly, the SEC's use of coding technologies, when considered in conjunction with *National Day Laborer*, which also touches on this issue to an extent,[179] poses an additional normative question: whether government bodies should also be allowed to use predictive coding in response to document requests. Government bodies are held to different standards than private parties,[180] and thus something that may be appropriate for a corporation or individual to employ may not be sufficient to meet certain higher burdens held by agencies. On the other hand, because the government has similar interests in producing relevant documents at low cost to personnel and budgets,[181] predictive coding provides parallel possibilities for increased speed, quality of

178. *See, e.g.*, Barry, *supra* note 42, at 364–72 (arguing that "[c]ourts [s]hould [a]dopt [p]redictive [c]oding" and providing solutions for the implementation of predictive coding in courts).

179. *See supra* notes 116–20 and accompanying text.

180. For example, they are bound to uphold the Constitution in their actions. U.S. CONST. art. VI, cl. 3.

181. *See* MARY GAY WHITMER, NASCIO, SEEK AND YE SHALL FIND? STATE CIOS MUST PREPARE NOW FOR E-DISCOVERY! 3 (2007), *available at* http://www.nascio.org/publications/documents/nascio-ediscovery.pdf ("What's at stake? If a state is involved in litigation, the outcome of the case could hinge upon the . . . retrieval of electronic information. In the event that the State CIO cannot . . . locate and retrieve discoverable information, the state could be penalized . . . [and,] [u]ltimately, a negative litigation outcome could cost substantial amounts of taxpayer dollars that might be spent on more pressing priorities." (emphasis omitted)).

review, and fiscal efficiency. However, the ways in which states and other branches of government do or do not use predictive coding will likely be influenced by the examples set by the courts and regulatory protocols. As such, increased transparency and awareness of the ramifications of these decisions are crucial for the proper development of predictive coding in a number of venues.

Another serious question is whether these technologies should be permitted in criminal cases.[182] Given that the Department of Justice is already accepting data produced through automated review,[183] this concern is far more than academic. It also brings together regulatory and judicial concerns—hypothetically, predictive coding technologies could be sufficient when an agency determines wrongdoing has occurred and pursues regulatory action against an entity but the court may not deem it acceptable for use in a judicial context. Additionally, the higher moral and punitive stakes in criminal cases particularly implicate the margin of error from automated retrieval. If the one document to prove innocence is missed because it is within the margin of error, is that acceptable in a criminal case? Should it be? Should an attorney be able to make that call for a client or should explicit understanding and consent be required from defendants? These are issues that connect regulatory and judicial action and raise salient issues that touch more than procedural concerns and may call for public commentary in addition to legal recommendations.

182. *See generally* Andrew D. Goldsmith & John Haried, *The New Criminal ESI Discovery Protocol: What Prosecutors Need to Know*, U.S. ATT'YS' BULL., Sept. 2012, at 3, *available at* http://www.justice.gov/usao/eousa/foia_reading_room/usab6005.pdf (discussing criminal ESI discovery protocol disseminated by the Department of Justice); Mabrey, *supra* note 140 (suggesting ways to employ predictive coding and automated technologies in a criminal government investigation).

183. *See supra* Part IV.A.

VI. COMPARISONS, DIFFERENCES, AND RECOMMENDATIONS FOR COURTS AND AGENCIES ADDRESSING THE USE OF PREDICTIVE CODING

A review of the seminal cases and limited information available about regulatory bodies' behavior reveals a number of parallel trends. Most fundamentally, predictive coding solves a common problem: ESI results in massive caches of data that are fundamentally incompatible with traditional human review but can be more readily and affordably accessed with technology. Moreover, among both courts and regulatory bodies, there is clear support for heightened transparency and cooperation between the parties who seek to employ predictive coding. Furthermore, shared areas of tension also arise. For example, can an unwilling party be mandated to engage in, or at the least consider engaging in, predictive coding? Is it better for decisions to be made by individuals or bodies with personal experience working with predictive coding? The latter question is particularly poignant given that while agencies may gain actual experience using, rather than regulating, coding, it is less likely that judges will do so. Additionally, actual use will help to enforce accurate expectations of what the technology is capable of but may also create other biases, including those in favor of vendors or in favor of decreased transparency by parties, who may fear that transparency could implicate privilege or attorney-work product concerns.[184] It also creates an imbalance between judges and agencies,

