Richmond Journal of Law & Technology Volume XXVI, Issue 3

HUMANS AGAINST THE MACHINES: REAFFIRMING THE SUPERIORITY OF HUMAN ATTORNEYS IN LEGAL DOCUMENT REVIEW AND EXAMINING THE LIMITATIONS OF ALGORITHMIC APPROACHES TO DISCOVERY[†]

By: Robert Keeling,[*] Rishi Chhatwal,[**] Peter Gronvall, & Nathaniel Huber-Fliflet[***]

Cite as: Robert Keeling et al., *Humans Against the Machines: Reaffirming the Superiority of Human Attorneys in Legal Document Review and Examining the Limitations of Algorithmic Approaches to Discovery*, 26 RICH. J.L. & TECH., no. 3, 2020.

[†] The authors wish to thank Adam Kleven of Sidley Austin LLP for his assistance in writing this article.

[*] Robert Keeling is a partner at Sidley Austin, LLP. He is an experienced litigator whose practice includes a special focus on electronic discovery matters. Robert is co-chair of Sidley's eDiscovery Task Force.

[**] Rishi Chhatwal is an Assistant Vice President and Senior Legal Counsel at AT&T Services, Inc., and heads AT&T's Enterprise eDiscovery group.

[***] Peter Gronvall and Nathaniel Huber-Fliflet are both Senior Managing Directors at Ankura Consulting Group. They advise law firms and corporations on advanced data analytics solutions and legal technology services.

ABSTRACT

Technological advancements are significantly influencing the legal services landscape in a myriad of respects. Automation, machine learning, and other advanced analytics are experiencing unprecedented acceptance and adoption across the legal industry. Discovery is no exception: over the past decade, the expansion of technology-assisted review has fueled speculation that attorneys will be largely replaced by machines. Commentators have prophesied an "artificial intelligence invasion" that brings about the "extinction of the legal profession." Specifically, in the e-discovery process, predictive coding has been tagged as this sort of disruptive, impactful technology that would make attorneys no longer necessary. Predictive coding is a type of technology-assisted review that employs algorithms to help classify documents (relevant or not, privileged or not, etc.). This technology has surged in popularity, supported by a conventional wisdom that posits that predictive coding is faster, cheaper, and more accurate than manual (i.e., human) review. While experience supports an emerging–and powerful–role for artificial intelligence in data review, research has not actually shown that predictive coding is simply 'better than' humans, or that predictive coding should ever be employed without human training, iteration, and final review.

This article enters this important discussion, challenging the prevailing wisdom around what predictive coding purports to do, and arguing that machines are simply not what they are promoted to be, especially in the discovery process. This study analyzes the results of prior research on predictive coding, revealing flaws and correcting misunderstandings. The article then examines new data that challenges the prevailing 'dim' view the market has towards the quality and utility of human review. The authors outline new research that tilts against that narrative; showing that human attorney review can significantly increase the quality of a document review. The data further reveals an important limitation of predictive coding: unlike humans, predictive coding cannot be a reliable tool for identifying key documents used in actual proceedings. The article concludes by surveying the significant risks inherent in relying on predictive coding to drive high-quality, legally defensible document reviews. Namely, the exclusive use of predictive coding can lead to unwanted disclosures, threatening attorney-client privilege and work-product protections. With a fulsome evaluation of predictive coding's capabilities, limitations, and drawbacks, the rise of the "robot overlords" seems less threatening.

TABLE OF CONTENTS

TABLE OF CONTENTS, CONT'D

I. INTRODUCTION

[1] Technological advancements over the past decade have fueled speculation that attorneys will soon be replaced by machines—at least when it comes to reviewing documents.[1] Parroting Shakespeare's "let's kill all the lawyers,"[2] commentators have prophesied an "[a]rtificial [i]ntelligence [i]nvasion"[3] that brings about the "extinction of the legal profession."[4] Visions of "virtual courts"[5] populated by "robot lawyer[s]"[6]

[1] *See* Dan Mangan, *Lawyers Could Be the Next Profession to Be Replaced by Computers*, CNBC (Feb. 17, 2017), https://www.cnbc.com/2017/02/17/lawyers-could-be-replaced-by-artificial-intelligence.html [https://perma.cc/J8PC-MR7L]; John Markoff, *Armies of Expensive Lawyers, Replaced by Cheaper Software*, N.Y. TIMES (Mar. 4, 2011), http://www.nytimes.com/2011/03/05/science/05legal.html [https://perma.cc/U23F-FBCL]; James O'Toole, *Here Come the Robot Lawyers*, CNN (Mar. 28, 2014), http://money.cnn.com/2014/03/28/technology/innovation/robot-lawyers/index.html [https://perma.cc/74HW-ZJ65]; Joe Palazzolo, *Why Hire a Lawyer? Computers Are Cheaper*, WALL STREET J. (June 18, 2012), https://www.wsj.com/articles/SB10001424052702303379204577472633591769336 [https://perma.cc/2K7B-U8CX]; Hugh Son, *JPMorgan Software Does in Seconds What Took Lawyers 360,000 Hours*, BLOOMBERG, https://www.bloomberg.com/news/articles/2017-02-28/jpmorgan-marshals-an-army-of-developers-to-automate-high-finance [https://perma.cc/6XXX-XYRR].

[2] WILLIAM SHAKESPEARE, THE SECOND PART OF KING HENRY THE SIXTH, act 4, sc. 2, 55 (Henry Bullen ed., Fall River Press 2012); *see* Elizabeth S. Fitch & Elizabeth Haeker Ryan, *The First Thing We Do, Let's Kill All the Lawyers*, IADC TECH COMMITTEE NEWSL. (Int'l Ass'n. Def. Couns.), Feb. 2017, at 1; *see also* Jason Koebler, *Rise of the Robolawyers*, ATLANTIC (Apr. 2017), https://www.theatlantic.com/magazine/archive/2017/04/rise-of-the-robolawyers/517794/ [https://perma.cc/3M7V-62E7] ("Let's kill all the lawyers."); Abdi Shayesteh & Elnaz Zarrini, *Man vs. Machine: Or, Lawyers vs. Legal Technology*, LAW 360 (Nov. 14, 2016), https://www.law360.com/articles/862058 [https://perma.cc/US9T-QMP6] ("The first thing we do, let's kill all the lawyers.").

[3] Fitch & Ryan, *supra* note 2, at 3.

[4] Shayesteh & Zarrini, *supra* note 2.

[5] Rachel Hall, *Ready for Robot Lawyers? How Students Can Prepare for the Future of Law*, GUARDIAN (July 31, 2017), https://www.theguardian.com/law/2017/jul/31/ready-

have caused some law firms to embrace new technologies for fear of becoming obsolete.[7] Today, many firms already use predictive coding for large document review tasks, and observers envision even less attorney involvement going forward as the technology improves.[8]

[2] The increased use of robots in the discovery process could be especially troubling for the thousands of contract attorneys in the United States.[9] Even clearinghouses for contract attorneys have acknowledged

for-robot-lawyers-how-students-can-prepare-for-the-future-of-law [https://perma.cc/27VC-YFSM].

[6] Shannon Liao, *'World's First Robot Lawyer' Now Available in All 50 States*, VERGE (July 12, 2017), https://www.theverge.com/2017/7/12/15960080/chatbot-ai-legal-donotpay-us-uk [https://perma.cc/4KXY-ZAZ6].

[7] *See* Steve Lohr, *A.I. Is Doing Legal Work. But It Won't Replace Lawyers, Yet.*, N.Y. TIMES (Mar. 19, 2017), https://www.nytimes.com/2017/03/19/technology/lawyers-artificial-intelligence.html [https://perma.cc/Q4AQ-TCUC].

[8] *See* Caroline Hill, *Deloitte Insight: Over 100,000 Legal Roles to Be Automated*, LEGALIT INSIDER (Mar. 16, 2016), https://www.legaltechnology.com/latest-news/deloitte-insight-100000-legal-roles-to-be-automated/ [https://perma.cc/Y5ZT-DP8U]; *see also* Jim Kerstetter, *Tech Roundup: Will Robots Replace Lawyers?*, N.Y. TIMES (Mar. 20, 2017), https://www.nytimes.com/2017/03/20/technology/robots-lawyers-automation-workers.html [https://perma.cc/6ZQF-XR4Z]; Kingsley Martin, *Artificial Intelligence: How Will It Affect Legal Practice—And When?*, THOMPSON REUTERS (Apr. 27, 2016), https://blogs.thomsonreuters.com/answerson/artificial-intelligence-legal-practice/ [https://perma.cc/8X9U-XUS8].

[9] *See* Anna Stolley Persky, *Under Contract: Temporary Attorneys Encounter No-Frills Assignments, Workspaces*, WASH. LAW. (Jan. 2014), https://www.dcbar.org/bar-resources/publications/washington-lawyer/articles/january-2014-contract-lawyers.cfm [https://perma.cc/Q2TM-7BJ7] (explaining that although total numbers are hard to estimate, The Posse List, which is just one clearinghouse for contract attorneys, had more than 14,000 contract attorneys actively seeking employment).

this apparent reality.[10] One job site for contract attorneys has recognized that "as the technology improved, the need for these large numbers [of contract attorneys] dwindled. But, there was an increase in the need for greater sophistication and expertise in [Electronically Stored Information] management and e-discovery."[11] The result? There are fewer contract attorneys but more forensic consultants and e-discovery companies.

[3] Whether contract attorneys are the canaries in the coal mine for document review generally remains to be seen. Attorneys are understandably risk averse and the risks of completely removing humans from document-review projects are significant. Inadvertent disclosures of confidential or privileged information to opposing counsel or compelled disclosures in the judicial or regulatory settings threaten crucial privilege and work-product protections.[12] At times, working with regulators and using predictive coding may risk forfeiting the ability to engage in meaningful manual review.[13] Other risks are more mundane but nonetheless reveal inefficiencies. For example, as a predictive-coding algorithm learns what is relevant based on key words, it can sweep in a host of wholly irrelevant documents.[14]

[10] *See generally What the Posse List Is About*, POSSE LIST, https://www.theposselist.com/the-posse-list/ [https://perma.cc/DLF5-XDRU].

[11] *Id.*

[12] *See* Dana A. Remus, *The Uncertain Promise of Predictive Coding*, 99 IOWA L. REV. 1691, 1722 (2014).

[13] *See id.*

[14] *See* Kevin D. Ashley, *Automatically Extracting Meaning from Legal Texts: Opportunities and Challenges*, 35 GA. ST. UNIV. L. REV. 1117, 1118 (2019).

[4] Current risks to attorney-client privilege and general inefficiencies aside, machine learning technology and artificial intelligence continue to march forward. "Machine learning" and "artificial intelligence" have essentially become buzzwords, so it is worth clarifying what they mean in the scope of this article. Put simply, in the e-discovery process, "predictive coding" refers to employing technology—usually an algorithm—to help classify documents (relevant to the case or not, attorney-client or work-product privileged or not, etc.).[15] "Machine learning algorithms are frequently used to build predictive models from historical data for making predictions and, by analyzing data, the algorithms can continue to improve their models and produce more accurate results."[16]

[5] In the document-review context, the hysteria surrounding artificial intelligence traces back, in large part, to early two studies that "essentially created the technology-assisted review field"[17] by positing through experiments that predictive coding, a form of technology-assisted review, could be faster, cheaper, and more accurate than manual review.[18] However, proponents of these studies have largely exaggerated the results

[15] *See* William W. Belt et al., *Technology-Assisted Document Review: Is it Defensible?*, 18 RICH. J. L. TECH., no. 3, 2012, at 1.

[16] Robert Keeling et al., *Separating the Privileged Wheat from the Chaff – Using Text Analytics and Machine Learning to Protect Attorney-Client Privilege*, 25 RICH. J. L. TECH., no. 3, 2019, at 24.

[17] Victor Li, *Maura Grossman: She's the Star of TAR*, LEGAL REBELS (Sept. 20, 2016), https://www.abajournal.com/legalrebels/article/maura_grossman_profile [https://perma.cc/R5LM-6N2Q].

[18] *See* 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., no. 3, 2011, at 11 (the "TREC Data Study"); Herbert L. Roitblat et al., *Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review*, 61 J. AM. SOC'Y FOR INFO. SCI. & TECH. 70, 70 (2010) (the "Four Team study").

of these studies in favor of artificial intelligence.[19] Research has not actually shown that predictive coding is better than the alternatives,[20] and it certainly has not shown that predictive coding can ever be employed without human inputs, monitoring, and final review.[21] Furthermore, as alluded to, relying on predictive coding can increase the risk of disclosing sensitive, confidential, or privileged information to opposing counsel,[22] thus threatening privilege and work-product protections. A closer look at the capabilities and drawbacks of predictive coding reveals that the automation of the legal profession may not be quite as imminent as it seems.

[6] This article challenges the prevailing wisdom of predictive coding's current capabilities, arguing that machines are not all they are purported to be in the discovery process. While some scholars have generally noted that predictive coding raises a host of questions that need to be answered,[23] or have specific concerns like predictive coding's

[19] *See* William C. Dimm, Predictive Coding: Theory and Practice 8 (Dec. 8, 2015) (unpublished manuscript), http://www.predictivecodingbook.com/ [https://perma.cc/75Y3-TQHB] (noting that the TREC Data study is often cited "to justify things that it did not actually measure or claim").

[20] *See id.* at 10 ("It would certainly be easier to justify the use of predictive coding if it could be proven to always produce results that are at least as good as exhaustive manual review. No such proof will ever exist.").

[21] *See id.* at 117–18 ("In any event, one should *not* conclude that the study proves that precision values for TAR [technology-assisted review] predictions are so high that it is unnecessary to review documents before producing them—the precision of the TAR predictions (without subsequent human review) wasn't measured in the study.").

[22] *See* Dana A. Remus, *The Uncertain Promise of Predictive Coding*, 99 IOWA L. REV. 1691, 1716–17 (2014).

[23] *See generally* Charles Yablon & Nick Landsman-Roos, *Predictive Coding: Emerging Questions and Concerns*, 64 S.C. L. REV. 633, 635–36 (2013).

impacts on due process norms,[24] few scholars have countered the notion that predictive coding is better than human review. To date, it is not, by most of the more important metrics.[25] To be sure, predictive coding has many benefits. In many cases, predictive coding can:

- reduce the total number of documents that need to be reviewed by removing clearly irrelevant ones;
- increase the percentage of documents that are actually responsive to a discovery request;
- improve the speed of a review compared to linear review;
- reduce the costs of the review; and
- return more consistent results.[26]

The authors of this article employ predictive coding in almost all of our respective matters. In our experience, predictive coding offers significant benefits for our clients. In addition, the authors have conducted substantial research into predictive coding using real-word data sets, including how to improve predictive coding settings[27] and deploy predictive coding to identify documents protected by attorney-client privilege.[28] Accordingly,

[24] *See generally* Seth Katsuya Endo, *Technological Opacity & Procedural Injustice*, 59 B.C. L. REV. 821, 823 (2018).

[25] *See infra* Part II (B)(2)(b) (showing that predictive coding struggles to identify a substantial portion of documents used as trial exhibits, which are currently identified by actual attorneys).

[26] *See, e.g.*, BOLCH JUDICIAL INSTITUTE, Technology Assisted Review (TAR) Guidelines, Duke Law School (Jan. 2019), at iv, 40; *see also* Keeling et al., *supra* note 16, at 27–28, 42–43.

[27] *See* Keeling et al., *Using Machine Learning on Legal Matters: Paying Attention to the Data Behind the Curtain*, 11 HASTINGS SCI. & TECH. L. J. 9, 14 (2020).

[28] *See, e.g.*, Keeling et al., *supra* note 16, at 24.

the goal of this article is to not to show that predictive coding is without value. Far from it. However, our more recent research suggests that human review maintains advantages over predictive coding in certain important respects. In particular, human review can greatly increase the quality of a review and confirm the limitations of predictive-coding models, *i.e.*, manual review identifies the documents predictive-coding models anticipate they will miss. In short, current studies are unambiguous that predictive coding is not a panacea for discovery; human review deserves proper billing as indispensable to the discovery process.

[7] This article proceeds in three parts. Part II provides a brief overview of predictive coding, its use in document review, and introduces two predictive-coding measures: recall and precision.[29] Part III analyzes the results of previous studies, revealing flaws and correcting misunderstandings that courts and regulators should bear in mind when assessing predictive coding's capabilities.[30] It then examines new data that challenges the idea that manual review hinders the quality of a document review that incorporates the use of predictive coding.[31] Here, the article highlights new research showing that manual review, especially with subject-matter experts, can achieve near-100% precision when combined with predictive coding.[32] Further, the data reveal an important limitation of predictive coding: unlike humans, predictive coding cannot be a reliable tool for identifying key documents used at depositions or trial.[33] Finally, Part IV turns to the significant risks underlying any use of predictive

[29] *See infra* Part I.

[30] *See infra* Part II (A).

[31] *See infra* Part II (B).

[32] *See id.*

[33] *See id.*

coding.[34] It explains that the exclusive use of predictive coding can lead to unwanted disclosures in litigation and regulatory enforcement proceedings, thus threatening attorney-client privilege and work-product protections.[35] The upshot is that there are a host of reasons to press pause before surrendering all of document review to the machines.

II. Using Predictive Coding for Document Review

[8] Predictive coding is a form of technology-assisted review (TAR)[36] in which supervised machine learning techniques are used to automatically classify documents into predefined categories of interest like relevant or non-relevant, or privileged or not privileged.[37] "Although there are different TAR software, all allow for iterative and interactive review."[38] Generally, the first step to developing the algorithm's 'intelligence' to conduct a document classification is for human attorneys to manually code a set of documents as relevant or non-relevant.[39] A supervised machine learning algorithm then analyzes these training documents to draw inferences (including words and combinations of words) that determine a

[34] *See infra* Part III.

[35] *See id.*

[36] *See, e.g.*, Keeling et al., *supra* note 16, at 23 (some of the quoted material uses the acronym TAR, but in an effort to avoid over using acronyms, this article uses "predictive coding.").

[37] *See id.* at 24; Grossman & Cormack, *supra* note 18, at 3; Maura R. Grossman & Gordon V. Cormack, *The Grossman-Cormack Glossary of Technology-Assisted Review*, 7 FED. CTS. L. REV. no. 1, 2013, at 26 [hereinafter *Grossman-Cormack Glossary*].

[38] BOLCH JUDICIAL INSTITUTE, *supra* note 26, at 1.

[39] *See* Keeling et al., *supra* note 16, at 24–25; BOLCH JUDICIAL INSTITUTE, *supra* note 26, at 1, 16.

document's relevance.[40] The result of this analysis phase is a predictive model that can be applied to each of the remaining documents in the data set to evaluate their relevance.[41] If necessary, additional human review can be used to help improve the predictive model's performance.[42] Each of these steps in the model-building process is human intensive; often attorneys will work with subject-matter experts to ensure that these early coding decisions are accurate.[43]

[9] If used correctly, predictive coding can be a powerful tool in the document-review process. It can save significant amounts of time in a variety of circumstances by quickly separating the documents most likely to be relevant from those that are not.[44] Even when attorneys are prepared to review the entire document set manually by hand, predictive coding helps prioritize the order in which the documents are reviewed.[45] By starting with the documents predicted to be most relevant, human reviewers can quickly develop an understanding of the case and decide how to proceed before expending more resources on document review.[46]

---

[40] *See* Keeling et al., *supra* note 16, at 24–25; BOLCH JUDICIAL INSTITUTE, *supra* note 26, at 1, 16; *see also* Maura R. Grossman & Gordon V. Cormack, *Comments on "The Implications of Rule 26(g) on the Use of Technology-Assisted Review"*, 7 FED. CTS. L. REV. no. 1, 2014, at 289–291 (explaining how an algorithm is continuously retrained as the human reviewer codes documents).

[41] *See* Keeling et al., *supra* note 16, at 24–25; BOLCH JUDICIAL INSTITUTE, *supra* note 26, at 1, 3, 17.

[42] *See* Grossman & Cormack, *supra* note 40, at 290.

[43] *See Grossman-Cormack Glossary*, *supra* note 37, at 31.

[44] *See* Dimm, *supra* note 19, at 3.

[45] *See* BOLCH JUDICIAL INSTITUTE, *supra* note 26, at 31.

[46] *See id.*

Generally, though, attorneys decide to forego exhaustive human review of documents that the predictive-coding process has predicted will be irrelevant, instead attorneys typically conduct sampling to make sure they are truly irrelevant to the matter at hand or otherwise validate the predictive coding model.[47]

[10] When using predictive coding to produce documents in litigation, the effectiveness of a predictive-coding process is generally evaluated by two metrics: "precision" and "recall." Precision is the portion of documents predicted to be relevant that actually are relevant.[48] A high precision ratio means that relatively few irrelevant documents were erroneously flagged by the predictive model as relevant.[49] Meanwhile, recall is a measure of how effectively the predictive model identified the relevant documents in the data set.[50] A high recall ratio means that most of the relevant documents were identified by the predictive model.[51] If a sample of the documents predicted by the model to be irrelevant includes a

[47] *See* Jennifer Kennedy Park & Scott Reents, *Use of Predictive Coding in Regulatory Enforcement Proceedings*, 81 U.S. L. WK. 191 (Aug. 7, 2012), https://www.clearygottlieb.com/-/media/organize-archive/cgsh/files/publication-pdfs/use-of-predictive-coding-in-regulatory-enforcement-proceedings-park-reents.pdf [https://perma.cc/H7TF-TJNY] (explaining that "documents that the computer predicts to be relevant are typically subjected to comprehensive human review prior to production, while documents predicted to be irrelevant are given less costly review treatment to verify irrelevance or withheld from production without further review"); *see also* BOLCH JUDICIAL INSTITUTE, *supra* note 26, at 20 (demonstrating more than one source of lawyers using this sampling technique).

[48] *See* BOLCH JUDICIAL INSTITUTE, *supra* note 26, at 6, 41; *see also Understanding Precision and Recall*, LEXBE, https://www.lexbe.com/blog/ediscovery-processing/understanding-precision-and-recall/ [https://perma.cc/7UXR-523U].

[49] *See Understanding Precision and Recall*, *supra* note 48.

[50] *See id.*

[51] *See id.*

high proportion of relevant documents, then recall would be considered rather low.[52] It usually follows in this scenario, that further coding and human inputs are needed to improve the performance of the model.

[11] As between the two measures of precision and recall, recall is more important to attorneys, regulators, and courts because it measures whether predictive coding is actually identifying the responsive documents.[53] In other words, recall is the truest indicator of whether the technology is doing what it is purposed to do. In order for predictive coding to be defensible in court, the resulting recall of a predictive model should be high enough to fulfil the requesting party's document needs.[54] Typically, a recall of 75% to 80% is appropriate for responding to discovery requests.[55] For example, using a predictive model with 75% recall would result in the requesting party receiving 75% of the documents in the collection that were relevant to the document request.

[12] However, precision is also crucial for controlling the costs of document review,[56] as well as reducing the risks associated with the inadvertent production of sensitive non-relevant documents.[57] A high precision ratio means that the predictive model correctly predicted most of the documents were relevant and less manual review is required to address

[52] *See id.*

[53] *See id.*

[54] *See id.*

[55] *See* Dimm, *supra* note 19, at 75.

[56] *See id.* at 84.

[57] *See infra* Part III.

the error of the predictive model's forecasts.[58] For example, a predictive model with a precision ratio of 95%, an excellent result, would result in an incorrect prediction for 5 out of every 100 documents. In a document review with 100,000 documents predicted to be relevant by the predictive model, this precision ratio amounts to 5,000 documents that attorneys need to find to overturn their coding to not relevant. A low precision ratio means that a predictive model requires attorneys to conduct *more* quality control review than a higher precision model to enhance the relevance of a production or identify sensitive irrelevant documents.[59]

III. Limitations of Past Research on Predictive Coding & New Research Showing Manual Review Is Better Where It Counts

[13] Previous studies on predictive coding have fundamental limitations, and the results of the prior studies have been misunderstood. Prior research does not support the notion that predictive coding holds an absolute advantage over manual attorney review.[60] In fact, prior studies are much more modest than currently perceived, and they contain built-in biases that skew their results to favor predictive coding.[61] Moreover, new research rebuts previous studies, suggesting opposite findings: human review can be better than predictive coding in particular settings.[62] Part II (A) addresses the two main studies relied on by predictive coding's

[58] *See Understanding Precision and Recall*, LEXBE, https://www.lexbe.com/blog/ediscovery-processing/understanding-precision-and-recall/ [https://perma.cc/XZ26-P6L7].