184. *See* Progressive Cas. Ins. Co. v. Delaney, No. 2:11-cv-00678-LRH-PAL, 2014 WL 3563467, at *10 (D. Nev. July 18, 2014) ("[L]itigators are loathe to reveal their methodological decisions [when training an automated system] for various reasons including assertions that: methodological decisions reveal work product; discovery about discovery exceeds the scope of Rule 26 of the Federal Rules of Procedure; revealing documents non-responsive to discovery requests exposes the producing party to unnecessary litigation risks; and the Federal Rules of Civil Procedure only require parties to conduct a reasonable search for responsive documents.").

as only agencies really have the power to gain this hands-on experience, since courts act as mediators while agencies can be parties to disputes involving actual data. Moreover, it raises the question of whether there should be differences between litigation and regulatory use. Finally, both contexts place significant emphasis on the need for proper training in order to create unbiased results.

There are also a number of differences between litigation and regulatory contexts. First, thus far, litigation has involved private parties and the potential for information and monetary imbalances,[185] whereas regulatory bodies can require private parties to produce information regardless of the individuals' finances and comfort.[186] This dynamic relegates the greater incentives to private parties, rather than government agencies, for incorporation of this new technology.[187] However, other factors make agencies seem to be the more attractive candidates for incorporating coding technologies. For example, agencies simply have better opportunities to gain experience around predictive coding: they deal with more eligible data caches than courts, they can use coding technology on their own initiatives, they are less bound by rigid rules of evidence and procedure, and the parties they work with have a high self-interest in promoting coding.[188] The interaction with agencies is distinguishable in that the

185. *Cf.* Losey, *supra* note 33, at 17–18 ("The use of computerized categorization techniques, such as predictive coding, will likely become the norm for large-scale reviews in the future, given the likelihood of increasing societal acceptance of artificial intelligence technologies .... The problem is that considerable sums of money are being spent unnecessarily today while attitudes slowly change over time." (citation and internal quotation marks omitted)).

186. *See* Lisa C. Wood, *Predictive Coding Has Arrived*, ANTITRUST, Fall 2013, at 93, 95.

187. *See supra* Parts III and IV (describing the use of predictive coding by private parties in litigation and government agencies, respectively).

188. *See supra* Part IV.

agencies themselves must peruse the produced data, whereas judges only act as umpires between the parties exchanging the data.[189] Judges, on the other hand, have strong interests in justice and efficiency, but mere context may render them less experienced or willing to make a determinative endorsement of particular technologies.[190] Additionally, courts are bound by the rules of evidence and procedure.[191] Judicial decisions that integrate predictive coding into a traditional and foreseeable structure have important ramifications both for the technology and how it develops—for example, what limits on its use will stunt growth in some areas and foster growth in others? A significant positive, additionally, is the great transparency judicial decisions have in this area.[192]

As such, courts and agencies should share their strengths with the other and complement the other's weaknesses. Both groups should also consider how and where to accept sufficient quality

189. Nelson, *supra* note 63 (illustrating how agencies' interactions are not a zero-sum game).

190. *See* Harrison M. Brown, Note, *Searching for an Answer: Defensible E-Discovery Search Techniques in the Absence of Judicial Voice*, 16 CHAP. L. REV. 407, 424 (2013) ("Courts have yet to embrace any of the new search technologies, instead only generally alluding to potential benefits they offer, but not going so far as to expressly endorse a particular method.").