[59] *See* Grossman & Cormack, *supra* note 18, at 9.

[60] *See infra* Part II (A).

[61] *See id.*

[62] *See infra* Part II (B).

proponents and explains why courts should not put too much stock in the studies as an accurate reflection of predictive coding's performance relative to human review.[63] Part II (B) turns to the new research that shows manual review, especially with subject-matter experts, can achieve near-100% precision when combined with predictive coding.[64] Importantly, this research indicates that manual review confirms the results of predictive-coding models by identifying documents that the models themselves anticipated they would probably miss.[65] A second set of data, meanwhile, exposes an important limitation on predictive coding's capabilities—the inability to reliably identify key documents for depositions and trial.[66]

A. Correcting Misunderstandings about Prior Research on Predictive Coding's Capabilities

[14] The idea that predictive coding is superior to manual review by attorneys came about largely because of two academic studies.[67] The first of these studies (the "Four Team study"),[68] concluded that predictive coding "was at least as accurate (measured against the original [manual]

[63] *See infra* Part II (A).

[64] *See infra* Part II (B).

[65] *Infra* Part II (B).

[66] *See infra* Part II (B).

[67] *See* Thomas C. Gricks III & Robert J. Ambrogi, *A Brief History of Technology Assisted Review*, L. TECH. TODAY (Nov. 17, 2015), http://www.lawtechnologytoday.org/2015/11/history-technology-assisted-review/ [https://perma.cc/7TXJ-CFXL].

[68] *See infra* Section II (A)(1) (comparing two human teams to two predictive-coding teams).

review) as that of a human re-review."[69] The second study (the "TREC Data study")[70], went a bit further and concluded that predictive coding "can (and does) yield more accurate results than exhaustive manual review, with much lower effort."[71] These studies significantly increased the legal profession's confidence in the use of predictive coding. Before turning to the details of the prior research, though, some general background on the monumental *Da Silva Moore v. Publicis Groupe* decision in 2012 is in order.

[15] The *Da Silva Moore* case involved a putative class action in which plaintiffs were suing "one of the world's 'big four' advertising conglomerates," Publicis Groupe.[72] Plaintiffs alleged that the company had engaged in "systematic, company-wide gender discrimination. . . ."[73] The 2012 Order came about as a result of a jurisdictional dispute between the parties and the jurisdictional discovery that ensued.[74] The parties had clashed over the appropriate methodology by which to evaluate over 3-million emails from the defendants.[75] Plaintiffs argued that the defendants had no standards for assessing the accuracy of defendants' predictive-

[69] Dimm, *supra* note 19, at 8.

[70] *See infra* Section II (A)(2) (This study utilizes data from the Text Retrieval Conference, or TREC).

[71] *See* Grossman & Cormack, *supra* note 18, at 48.

[72] Da Silva Moore v. Publicis Groupe, 287 F.R.D. 182, 183 (S.D.N.Y. 2012).

[73] *Id.*

[74] *See id.*

[75] *See id.* at 184–87.

coding process, rendering it unreliable, and they objected to a plan to review only the top 40,000 emails from the predictive-coding process.[76]

[16] The court first held that plaintiffs' concerns regarding reliability were premature because of the dearth of information about how many relevant documents would be produced and at what cost.[77] The court then went into a discussion of "further analysis and lessons for the future," concluding that it "appear[ed] to be the first . . . Court [that] ha[d] approved the use of computer-assisted review."[78] The court had "determined that the use of predictive coding was appropriate considering" five factors: "(1) the parties' agreement, (2) the vast amount of ESI to be reviewed . . . (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 . . . and (5) the transparent process proposed by [defendant]."[79] The court did not, however, wholly adopt certain specifics in the defendants' proposal, namely their 40,000 document cut off and their proposal to decide in advance that seven rounds of training for the predictive model would be sufficient.[80] Again, these issues related to proportionality, "better decided

[76] *See id.* at 185–87.

[77] *See id.* at 189. The court also dismissed plaintiffs' arguments under Federal Rule of Civil Procedure 26(g) (regarding certifying that a document production is "complete"), and that Federal Rule of Evidence 702 and Daubert standards applied to predictive coding. *See id.* at 188–89.

[78] *Da Silva Moore*, 287 F.R.D., at 189–93, 93.

[79] *Id.* at 192.

[80] The court accepted the proposals for seven initial iterative rounds of training, with the caveat that additional rounds may be required. *See id.* at 185, 187.

'down the road,' when real information [would be] available to the parties and the Court."[81]

[17] The Southern District of New York's *Da Silva Moore* decision relied heavily on the Four Team Study and the TREC Data Study.[82] For the first time, a court approved the use of predictive coding for document review, "opening the door to a sea of change in how lawyers conduct e-discovery."[83] Emboldened by *Da Silva Moore*, an increasing number of courts and commentators began promoting predictive coding and encouraging its use in litigation.[84] Courts have also relied on the TREC Data Study in particular, as well as *Da Silva Moore*'s discussion of various studies, in endorsing predictive coding,[85] and "many other courts have encouraged its use, or commented on its availability to potentially

[81] *Id.* at 187.

[82] *See* Gricks & Ambrogi, *supra* note 67; *see also Da Silva Moore*, 287 F.R.D., at 189–190.

[83] *See* Gricks & Ambrogi, *supra* note 67.

[84] *See* Robert Hilson, *A Visual Representation of Predictive Coding Case Law*, LOGIKCULL (Sept. 23, 2015), http://blog.logikcull.com/a-visual-representation-of-predictive-coding-case-law (depicting *Da Silva Moore* as the center of the predictive coding "solar system") [https://perma.cc/R6JD-GXRK].

[85] *See, e.g.*, City of Rockford v. Mallinckrodt ARD Inc., 326 F.R.D. 489, 492–93 (N.D. Ill. 2018); Progressive Cas. Ins. Co. v. Delaney, No. 2:11–cv–00678-LRH-PAL, 2014 WL 3563467, at *8 (D. Nev. May 19, 2014); Dynamo Holdings Ltd. Psh'p v. Comm'r of Internal Rev., No. 2685-11, 2016 WL 4204067, at *4–6 (U.S Tax Ct. July 13, 2016); *cf.* Nat'l Day Laborer Organizing Network v. U.S. Immigration & Customs Enforcement Agency, 877 F. Supp. 2d 87, 107 n.103 (S.D.N.Y. 2012) (noting that "verification tests using sophisticated search techniques . . . would have given the Court significantly more confidence regarding the adequacy of" a party's manual review).

reduce cost and burden" even if not specifically mandating "its use in the particular case."[86]

[18] The studies, however, contain a number of limitations that cast doubt on the full extent of their conclusions.[87] Indeed, a few thoughtful observers have noted that predictive coding's rise in popularity may be based, at least in part, on a faulty understanding of predictive coding's capabilities.[88] For example, the Four Team study did not establish a reliable "gold standard" to which human and predictive coding results could be compared.[89] The TREC Data study, meanwhile, created a number of built-in advantages in their study, resulting in an inherent bias in favor of predictive coding over human review.[90] Many of these limitations were noted by the studies' authors but have rarely been mentioned by those pointing to the studies as proof of predictive coding's superiority.[91]

[86] *See* The Sedona Conference TAR Case Law Primer, 18 SEDONA CONF. J. 1, 15, 16–17 (2017) (surveying post-*Da Silva Moore* cases).

[87] *See* Dimm, *supra* note 19, at 8; William Webber, *Re-examining the Effectiveness of Manual Review*, PROC. SIGIR 2011 INFO. RETRIEVAL FOR E-DISCOVERY (SIRE) WORKSHOP 1–8 (July 28, 2011), http://users.umiacs.umd.edu/~oard/sire11/papers/webber.pdf [https://perma.cc/74AR-SS4K].

[88] *See* Bill Speros, *Despite Early Success, Technology Assisted Review's Acceptance is Limited by Lack of Definition*, ACEDS NEWS (Aug. 31, 2016), http://www.aceds.org/news/305930/ [https://perma.cc/24B9-4V38].

[89] *See* Dimm, *supra* note 19, at 8.

[90] *See id.* at 121.

[91] *See* Speros, *supra* note 88.

[19] In fact, proponents of predictive coding have often exaggerated the results of the predictive-coding studies.[92] The influential *Da Silva Moore* opinion relied heavily on the TREC Data study, but the court often categorized the study as "proof-of-capability rather than as proof-of-concept."[93] *Da Silva Moore* slightly embellished the study's results, describing predictive coding as "better than the available alternatives," and repeatedly arguing that it "should be used."[94] The design defects in the study, however, prevent it from going quite that far.[95] The study stems from another separate experiment and concerned whether predictive coding could ever produce results superior to human review in a particular, largely unrealistic setting.[96] The TREC Data study does not conclude that predictive coding holds an absolute advantage over human review.[97] To be fair, *Da Silva Moore* also included some language indicating that predictive coding was not yet perfect and was not necessarily appropriate for every case.[98]

---

[92] *See* Dimm, *supra* note 19, at 8 (explaining potential shortcomings of an often-cited study which claims technology-assisted review yields increased accuracy).

[93] *See* Speros, *supra* note 88.

[94] *Da Silva Moore*, 287 F.R.D., at 183, 189, 191.

[95] *See* Speros, *supra* note 88 ("[W]hile TAR has been patented, promoted, and demonstrated, TAR lacks a definition of what it does, in what conditions it works…").

[96] *See generally* Bruce Hedin et al., *Overview of the TREC 2009 Legal Track*, NAT'L INST. OF STANDARDS & TECH. (2009), http://trec.nist.gov/pubs/trec18/papers/LEGAL09.OVERVIEW.pdf [https://perma.cc/5H7D-AF7Q].

[97] *See infra* Part II (A)(2).

[98] *See Da Silva Moore*, 287 F.R.D., at 190–191 (describing keyword searches as a "child's game of 'Go Fish,'" "often . . . over-inclusive" by including too many irrelevant documents, and "usually . . . not very effective.")..

[20] However, advocates of artificial intelligence[99] overstated *Da Silva Moore*'s conclusions as well.[100] They reported that *Da Silva Moore* had "ordered" or "required" the use of predictive coding.[101] This kind of exaggeration about the results of the predictive-coding studies and about the outcome of judicial decisions has led to widespread view of the superiority of predictive coding, even though the technology's shortcomings are not widely understood, or even acknowledged.[102]

[21] Neither of the prior predictive-coding studies provides support for the idea that predictive coding is always superior to manual review; in fact, neither study was designed to answer that question.[103] The Four Team study seeks to determine whether "using machine categorization can be a reasonable substitute for human review."[104] Although the TREC Data study is concerned with whether predictive coding "can also yield results superior to those of exhaustive manual review,"[105] its conclusion is not absolute and is rooted in certain design flaws of the research. A final important issue is worth raising: these studies do not reflect the reality on

[99] *See* Speros, *supra* note 88 ("not only vendors who sell [predictive-coding] products or services but those whose reputation and celebrity is enhanced by [predictive coding's] promotion").

[100] *See id.*

[101] *See id.*

[102] *See id.*

[103] *See generally* Roitblat, et al., *supra* note 18 (comparing two human teams to two predictive-coding teams in the Four Team study); Grossman & Cormack, *supra* note 18 (utilizing data from the Text Retrieval Conference, or TREC in the TREC Data study).

[104] Roitblat, et al., *supra* note 18, at 70.

[105] Grossman & Cormack, *supra* note 18, at 2.

the ground for document review.[106] That is, though they compare *exhaustive* manual review to predictive coding, in fact, nobody really engages in exhaustive manual review.[107] A more relevant comparison would be predictive coding alone compared to human review using search terms and/or predictive coding augmented by analytics (about which there are no studies).

[22] A closer look at each study paints a clearer picture of what has actually been proven about predictive coding's capabilities.

1. The Four Team Study Does Not Show that Predictive Coding Is Superior to Manual Review

[23] The Four Team study's goal was far more modest than discerning whether predictive coding is superior to manual review. Rather, it sought to understand whether predictive coding could save time and money in legal discovery.[108] Under the Federal Rules of Civil Procedure, discovery processes have to be reasonable and not unduly burdensome.[109] The study "intended to investigate whether the use of technology is reasonable in this sense,"[110] i.e., whether the use of technology would be reasonable and make discovery less burdensome. The study argued that if exhaustive human review is commonly accepted as reasonable, then a predictive-

[106] *See* Roitblat, et al., *supra* note 18, at 70; Grossman & Cormack, *supra* note 18, at 2.

[107] *See* Roitblat, et al., *supra* note 18, at 70; Grossman & Cormack, *supra* note 18, at 2.

[108] Roitblat et al., *supra* note 18, at 70.

[109] FED. R. CIV. P. 26.

[110] Roitblat et al., *supra* note 18, at 72.

coding review that could achieve comparable results at lower cost should also be considered acceptable.[111]

[24] The study's conclusion that the use of predictive coding is reasonable is well taken. Although this study further concludes that predictive coding is just as accurate as human review, that conclusion is flawed for two reasons. First, the conclusion stems from a faulty measurement. Second, the study's own findings, which flow from that faulty measure, do not establish that predictive coding is better than human review on the metric that matters most (recall). This is so, even despite the fact that predictive coding had an inherent advantage in the design of the study. Accordingly, before exploring these limitations, a review of the study's design is necessary.

[25] The study compared five document reviews, one "real" and four experimental, to evaluate whether predictive coding could produce results similar to manual review.[112] Attorneys conducted the "real" review as part of an antitrust discovery request concerning a large corporate merger.[113] The reviewing attorneys looked at about two-million documents by hand and determined that 9.46% of them were relevant.[114]

[26] For the study, researchers took a sample of 5,000 documents from the original review, 9.8% of which were originally identified as relevant.[115] Two human teams ("Team A" and "Team B") then reviewed

[111] *See id.* at 79.

[112] *See id.* at 73.

[113] *See id.*

[114] *See id.*

[115] Roitblat et al., *supra* note 18, at 73.

this sample of 5,000 documents.[116] Two predictive-coding service-provider teams ("Team C" and "Team D") reviewed the majority of the two-million documents from the original review.[117] Yet, because Teams A and B "both reviewed the same 5,000 documents in preparation for one of the processes of one of the two service providers," the decisions made by that service provider were "not completely independent of the decisions made by the re-review teams."[118]

[27] After Teams A and B had completed their reviews, the documents that they disagreed upon were sent to a neutral adjudicator.[119] This adjudicator was a senior litigator for one of the corporations involved in the merger, but he did not participate in the original review.[120] The senior litigator chose which documents he thought were relevant, providing an authoritative benchmark, a gold standard, against which to measure the review teams.[121] Compared against the adjudicated results, Team A had a recall of 77.1% and a precision of 60.9%.[122] Team B had better recall, 83.6%, but worse precision, 55.5%.[123] These measures for Teams A and B

[116] *See id.*

[117] *See id.*

[118] *id.*

[119] *See id.* at 74.

[120] *See id.*

[121] *See* Roitblat et al., *supra* note 18, at 74; Grossman & Cormack, *supra* note 18, at 15.

[122] *See* Grossman & Cormack, *supra* note 18, at 16.

[123] *See id.*

are both generally acceptable for human reviews and would likely be defensible in court.[124]

	Recall	**Precision**
Team A (human)	77.1%	60.9%
Team B (human)	83.6%	55.5%

Table 1: Recall and precision of human review teams in the Roitblat, Kershaw, and Oot study, measured against the adjudicated results from the senior litigator.[125]

[28] The study, however, does not provide a true comparison between human review and predictive-coding review; namely, it does not report how Teams C and D performed relative to the adjudicated results from the senior litigator. According to the study, one of the predictive-coding teams had access to the adjudicated results *from the human review* and based its coding decisions on that information.[126] Thus, any comparison against the adjudicated results is severely skewed towards this "compromised" team, which knew and could then build off of the adjudicated results.[127] Notably, the study does not report how the other, "non-compromised" predictive-coding team performed against the adjudicated results. The study instead used the original review, rather than the adjudicated results by the senior litigator, as the "gold standard" against which to measure Teams A, B, C, and D.[128] From this faulty standard, the study concluded that "[o]n every measure, the performance of the two computer systems was at least as

[124] *See id.* at 15.

[125] *See id.* at 16 (citing Roitblat, et al., *supra* note 18, at 74).

[126] *See id*. at 7.

[127] *See id*.

[128] *See id.* at 13; Webber, *supra* note 87, at 2.

accurate (measured against the original review) as that of a human re-review."[129]

[29] Most importantly, the study's conclusion is based on just that: a faulty measurement taken against the original human review rather than the adjudicated results from the senior litigator.[130] The original review is a poor benchmark and can hardly be considered a "gold standard." Given the variability of human reviewers, the best benchmark against which to measure a document review is a determination of relevance by a small team of knowledgeable attorneys.[131] Using a small group of individuals with some expertise in the subject matter provides a more consistent, authoritative criterion for evaluating the review. Instead, the study's gold standard is a compilation of results from more than 200 attorneys from the original review, which is most likely "subject to the same variability in human assessment that the study itself demonstrates."[132] Because of this, "it is impossible to say whether any disagreements between the predictive coding results and human review should be considered failures of the predictive coding systems or mistakes by the human reviewers."[133]

[30] Furthermore, the study's findings do not support the idea that predictive coding is always better than manual review, and the study does not even find that predictive coding is better on the metric on which predictive coding should excel.[134] As seen in Table 2, Team B actually

[129] *See* Grossman & Cormack, *supra* note 18, at 17.

[130] *See id*.

[131] *See* Webber, *supra* note 87, at 6.

[132] *See id.* at 2.

[133] Dimm, *supra* note 19, at 8.

[134] *See* Grossman & Cormack, *supra* note 18, at 17–18.

achieved a slightly *higher* recall than either of the two predictive-coding teams.[135] The predictive-coding teams had better precision than either of the human reviewers, but recall is the more important factor for crafting a review that will be defensible in court.[136] Recall is more important because it is a measure of whether predictive coding is accurately categorizing documents as relevant.[137] As mentioned, the fact that one of the predictive-coding teams had access to the adjudicated results of the human teams' reviews casts serious doubt on the data in favor of that predictive-coding team; importantly, this "compromised" team used the adjudicated results to train the computer to make coding decisions.[138] In other words, the human-review teams were on-par or better than the predictive-coding teams on recall, and that remained so even despite the fact that one predictive-coding team used adjudicated results *from the human review* to build their coding algorithm.

[135] *See* Roitblat et al., *supra* note 18, at 76.

[136] *See id.;* Dimm, *supra* note 19, at 75 (noting precision is also important because it typically impacts the efficiency and cost of the review).

[137] *See Understanding Precision and Recall*, *supra* note 58.

[138] *See* Roitblat et al., *supra* note 18, at 74.

	Recall	**Precision**
Team A (human)	48.8%	19.7%
Team B (human)	53.9%	18.2%
Team C (Predictive Coding)	45.8%	27.1%
Team D (Predictive Coding)	52.7%	29.4%

Table 2: Recall and precision of human and predictive-coding review teams in the Roitblat, Kershaw, and Oot study, measured against the results of the original review.[139]

[31] Given the flaws of the Four Team study, its "proof" of predictive coding's superiority is uncertain at best. The study shows that predictive coding could produce results that are comparable to human review, but only under the right circumstances. Specifically, these circumstances would include built-in advantages for the predictive-coding teams, and when an unstable gold standard is used to measure performance.[140] Simply put, the study does not show that predictive coding is superior to human attorneys.

2. The TREC Data Study Does Not Show that Predictive Coding Is Superior to Manual Review

[32] Similarly, despite what proponents have said about the TREC Data study, it does not demonstrate that predictive coding holds an unequivocal advantage over human review.[141] Limitations in this study have led courts

[139] *See id.* at 76.

[140] *See id.* at 74, 77, 79.

[141] *See* Grossman & Cormack, *supra* note 18, at 44*;* see also Remus, *supra* note 22, at 1703 (explaining that "advocates of predictive coding—lawyers, judges, and vendors alike"—ignored the study's limitations "and began using the study to argue for widespread adoption" of predictive coding).

and others to misunderstand its conclusions. Chief among these limitations is that human review assisted the predictive-coding teams,[142] resulting in a misleading comparison between human review and predictive coding. Moreover, the TREC Data study suffers from other design flaws, including the skewed selection of the predictive-coding teams that were compared to human reviewers, a biased appeals process, the lack of experience of certain volunteers who performed the human review, and others.[143] As an initial matter, however, it is important to provide an overview of the study and the experiment on which it was based.

a. The TREC Legal Track Data

[33] The TREC Data study did not conduct original experimentation, and most of the limitations in the study stem from adapting the original TREC data. The study used data from the Interactive Task of the 2009 Text Retrieval Conference (TREC) Legal Track.[144] TREC's Legal Track is designed to "create test collections with enduring value, to report results against which future work can be compared, and to foster the development of a research community that is well prepared to continue that work."[145] Given those goals, using the TREC data to study predictive coding's

[142] See *infra* Section II (A)(2)(b).

[143] *See id.* To be fair, the authors of the TREC Study performed another study that reached similar results. *See, e.g.*, Gordon V. Cormack & Maura R. Grossman, *Navigating Imprecision in Relevance Assessments on the Road to Total Recall: Roger and Me*, SIGIR '17: PROC.'S OF THE 40TH INT'L ACM SIGIR CONF. ON RES. AND DEV. IN INFO. RETRIEVAL 5, 5–14 (2017), https://dl.acm.org/doi/pdf/10.1145/3077136.3080812 [https://perma.cc/7DPY-3J5A]. This more-recent study is not the focus of this article because it has not been used in the authors' experience by proponents of predictive coding to the same extent as the TREC Study.

[144] *See* Grossman & Cormack, *supra* note 18, at 2.

[145] Hedin et al., *supra* note 96.

capabilities is perfectly reasonable. Yet, although the 2009 Interactive Task involved the use of predictive coding, it was not designed to compare predictive coding to human review and did not attempt to conduct an exhaustive human review.[146]

[34] The TREC Legal Track is a conference that focuses on advancements in e-discovery.[147] In 2009, TREC designed its "Interactive Task" to evaluate and compare how effectively various predictive-coding tools conducted document review.[148] TREC invited predictive-coding teams to conduct a mock review of more than 800,000 documents collected by the Federal Energy Regulatory Commission during its investigation of the Enron Corporation.[149] Eleven predictive-coding teams participated, with each team completing between one and four "runs," or searches.[150] Each run was given one of seven topics, and the teams were instructed to code each document as relevant or non-relevant to the given topic.[151] Each team was allowed to consult with a designated TREC "Topic Authority" for up to ten hours "for purposes of clarifying the scope

[146] *See id.* at 1; *see also* Dimm, *supra* note 19, at 8–9.

[147] *See* Hedin et al., *supra* note 96, at 1.

[148] *See id.* at 1–2.

[149] *See id.* at 5 (explaining that the documents were 569,034 Enron emails and their 278,575 attachments).

[150] *See id.* at 6.

[151] *See id.* at 5.

and intent of a topic."[152] In total, the eleven predictive-coding teams submitted twenty-four unique runs across the seven topics.[153]

[35] Once the predictive-coding teams submitted their runs, TREC drew samples of the results and conducted a two-step evaluation: a "first-pass" review of the sample documents followed by an appeals process.[154] Human assessors, primarily *law student volunteers*, conducted the first-pass review of the samples, which amounted to nearly 50,000 documents over the seven topics.[155]

[36] TREC then gave the predictive-coding teams the results of the first-pass assessments and allowed them to appeal any assessments that conflicted with their predictive-coding results.[156] The predictive-coding teams submitted "documents detailing the grounds for each appeal they were submitting."[157] The first-pass reviewers did not participate in the appeals process at all.[158] The Topic Authorities then adjudicated the appeals for their designated topics, providing an authoritative determination of relevance for each of the appealed documents in the sample.[159] Using this two-step evaluation of the sample documents, the

[152] *Id.* at 3.

[153] *See* Hedin et al., *supra* note 96, at 7, Table 1.

[154] *See id.* at 8, 13.