191. *See, e.g.*, Nathan M. Crystal, *Inadvertent Production of Privileged Information in Discovery in Federal Court: The Need for Well-Drafted Clawback Agreements*, 64 S.C. L. REV. 581, 584–85, 591–92 (2013) (discussing privilege and work product); Waxse & Yoakum-Kriz, *supra* note 56, at 214–19 (addressing the interaction between automated technologies and experts in the federal rules of evidence).

192. *See* NICHOLAS M. PACE & LAURA ZAKARAS, RAND, WHERE THE MONEY GOES: UNDERSTANDING LITIGANT EXPENDITURES FOR PRODUCING ELECTRONIC DISCOVERY 98–99 (2012) (stating that "the legal world has been reluctant to embrace [predictive coding] . . . [because] of the absence of widespread judicial approval" and asserting that "the best catalyst for more-widespread use of predictive coding would be well-publicized instances of successful implementation in cases in which the process has received *close judicial scrutiny*" (emphasis added)).

control levels—at what point is the program sufficiently trained and reliable? As previously discussed, this may vary significantly depending on the legal context. This means that while agencies and courts can learn from each other's experiments, they also should be careful to acknowledge the differences between them and how these differences may make certain rules either inapplicable or materially harmful to a particular context.

CONCLUSION

Predictive coding is still in its trial stages—in all likelihood, it will not work in all contexts, and it will need guidelines in order to achieve its highest potential in appropriate cases. Because it is still developing, many questions remain as to what predictive coding *can* do and what it *should* do. The debate over how to answer these questions will only strengthen understandings of its possibilities and highlight areas of concern. Raising questions is a positive thing: by identifying early on the areas where tension and issues may arise, courts and agencies can create rules in light of these issues and play a preventative role. They can also highlight the issue areas that they may be most concerned about—such as transparency, companies' and attorneys' desire to protect confidential, non-responsive documents, and cooperation. When determining both what areas to focus their energies on and how to achieve these goals, courts and agencies should also strongly consider research by other groups, such as the Sedona Conference's *Sedona Principles Addressing Electronic Document Production*,[193] as the informational exchange will be best

193. WORKING GRP. ON ELEC. DOCUMENT RETENTION & PROD., THE SEDONA CONFERENCE, THE SEDONA PRINCIPLES: BEST PRACTICES RECOMMENDATIONS & PRINCIPLES FOR ADDRESSING ELECTRONIC DOCUMENT PRODUCTION ii (2d ed. 2007), *available at* https://thesedonaconference.org/download-pub/81.

served through this sort of transparent exchange.[194] By showing each other what predictive coding has achieved and in what contexts it has done so, both courts and regulatory bodies can use this knowledge to promote rules that identify and create the best guidelines that allow predictive coding to grow without stifling justice, impartiality, or efficiency. In order to maintain existing legal principles, or to make a conscious choice to overhaul existing norms, all bodies should be hyper-vigilant when addressing new technology to be sure that all regulatory and guiding choices prospectively promote desired future reforms, rather than resulting in forced, piecemeal changes. Predictive coding holds great promise for the future of the legal field, but it is up to the community as a whole to ensure a healthy growth environment so it can achieve its full potential while not alienating existing moral and procedural standards.

CHRISTINA T. NASUTI**

194. *See* Murphy, *supra* note 138, at 657. For example, even early on in *Da Silva Moore*, Judge Peck cited the Sedona Cooperation Model (original cited in his academic writings) when addressing the proper ways to handle predictive coding and ESI in that case. Da Silva Moore v. Publicis Groupe, 287 F.R.D. 182, 184 (S.D.N.Y. 2012) (citing Andrew Peck, *supra* note 75, at 29).

** Special thanks go to my loving family and fiancé, who have supported me through every stage of this project. I would also especially like to thank Professor Dana Remus, who introduced me to predictive coding and was the inspiration for this Comment's topic. Finally, thank you to the *North Carolina Law Review* Board and Staff, and Jennifer Little in particular, for your thorough edits and help along the way. I am sincerely grateful to you all.