[155] *See id*.

[156] *See id*. at 13.

[157] *See id*.

[158] *See id*.

[159] *See* Hedin et al., *supra* note 96, at 13.

first-pass review and the appeal, TREC estimated the recall and precision for each of the predictive-coding teams' runs.[160] The teams' results are reproduced in Table 3.

Topic	Run	Recall	Precision
201	UW	77.8%	91.2%
	CB	20.4%	69.0%
	CS	48.9%	21.5%
	UP	16.7%	11.7%
202	UW	67.3%	88.4%
	CS	57.9%	66.4%
203	UW	86.5%	69.2%
	ZL-NoCull	17.5%	89.5%
	UB	59.2%	11.1%
	ZL-Cull	2.9%	61.3%
204	H5	76.2%	84.4%
	CB	19.8%	16.9%
	AD	30.5%	7.7%
205	EQ	46.3%	91.5%
	CS	67.3%	32.1%
	IN	29.2%	25.1%
206	CB-High	7.6%	3.8%
	LO	4.2%	2.6%
	CB-Mid	1.1%	60.8%
	CB-Low	0.9%	61.2%
207	UW	76.1%	90.7%
	CB	76.8%	83.4%
	EQ	48.3%	72.5%
	LO	53.8%	18.3%
Average for All 24 Runs:		**41.4%**	**51.3%**

Table 3: Post-adjudication estimates of recall and precision for the Predictive-Coding Teams participating in the TREC 2009 Legal Track Interactive Task.[161]

[160] *See id*. at 17, Table 6.

[161] *See id.* at 6–7, 17 Table 6 (explaining that each of the eleven predictive-coding teams was given a two-letter team ID and completed 1–4 single-topic runs; for Topic 203, Team

b. The TREC Data Study

[37] The TREC Data study compared the results of some of these predictive-coding runs with the results of TREC's first-pass assessments, using the assessments as a rough substitute for exhaustive human review.[162] Since the first-pass assessment reviewed only the sample documents, the sample results were extrapolated to derive an estimate for the entire document population.[163] Not all eleven predictive-coding teams were evaluated, but instead the analysis was restricted to the two highest-performing teams, labeled UW and H5 in Table 4.[164] Accordingly, only the five topics reviewed by one of those two teams were considered in the TREC Data study.[165] The results of their analysis are reproduced in Table 4.

ZL submitted two runs—one with pre-culling by a document custodian, and one with no culling (i.e., coding the entire document set), and for Topic 206, Team CB submitted three runs, each with "a different level of effort (low, medium, high) in preparing the submission).

[162] *See* Grossman & Cormack, *supra* note 18, at 15.

[163] *See id.* at 16–17.

[164] *See id.* at 14–15.

[165] *See id.* at 36, 37 and accompanying Table.

Topic	Run/Assessment	Recall	Precision
201	UW	77.8%	91.2%
	First-pass (Law Students)	75.6%	5.0%
202	UW	67.3%	88.4%
	First-pass (Law Students)	79.9%	26.7%
203	UW	86.5%	69.2%
	First-pass (Professionals)	25.2%	12.5%
204	H5	76.2%	84.4%
	First-pass (Professionals)	36.9%	25.5%
207	UW	76.1%	90.7%
	First-pass (Professionals)	79.0%	89.0%
Average:	**H5/UW (all runs)**	**76.7%**	**84.7%**
	First-pass (all assessments)	**59.3%**	**31.7%**

Table 4: Post-adjudication estimates of recall and precision for UW and H5's runs from TREC 2009, compared with the extrapolation of the first-pass assessments.[166]

[38] A core issue with the TREC Data study is that manual human review artificially inflated the performance of the predictive-coding teams.[167] By conducting their own reviews after the predictive-coding systems produced results, humans on the predictive-coding teams corrected the machines' mistakes.[168] That is, humans themselves assisted the predictive-coding teams. Although combining manual review and predictive coding is an advisable approach to discovery in litigation, doing so in an experiment or study and labeling this hybrid approach as

[166] *Id.* at 37 and accompanying Table.

[167] *See* Dimm, *supra* note 19, at 116.

[168] *See id.*

"predictive coding only" can (and has) lead to mistaken and misunderstood conclusions.[169]

[39] The study therefore cannot be relied upon to show how predictive coding, by itself, compares against manual human review. A significant effect of human review on predictive coding results was to increase precision by removing non-relevant documents that the predictive model flagged incorrectly as relevant.[170] The study found that predictive coding held a statistically significant advantage over human review on just precision; no other value, including recall, exhibited statistically significant results.[171] Since the predictive-coding teams used manual review as well, "[i]t is impossible to determine how much of the precision is attributable to the performance of the [predictive-coding] algorithms, and how much is attributable to the humans that removed the software's bad predictions."[172] Perhaps the increase in precision was primarily due to "the [predictive-coding] systems making it easier for the human reviewers to produce good results by presenting documents in an order that made the review easier."[173] Perhaps, the humans on the predictive-coding teams were simply better at reviewing documents than the first-pass assessors.[174] Either way, the predictive-coding teams' precision ratios were seemingly inflated to an interminable extent via human review of the predictive-coding results.

[169] *See id.* at 117.

[170] *See id.*

[171] *See id.* at 116–17.

[172] Dimm, *supra* note 19, at 117.

[173] *Id.*

[174] *See id.*

[40] Nonetheless, from the comparisons of the predictive-coding runs to the human reviews, the TREC Data study observed that "the average efficiency and effectiveness of the five technology-assisted reviews surpasses that of the five manual reviews."[175] Specifically, the study found that the predictive-coding systems yielded higher precision than the first-pass assessments.[176] It noted that "[t]he measurements also suggest that the technology-assisted processes may yield better recall, but the statistical evidence is insufficiently strong to support a firm conclusion to this effect."[177] Finally, the study concluded that "[t]echnology-assisted review can (and does) yield more accurate results than exhaustive manual review, with much lower effort."[178]

[41] For a number of reasons, beyond the core issue discussed above, the idea that predictive coding always produces better results than human review is not supported. First, even when using the first-pass assessments as a rough approximation of an exhaustive human review, predictive coding does not produce consistently superior results.[179] The study found no statistically significant difference between the recall ratios of human reviewers and predictive coding for three of the topics.[180] The predictive-coding teams tended to have better precision ratios than the human

[175] *See* Grossman & Cormack, *supra* note 18, at 43.

[176] *See id.* at 43–44.

[177] *Id.* at 44.

[178] *Id.* at 48.

[179] *See id.* at 37, Table 7.

[180] *See id.*

assessors, but the first-pass review essentially tied one of the highest-performing predictive-coding team's precision for one of the topics.[181]

[42] Of course, only the top two predictive-coding teams from TREC 2009 were taken into account, while the results of the other nine were ignored.[182] Compared to the *average* performance of TREC's 24 predictive-coding teams' runs, the first-pass assessments still have worse precision (51.3% to 31.7%), but have better recall (59.3% to 41.4%).[183] Thus, "[i]t would be absolutely wrong to claim that [predictive coding] has been shown to beat human review at finding relevant documents in general, though [predictive coding] does achieve its results at much lower cost."[184]

[43] Notably, the study was not designed to be a fair fight between predictive coding and human review. That is because the top two predictive-coding teams were picked precisely "because they were considered most likely to demonstrate that technology-assisted review can improve upon exhaustive manual review."[185] The question the study sought to answer was simply "*whether* [predictive coding] *can* improve on manual review," not whether it always does.[186] Drawing conclusions from

[181] *See* Grossman & Cormack, *supra* note 18, at 37, Table 7.

[182] *See id.* at 48.

[183] *Id.* at 37.

[184] Dimm, *supra* note 19, at 9–10.

[185] Grossman & Cormack, *supra* note 18, at 48.

[186] *See id.*

the study about how predictive coding "always" performs is misguided.[187] This study looked at predictive coding at its best to see if it could beat human review at its worst.

[44] For example, the study relied upon TREC's first-pass assessments of sample documents in place of an actual manual review.[188] First, real document reviews typically involve more than one "pass" before production for at least a percentage of the documents at issue.[189] However, even if the first-pass processes were comparable to an actual review, it seems highly unlikely that volunteer law students are comparable to professional attorneys for obvious reasons, namely experience and the fact that this is volunteer work and not a real case.[190] Research in a number of fields has shown a significant qualitative difference between using student and professional subjects, casting serious doubt on the generalizability of student performance to attorneys at large.[191] Here, no adjustments were

[187] *See* Dimm, *supra* note 19, at 10 ("It would certainly be easier to justify the use of predictive coding if it could be proven to always produce results that are at least as good as exhaustive manual review. No such proof will ever exist.").

[188] *See* Grossman & Cormack, *supra* note 18, at 28–29.

[189] *See* Webber, *supra* note 87.

[190] *See* Dimm, *supra* note19, at 116.

[191] *See* Michael E. Gordon et al., *The "Science of the Sophomore" Revisited: From Conjecture to Empiricism*, 11 ACAD. MGMT. REV. 191, 192–93, 199–200 (1986) (examining thirty-two studies in which student and non-student subjects participated under identical experimental conditions and finding at least one qualitative difference in twenty-two of statistical studies, twelve of which indicated statistically significant differences); David O. Sears, *College Sophomores in the Laboratory: Influences of a Narrow Data Base on Social Psychology's View of Human Nature*, 51 J. PERSONALITY & SOC. PSYCHOL. 515 (1986) (finding that research based on student subjects is only somewhat applicable to the general public and that overdependence on students as subjects has distorted psychology's understanding of human behavior); Wayne Weiten & Shari Seidman Diamond, *A Critical Review of the Jury Simulation Paradigm*, 3 L. &

made to the first-pass assessors' results to more accurately reflect the performance of experienced lawyers.[192]

[45] In fact, research suggests that the students' assessments are not of the quality of professional attorneys'.[193] In an effort to determine what caused the wide variability in recall scores, the TREC Data study sampled a number of documents that the first-pass reviewers incorrectly coded as irrelevant.[194] It found that 65% of those documents were clearly relevant and that only 4% of the documents were of debatable relevance.[195] It seems likely, then, that the many of the errors were due to "review teams being especially careless, rather than it being inherently difficult . . . for humans to identify relevant documents."[196]

[46] Analyzing the TREC 2009 data for himself, information scientist William Webber concluded that the high variability of the first-pass

HUM. BEHAV. 71, 81–82 (1979) (noting that student subjects can be a poor substitute for real jurors and that research using students can produce results that are not generalizable to real cases); David F. Bean & Jill M. D'Aquila, *Accounting Students as Surrogates for Accounting Professionals When Studying Ethical Dilemmas: A Cautionary Note*, 7 TEACHING BUS. ETHICS 187 (2003) (finding a significant qualitative difference between the performance of students and experienced accounting practitioners, and questioning the external validity of research based on student subjects); Robert A. Peterson & Dwight R. Merunka, *Convenience Samples of College Students and Research Reproducibility*, 67 J. BUS. RES. 1035, 1040 (2014).

[192] *See* Hedin et al., *supra* note 96.

[193] *See* Webber, *supra* note 87.

[194] *See* Grossman & Cormack, *supra* note 18, at 37–38.

[195] *See id.* at 43.

[196] *See* Dimm, *supra* note 19, at 119.

assessments "suggest[s] a lack of the quality control and direction that might be expected in a true, professional review effort."[197] Webber recalculated the TREC Data study's comparisons, "exclud[ing] those reviewers whose proportion relevant are significantly different from the median, and re-apportion[ing] their work to the more reliable reviewers."[198] Webber found that this change generally improved the first-pass reviewers' results for every topic except one, suggesting that the review of that topic was conducted with the supervision and discipline typically expected of a professional document review.[199] Of course, the human reviewers performed well on that topic even under the TREC Data study's numbers—tying the scores of the predictive-coding team on both recall and precision.[200]

[47] Even assuming the first-pass assessments were comparable to a document review performed by the average contract attorney, the TREC Data study presents no reason to believe the results reflect what the best lawyers could do.[201] Some reviewers are more experienced, more attentive, more accurate, and more knowledgeable than others. Given that only the best predictive-coding teams were evaluated, a review conducted or at least supervised by an experienced attorney would have allowed for a more helpful comparison between predictive coding and manual review.[202]

[197] Webber, *supra* note 87, at 6.

[198] *Id.* at 6.

[199] *See id.* at 6–7.

[200] *See supra* Table 4

[201] *See* Ralph Losey, *Secrets of Search - Part II*, E-DISCOVERY TEAM: L. & TECH. (Dec. 18, 2011), https://e-discoveryteam.com/2011/12/18/secrets-of-search-part-ii/ [https://perma.cc/VQ4Z-JUGM].

[202] *See* Webber, *supra* note 87, at 8.

Instead, this assessment of the TREC Data merely shows that "a highly resourced and incentivized [predictive-coding] review can perform as well as or better than a poorly budgeted and un-incentivized human review."[203]

[48] Furthermore, the extrapolation of the sample documents likely skewed the numbers in favor of the predictive-coding teams. TREC conducted its sampling to provide the best indication of how the predictive-coding teams compared with each other, not to compare predictive coding with human review.[204] "Documents predicted to be relevant by [at least one predictive-coding team] were sampled heavily," while documents that no predictive-coding team flagged as responsive were lightly sampled.[205] As a result, TREC did not have a clear idea of "the total number of relevant documents" in the data set.[206] This means that when comparing predictive coding to human review, "the recall values have large uncertainty due to the uncertain total number of relevant documents in the denominator."[207] This uncertainty does not affect the comparison of predictive-coding teams to each other, "because they all have the same (uncertain) denominator."[208]

[203] Gerard J. Britton, *Does Uncritical Judicial Acceptance of Unsupported and Potentially Erroneous Technology-Assisted Review Assertions Frustrate the Objectives of Discovery?*, ASS'N OF CERTIFIED E-DISCOVERY SPECIALISTS (2014).

[204] *See* Dimm, *supra* note 199, at 119.

[205] *See id*. at 120.

[206] *See id*. at 119.

[207] *Id*.

[208] *Id*.

[49] Extrapolating the sample data to the entire population only amplifies the sampling problems.[209] Since the documents that no predictive-coding team found relevant were sampled lightly, each one was weighted significantly heavier in the extrapolation.[210] A successful appeal of one of these documents was worth up to 150 times as much as appealing other documents.[211] Because of this, "even a slight appeal-induced bias would greatly harm the apparent precision of the reviewers, and boost the recall of the teams."[212]

[50] TREC's appeal process contained some bias toward the predictive-coding teams.[213] The first-pass reviewers had no involvement in the process once their assessments were complete, but the predictive-coding teams were allowed to file written appeals with the "Topic Authorities."[214] In addition, "since appeals always came from the [predictive-coding] teams, the Topic Authorities knew whether adjudicating a document to be relevant would favor a [predictive-coding] team or a human reviewer, so the Topic Authorities could intentionally or subconsciously bias the results."[215] The rather high success rate of the appeals further suggests that

[209] *See* Webber, *supra* note 87, at 4.

[210] *See id*.

[211] *See id*.

[212] *Id*. Webber also calculated that comparing TAR teams to the first-pass assessments *without* extrapolating the sample data improves the human reviewers' performance significantly. For one topic, for example, the law students' precision improves from 5% to 41%. *See id*.

[213] *See* Dimm, *supra* note 19, at 121.

[214] *Id*.

[215] *Id.*

the predictive-coding teams were able to re-align the Topic Authorities' "conception[s] of relevance" to fit their own.[216] On average, the predictive-coding teams won 89.7% of their appeals.[217]

[51] Moreover, the two teams hand-picked for evaluation benefited from the biased appeals.[218] Although the eleven predictive-coding teams appealed an average of 31.8% of conflicts with the first-pass assessments, one of the best predictive-coding teams appealed more than 50% of its conflicts, and the other appealed 97.5% of the time.[219] Of course, the biased nature of the appeals procedure would not have mattered as much for TREC 2009's original purposes. The first-pass reviewers were not involved in the appeals process because they had no true investment in the process. By comparing the first-pass reviews to the predictive-coding teams without controlling for the biased appeals, the TREC Data study further shifted in favor of predictive coding.

[52] The reliance on the Enron emails also limits its usefulness.[220] Released to the public by FERC in 2003, the emails are relics of Enron's culture in its final days, and they are not necessarily representative of

[216] Webber, *supra* note 87, at 2.

[217] *See* Hedin et al., *supra* note 96, at 14 Table 4.

[218] *See* Webber, *supra* note 87, at 2–3, 3 Table 2.

[219] *See* Hedin et al., *supra* note 96, at 14 Table 4.

[220] *See* Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 WASH. L. REV. 579, 610–11 (2018); April Glaser, *Who Trained Your A.I.?*, SLATE (Oct. 24, 2017, 5:51 PM), http://www.slate.com/articles/technology/technology/2017/10/what_happens_when_the_data_used_to_train_a_i_is_biased_and_old.html [https://perma.cc/523E-FSWS].

contemporary corporate email.[221] In the early 2000s, email was still fairly new and housed discussions that have since migrated to text messaging or social media.[222] Notably, the emails were released to the public because they were good evidence of Enron's wrongdoing.[223] The data set is thus based on the conversations of criminals and fraudsters, and "researchers have used the Enron emails *specifically* to analyze gender bias and power dynamics."[224] Consequently, a review of the Enron emails does not necessarily reflect predictive coding's ability to handle the documents and data typically produced by a corporation today.

[53] Taking all of the study's limitations into consideration, it is difficult to conclude how predictive coding performs relative to humans. Even with all of the advantages discussed above, the predictive-coding teams, were still unable to consistently produce higher recall than the first-pass assessors.[225] Although the predictive-coding teams had higher precision ratios than the human reviewers, it remains unclear how much of that precision was actually achieved by the predictive-coding tools themselves. At best, given the right data set, a biased appeals process, and help from human reviewers, *some* predictive-coding tools could yield higher precision ratios than undertrained human reviewers, some of whom

[221] Nathan Heller, *What the Enron E-Mails Say About Us*, NEW YORKER (July 17, 2017), https://www.newyorker.com/magazine/2017/07/24/what-the-enron-e-mails-say-about-us [https://perma.cc/CX64-282F].

[222] *See id.*

[223] *See id.*

[224] Levendowski, *supra* note 220, at 611.

[225] *See* Grossman & Cormack, *supra* note 18, at Table 7.

were law students, who reviewed a cherry-picked sample of the same documents.[226]

B. New Research Reveals the Benefits of Manual Review & the Limits of Predictive Coding

[54] Through our own research, the authors of this article enter the narrative by challenging the prevailing thinking that has resulted from misunderstandings of the prior studies. One experiment shows that combining manual review and predictive coding can yield significant benefits.[227] In particular, incredibly high precision can be achieved, while minimizing the unnecessary production of non-responsive documents.[228] Another set of data indicate that predictive coding, unlike humans, cannot be a reliable tool for identifying the most important documents—those used at depositions and trial.[229] Furthermore, this research elucidates the

[226] *Id.* at Table 7, ¶ 52. Grossman and Cormack have since published further research on TAR's capabilities, but the majority of their recent work has focused on comparing various predictive-coding systems with each other. *See, e.g.*, Gordon V. Cormack & Maura R. Grossman, *Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery*, 2014 PROC. 37TH INT'L ACM SIGIR CONF. ON RES. & DEV. INFO. RETRIEVAL 153, https://dl.acm.org/doi/pdf/10.1145/2600428.2609601?download=true [https://perma.cc/Y9FN-VCZJ]; Maura R. Grossman & Gordon V. Cormack, *Continuous Active Learning for TAR*, PRAC. L., Apr./May 2016, at 32. At any rate, none of their other works has been nearly as influential as their 2011 study—one of the leading articles in the entire field of e-discovery. *See* Li, *supra* note 17.

[227] *See infra* Part II (B)(2)(a).

[228] *See id*.

[229] *See infra* Part II (B)(2)(b).

costs and benefits of manual review vs. predictive coding, thus presenting a more nuanced (and realistic) picture than the literature to date.[230]

[55] Arguments have been made that manual attorney review is not as accurate as a well-trained predictive model.[231] Anecdotally, the authors of this article know that such a statement is false. In our experience, predictive models are not perfect and typically result in false positives (lower precision) at any given recall rate and in certain circumstances these false positives require production too. The authors of this article also know, again anecdotally, that subject-matter experts can drastically influence quality by reducing the impact of a predictive model's false positives on the document review by establishing a consistent and thoughtful first-level review framework bolstered by a robust quality-control protocol. Resolving these false positives means documents that a model identified as likely responsive are instead determined by a human lawyer as not responsive. At first glance, this suggests that the human lawyer is "overturning" the decision of the predictive coding model. However, as discussed further below, the opposite is typically the case. Perhaps counterintuitively, by coding documents as not responsive, the human attorney is confirming the existence of non-responsive documents consistent with the model's estimates.

[56] The authors conducted an experiment to provide an *empirical* assessment of this anecdotal experience. It evaluates the impact that incorrectly overturned responsive documents have on a document review, as well as the impact of a manual review guided by subject-matter experts. The authors used a data set from a real legal matter (referred to as "Project A" here) for our experiment. The authors also analyzed data from Project A to test predictive coding's capability of finding key documents that make their way onto exhibit lists.

[230] *See infra* Part II (B)(2)(a)–(b).

[231] *See* Grossman & Cormack, *supra* note 225, ¶ 61.

1. Data Sets and Predictive-Coding Models

[57] Project A's data consisted of email, Microsoft Office documents, PDFs, and other types of business documents. The initial review population totaled more than 10 million documents. The Project A experiment first used predictive coding to identify the likely responsive document set that would meet 75% recall, and then it applied manual review to that likely responsive set to remove false positives (non-responsive documents) from production.

2. Experiment Procedures and Results

[58] To evaluate the impact that incorrectly overturned responsive documents have on a review, as well as the impact that manual review performed by subject-matter experts, can have on the results of a document review, the authors created a predictive model for the data set and a validation set to establish the recall and precision of the predictive model. Table 5 contains the predictive modeling data set statistics.

Document Class Distribution	Project A
Training - Responsive	2,341
Training - Not Responsive	7,018
Training - Total	**9,359**
Validation - Responsive	525
Validation - Not Responsive	1,071
Validation - Total	**1,596**

Table 5: Predictive Model Training and Validation Data

a. Project A Details

[59] Project A established the responsive review population using a 75% recall cutoff score. At this recall, the predictive model achieved 80.69% precision before the review started, meaning 19.31% of the review population was known to be not responsive. Subject-matter experts created

a review protocol for the first-level review team, and then those experts applied quality-control methods to implement a defensible review process and minimize manual coding errors. To test the quality of the review and confirm the number of overturn errors, a random sample was created from the review population above the cutoff score and had subject-matter experts blindly review the sample for responsiveness. Table 6 contains the results of that blind review.

	First Level Review	Subject Matter Expert: Responsive	Subject Matter Expert: Not Responsive
Responsive	1,384	N/A	N/A
Not Responsive	218	57	161
Total	1,602	N/A	N/A

Table 6: Project A: Results of Blind Subject Matter Expert Review

[60] Of the 1,602 documents sampled from the review population, first-level reviewers coded 1,384 documents as responsive and 218 documents not responsive. Subject-matter experts confirmed that 161 of these 218 documents were actually not responsive, confirming that the overall results of this review achieved 96.4% precision. The results of the manual review process combined with the predictive model improved the overall quality of review by 15.75%, meaning attorneys performing the manual review correctly identified 15.75% (or more than 700,000 documents) of the known non-responsive documents. Additionally, subject-matter experts coded 57 of the manually reviewed non-responsive documents in the sample as responsive. This results in a 3.56% not-responsive overturn rate (57/1602) and comes at a small cost of 2.67% recall.

[61] In sum, manual review guided by subject-matter experts and combined with predictive coding allows legal teams to achieve precision levels over 95% while minimizing the unnecessary production of non-responsive documents. Achieving such high precision while managing

client risk is nearly impossible *without combining* manual review with predictive coding.[232] Human attorneys performing the manual review also confirm the math of the predictive-coding model. In other words, they find the documents that the predictive-coding model itself anticipated it would not find, thereby enhancing the quality of the document review. This research thus rebuts the prevailing wisdom (held by many, including regulators) that predictive coding alone produces "better and more consistent [results] than a manual review."[233]

b. Project A Exhibit Document Data

[62] The second analysis looked at predictive coding's ability to narrow responsive documents down to key documents. Whether a document is responsive to a discovery request says little about the document's actual importance in a case. An older study using 2008 survey data showed that on average, almost 5 million pages of documents "were produced in discovery in major cases that went to trial but only 4,772 pages actually were marked" as exhibits.[234] This issue has only grown more pronounced in more recent times with the continuing growth of data. In complex matters, in the authors' experience, parties typically produce hundreds of

[232] See *How to Make the E-Discovery Process More Efficient with Predictive Coding*, THOMSON REUTERS LEGAL, https://legal.thomsonreuters.com/en/insights/articles/how-predictive-coding-makes-e-discovery-more-efficient [https://perma.cc/8HM2-3BE7].

[233] Tracy Greer, *Avoiding E-Discovery Accidents & Responding to Inevitable Emergencies: A Perspective from the Antitrust Division* 6, ABA SPRING MEETING (Mar. 2017), https://www.justice.gov/atr/page/file/953381/download [https://perma.cc/7RU6-QEVH] [hereinafter Greer, *Avoiding E-Discovery Accidents*].

[234] LAWYERS FOR CIVIL JUSTICE, ET AL, LITIGATION COST SURVEY OF MAJOR COMPANIES, STATEMENT FOR PRESENTATION TO COMMITTEE ON RULES OF PRACTICE AND PROCEDURE JUDICIAL CONFERENCE OF THE UNITED STATES 3 (2010), *available at* https://www.uscourts.gov/sites/default/files/litigation_cost_survey_of_major_companies_0.pdf [https://perma.cc/LN48-SDTA].

thousands or even millions of documents.[235] If the case goes to a jury, that large pool will be winnowed down to, at most, about 150 documents; even for a bench trial, no more than 300-400 documents will be used.[236] In any case, it is a small fraction of the technically responsive documents in discovery. Again, in the authors' experience, the number of deposition exhibits, though higher than the number of trial exhibits, is likewise small. That is the primary concern of this analysis.

[63] This analysis also used documents from Project A. The sample consisted of the deposition exhibits that had been scored by the predictive-coding model. These documents were scored on a scale of 0 to 100, which to generalize, measures how similar a document is to a responsive document in the sample population. It should be noted that a document with a score of 50 does not mean, however, that the document has a "50% chance" of being coded as responsive. The data below represent the predictive-coding model's scores from Project A.

[64] Figure 1 below shows that most documents in a collection will likely be non-responsive. The scores are displayed across ten "buckets" (0-10, 10-20, etc.), and Figure 1 represents how many documents fall into each bucket.

[235] Letter from David M Howard, Jonathan Palmer & Joe Banks to Hon. David G. Campbell (August 31, 2011), *in* ADVISORY COMM. ON CIVIL RULES, Agenda Book, at 411 (Nov. 7–8, 2011), *available at* https://www.uscourts.gov/sites/default/files/fr_import/CV2011-11.pdf [https://perma.cc/F6HV-NRXY] ("[F]or each one-page trial exhibit, Microsoft produces an average of 1000 pages, manually reviews more than 4500 pages, collects and processes more than 90,000 pages").

[236] *Id*. at 410–11.

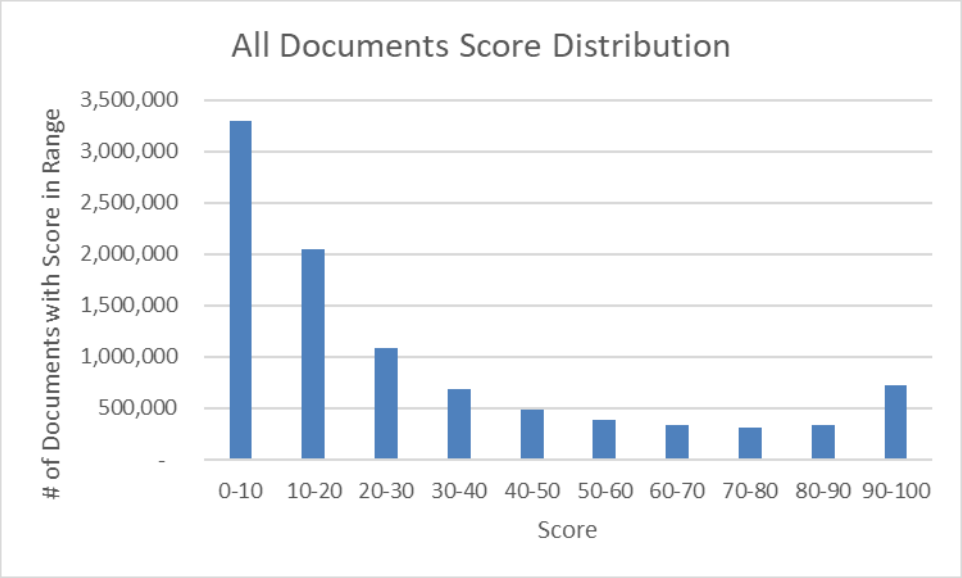

Figure 1.

[65] As this Figure illustrates, very few documents fall into the 90-100 bucket. Conventional wisdom suggests that the deposition and trial exhibits for a case would correspond with the documents scored mostly highly by the machine. It turns out, though, that this is not necessarily the case.

[66] To demonstrate, this article will now turn to the actual deposition and witness exhibits. The 1,015 exhibit documents have an average score of only 79.4, with a standard deviation of 27.87. The median score is 95.09, indicating half of the exhibit documents are with scores equal to or above 95.09.[237] The minimum score and maximum score are 1.31 and 100

[237] *See infra* Figure 2.

respectively.[238] Figure 2 also shows the distribution of scores over ten buckets.

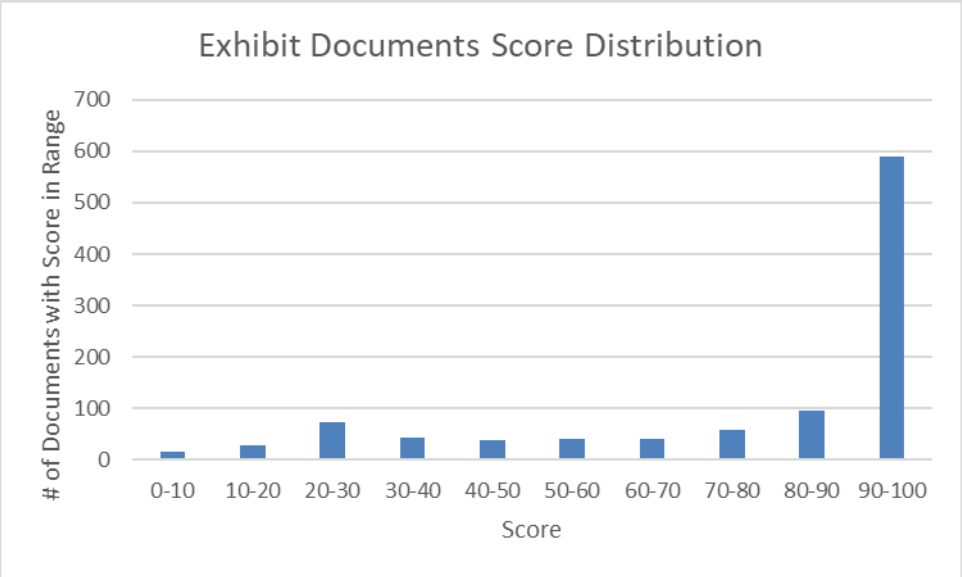

Figure 2.

[67] This data shows that predictive coding still struggles to reliably identify a substantial portion of the exhibit documents. Although 590 documents fall in the 90-100 range, about 42%, or 425 documents, fall below the 90-mark.[239] This means that predictive coding is not as reliable as one would expect when it comes to finding important documents, given predictive-coding models' ability to separate responsive from non-responsive documents. If predictive coding could do the work of locating the truly important documents, a job done quite well by actual attorneys,

[238] See *infra* Figure 2.

[239] *See supra* Figure 2.

one would expect far more documents in the 90-100 range for the predictive-coding model.

[68] The findings of this new research present an opportunity to outline a better picture of the pros and cons of predictive coding compared to human review. Predictive coding, to be sure, is valuable in certain circumstances: it efficiently identifies non-responsive documents, it can be more consistent, and it is an improvement on search terms (because predictive coding does not rely on "magic words"). Humans, meanwhile, score better on precision.[240] Most importantly, humans are better at identifying the documents that really matter: the few, but critical, documents used as evidence in actual proceedings. Predictive coding, as the foregoing data reveals, still has a long way to go in that regard. Furthermore, as this article's findings show, combining human review with predictive coding yields substantial improvements to the quality of document review, quality that predictive coding alone cannot achieve. In short, predictive coding has its benefits, but also its limits.

IV. PREDICTIVE CODING WITHOUT HUMAN REVIEW RISKS THE DISCLOSURE OF SENSITIVE AND CONFIDENTIAL INFORMATION

[69] In addition to the foregoing concerns, there are other problems with predictive coding without any corresponding human review, namely the increased risk of disclosing sensitive and confidential information. This risk comes in two forms: (A) inadvertent disclosures of confidential information that predictive coding nonetheless flags as relevant[241] and (B) compelled disclosures, *i.e.*, courts and regulatory agencies forcing the disclosure of certain information that producing parties would otherwise

[240] *See infra* Part III (A)–(B).

[241] *See* ARI KAPLAN, ADVICE FROM COUNSEL: CAN PREDICTIVE CODING DELIVER ON ITS PROMISE? 7 (FTI Consulting Tech. 2014), https://static2.ftitechnology.com/docs/white-papers/advice-from-counsel-predictive-coding-study.pdf [https://perma.cc/CXG7-RKRJ].

not produce because it is not responsive.[242] These disclosures also threaten core attorney-client privilege and work-product protections.[243]

[70] The result is that although predictive coding has the potential to make document review faster and cheaper, the risk of unwanted disclosures increase without human review to identify and fix such problematic disclosures. Especially when documents are produced to opposing counsel without any manual review, sensitive information can be released inadvertently. As sophisticated as today's technology is, "predictive coding is not magic and the software is not as smart as a human reviewer."[244]

A. Predictive Coding Without Human Review Increases the Risk of Inadvertent Disclosures

[71] The risk of accidently disclosing sensitive information is one of the principal dangers of using predictive coding. A large data set is bound to contain privileged information, trade secrets, personal health information, source code, and other confidential documents that predictive coding may nonetheless flag as relevant. Less "eyes-on" human review of a data set means fewer chances to find and remove sensitive information before producing documents to the other side. In a 2012 survey of top law firms and in-house counsel, 66% of respondents reported that the "risk of inadvertent productions would inhibit their use of predictive coding."[245] Some respondents even stated that because of this risk, "they probably

[242] *See* Remus, *supra* note 22, at 1716–17; Park & Reents, *supra* note 47.

[243] *See infra* Part III (A)–(B).

[244] Dimm, *supra* note 19, at 6.

[245] Kaplan, *supra* note 241, at 7.

would never use predictive coding for anything more than prioritizing documents for review."[246]

[72] When using predictive coding to respond to discovery requests, "[p]recision tends to decrease as recall increases," because finding "additional relevant documents means including documents that the [predictive model] predicted were less likely to be relevant."[247] This means that with a recall target as high as 75%, predictive coding will probably produce a number of false positives—documents incorrectly predicted as responsive.[248] A larger overall production of documents increases the risk that some of those documents, whether actually relevant or not, contain sensitive or harmful information. For example, a document incorrectly coded as responsive may reveal evidence exposing the producing party to liability unrelated to the matter at hand. Even a document that a predictive model correctly flags as relevant may contain privileged information that a human reviewer would have withheld from production.

[73] The negative consequences of disclosing privileged information are potentially devastating.[249] Even an inadvertent disclosure can waive privilege, either for the disclosed document alone or for all documents dealing with the same subject matter.[250] Some courts have found privilege

[246] *Id.*

[247] Dimm, *supra* note 19, at 84.

[248] *See id.* at 88.

[249] *See, e.g.*, Keeling et al., *supra* note 16, ¶ 6 ("Inadvertently disclosing privileged information can undermine a client's position and jeopardize her legal claim altogether.").

[250] *See* Manfred Gabriel et al., *The Challenge and Promise of Predictive Coding for Privilege* 2, ICAIL 2013 WORKSHOP ON STANDARDS FOR USING PREDICTIVE CODING, MACHINE LEARNING & OTHER ADVANCED SEARCH & REV. METHODS IN E-DISCOVERY

waivers even when "clawback" agreements were in place.[251] Federal Rule of Evidence 502 provides some measure of protection from the waiver of privilege in court proceedings, but the rule is less helpful in the context of government investigations and regulatory enforcement proceedings.[252] Moreover, privileged documents tend to include information that helps opposing counsel even when the documents are eventually destroyed or returned. For example, an inadvertently disclosed "document may lay out an attorney's assessment of the chances of success in litigation, weaknesses in the client's case, unfavorable case law, litigation strategy and the arguments to be relied on at various points in the litigation."[253]

(DESI V WORKSHOP) (June 14, 2013), http://www.umiacs.umd.edu/~oard/desi5/research/Gabriel-final2.pdf [https://perma.cc/XD9G-ZH4N].

[251] *See, e.g.*, Irth Sols., LLC v. Windstream Commc'ns LLC, No. 2:16-cv-219, 2017 WL 3276021, at *1, *16 (S.D. Ohio Aug. 2, 2017) (finding that the unintentional production of forty-three privileged documents constituted a waiver of privilege even when the parties had a clawback agreement in place that stated "[i]nadvertent production of privileged documents does not operate as a waiver of that privilege"), *appeal docketed*, No. 18-3740 (6th Cir. Aug. 7, 2018).

[252] *See* Elizabeth E. McGinn & Tihomir Yankov, *Guarding Against Privilege Waiver in Federal Investigations*, LAW 360 (Sept. 20, 2016, 12:46 PM), https://www.law360.com/articles/841643/guarding-against-privilege-waiver-in-federal-investigations [https://perma.cc/9XZX-DN32] (explaining that "textually, the thrust of Rule 502 governs the existence and reach of privilege waiver only in federal or state proceedings, to the exclusion of agency proceedings or investigations, even in cases where the privileged documents have been produced to a federal office or agency. Put differently, the rule may govern privilege waiver in cases where parties are subject to parallel (or sequential) federal investigation and civil litigation, but it does not address the scope of waiver—or the threshold question of whether there has been a waiver—with respect to the federal agency itself" (footnote omitted)).

[253] Gabriel et al., *supra* note 250, at 8; *see also* Keeling et al., *supra* note 16, ¶ 11.

B. Predictive Coding Without Human Review Increases the Risk of Compelled Disclosures

[74] The use of predictive coding without any corrective human review also results in forced disclosures of information that litigants would otherwise keep confidential. The issues at stake are attorney-client privilege and work-product doctrine protections. This forced disclosure occurs in the judicial setting, in which courts often encourage transparency and cooperation in discovery. Forced disclosure also occurs in the regulatory setting, in which agencies utilize their leverage in having the final say on the document-review process to gain significant control over parties' internal decisions about a document's relevance.

[75] Courts that approve of the use of predictive coding sometimes require (or at least strongly encourage) a heightened degree of "transparency" and "cooperation" in the discovery process.[254] In the landmark case *Da Silva Moore*, for example, the court required the defendants to disclose their seed-set documents (both relevant *and irrelevant* documents), coding decisions, and quality-control processes to the plaintiffs.[255] The court explained that "such transparency allows the opposing counsel (and the Court) to be more comfortable with computer-assisted review, reducing fears about the so-called 'black box' of the technology."[256] As more courts adopt the reasoning of *Da Silva Moore*, litigants hoping to use predictive coding may be required to disclose seed-

[254] *See, e.g.*, *Da Silva Moore*, 287 F.R.D., at 192; *See* Minutes of Status Conference re: Quality Assurance in Predictive Coding Regimen for City's Production, *Indep. Living Ctr. of S. Cal. v. City of Los Angeles*, No. 2:12-cv-00551 (C.D. Cal. June 26, 2014), ECF No. 375, at 2–3; *Kleen Prods. LLC v. Packaging Corp. of Am.*, No. 10 C 5711, 2012 WL 4498465, at *1, *8 (N.D. Ill. Sept. 28, 2012).

[255] *See Da Silva Moore*, 287 F.R.D., at 192.

[256] *Id.* at 192.

set documents and other potentially-sensitive data at Rule 16(b) conferences.[257]

[76] Similarly, parties seeking to use predictive coding in regulatory enforcement proceedings may be required to obtain agency approval of "specific methodological details, such as how the seed set is generated, how many training iterations are used, and what sampling is done to confirm the accuracy of the review."[258] Without a neutral judge to resolve disputes over the use of predictive coding, "the regulator has the final say on the way in which a document review is conducted."[259] Agencies can use this latitude to gain significant access to, and control over, parties' internal decisions about coding documents.[260]

[77] Although extensive cooperation with regulators may buy parties the right to use predictive coding, "this level of transparency, which is not typical in a linear review, comes with risks for producing parties."[261] One

[257] *See* FED. R. CIV. P. 16(b)(3)(B)(iii) (stating that the scheduling order can "provide for disclosure . . . of electronically stored information"); *see also* John M. Barkett, *More on the Ethics of E-Discovery: Predictive Coding and Other Forms of Computer-Assisted Review* 39–40, DUKE L. SCH. (2012), https://law.duke.edu/sites/default/files/centers/judicialstudies/TAR_conference/Panel_5-Original_Paper.pdf [https://perma.cc/5N5Q-KRHK] (explaining that "if a Rule 16(b) conference is held, it may result in questioning about e-discovery production protocols that might implicate the ethical duty of candor to the court," and that under the ABA Model Rules of Professional Conduct, "the duty of candor to a tribunal trumps the duty of confidentiality" (footnote omitted)).

[258] Park & Reents, *supra* note 47.

[259] *Id*. ("While it is possible for a party to seek court intervention with respect to a regulatory subpoena, that is highly unusual since the investigated party has strong incentives to accede to demands from the regulator in the interests of being perceived as cooperative.").

[260] *See id*.

[261] *Id*.

such risk is the "potential expansion of the regulator's investigation and document requests into new areas as a result of reviewing the non-responsive documents in the seed sets."[262] Compounding this risk is the fact that regulators may be eager to collect a great deal of documents from parties, both relevant and irrelevant. In a recent white paper, a senior attorney for the Department of Justice's Antitrust Division stated that "the Division will not agree to a party conducting essentially a second responsiveness review of the production during the privilege review process."[263] In other words, if attorneys using predictive coding "encounter obviously non-responsive documents in the course of a privilege review—such as employee's medical records or pictures of the employee's children—the division's policy suggests that the human attorneys will not be allowed to exercise their own independent judgment to mark such documents as non-responsive."[264] This is one example of the increasing (and unwarranted) tendency to elevate predictive coding over *any* human review, including by subject-matter experts, to provide a necessary check.

[78] In a sense, "requiring seed-set transparency threatens core protections for attorney work product, attorney-client privilege, and

[262] *Id*.

[263] Greer, *supra* note 233, at 6; *cf.* TRACY GREER, U.S. DEP'T OF JUSTICE, TECHNOLOGY-ASSISTED REVIEW AND OTHER DISCOVERY INITIATIVES AT THE ANTITRUST DIVISION 5 (Mar. 26, 2014), https://www.justice.gov/sites/default/files/atr/legacy/2014/03/27/304722.pdf [https://perma.cc/C3AB-SS5F] ("In some instances, the producing party preferred to use traditional manual review, rather than [predictive coding], for the documents of *the most senior executives. Although the Division did not object to this approach [for the most senior executives]*, judgments about responsiveness during manual review are less accurate and almost certainly are not consistent among reviewers." (emphasis added)).

[264] Robert Keeling, *Using Predictive Coding for HSR Second Requests*, 110 ANTITRUST & TRADE REG. REP. 716 (July 7, 2017).

confidentiality,"[265] because it forces litigants to reveal their decision-making processes to the other side. Disclosing internal reasoning about the selection of documents used to train the computer for relevance or privilege review could help opposing counsel figure out the best areas to probe in subsequent discovery requests.[266] Furthermore, even the "[n]on-privileged, non-responsive documents in a seed set could include information that reveals unethical or criminal activity by a party, embarrasses an officer or employee, or aids the requesting party in an unrelated cause of action."[267]

[79] Fortunately, a few courts have recognized some of the risks associated with predictive coding and have taken steps to mitigate them. In *Biomet*, a multi-district products liability case, the court rejected the plaintiffs' request for access to the irrelevant documents in the defendant's seed set.[268] The court noted that the plaintiffs' request reached "well beyond the scope of any permissible discovery by seeking irrelevant or privileged documents used to tell the algorithm what *not* to find."[269] Citing Rule 26 of the Federal Rules of Civil Procedure, the court concluded that it was "self-evident" that the plaintiffs had "no right to discover irrelevant or privileged documents."[270]

[265] Remus, *supra* note 22, at 1716.

[266] *See id*. at 1716, 1716 n.125.

[267] *Id*.

[268] *See In re* Biomet M2a Magnum Hip Implant Prods. Liab. Litig., No. 3:12-MD-2391, 2013 WL 6405156, at *1 (N.D. Ind. Aug. 21, 2013).

[269] *Id.* (emphasis added).

[270] *Id.* at *1–2.

[80] Cases like *Biomet* indicate that some litigants are not above exploiting their opponents' use of predictive coding in order to get their hands on as many documents as possible. Although the "fishing expedition" approach to discovery is nothing new, attorneys should be aware that the use of predictive coding can make it easier for opposing counsel to access sensitive documents. This is especially true when using predictive coding without any subsequent human review.[271] A growing number of plaintiffs have tried to require defendants to produce all documents flagged as responsive by the predictive model,[272] sometimes even trying to prevent the defendants from conducting a manual privilege review.[273] Litigants should keep in mind that utilizing some form of human review, whether instead of or in addition to predictive coding, can help to greatly reduce the risk of unwanted disclosures.[274]

V. CONCLUSION

[81] With a better understanding of predictive coding's capabilities, limitations, and drawbacks, the rise of the "robot overlords"[275] seems less threatening. Manual document review is not obsolete, and those claiming otherwise are overstating what has been proven about predictive coding's performance. Although in the right circumstances predictive coding can

[271] *See* Dimm, *supra* note 19, at 4.

[272] *See, e.g.*, Chen-Oster v. Goldman, Sachs & Co., No. 10 Civ. 6950, 2014 WL 716521, at *1 (S.D.N.Y Feb. 18, 2014).

[273] *See, e.g.*, Good v. Am. Water Works Co., No. 2:14-01374, 2014 WL 5486827, at *1–2 (S.D. W. Va. Oct. 29, 2014).

[274] *See* Park & Reents, *supra* note 47; Kaplan *supra* note 241, at 7.

[275] *See* Robert Ambrogi, *Fear Not, Lawyers, AI is Not Your Enemy*, ABOVE THE L. (Oct. 30, 2017, 3:00 PM), https://abovethelaw.com/2017/10/fear-not-lawyers-ai-is-not-your-enemy/ [https://perma.cc/2GT2-98TC].

improve on human review alone, predictive coding is still not advanced enough to replace manual review. Instead, human attorneys will continue to perform document reviews, sometimes aided by predictive coding and sometimes not.[276] Additionally, subject-matter experts certainly still have a critical role to play. Even if predictive coding is gaining ground in the document-review space, however, it still lags where it matters most: reliably identifying the most important documents in a case that will be used at depositions and trial. Humans will continue to handle that task. The future looks more like a co-existence of humans and machines, not complete replacement of the former with the latter.

[82] It is critical that courts and litigants should be mindful of the limitations and flaws in prior studies about the effectiveness of predictive coding, as well as the risks posed by predictive coding. As this article shows, predictive coding's performance relative to human review has been artificially inflated due to defects in the designs of prior research, as well as a misunderstanding of the goals and findings of the research. These concerns are only heightened by the fact that predictive coding risks infringing on the core protections of the attorney-client privilege and work-product doctrine, as well as certain protections for producing parties in the Federal Rules of Civil Procedure. This article's research, moreover, reveals that manual review, after the application of predictive coding, can significantly increase the quality of a document review and production. Further, and equally important, it confirms the math of predictive-coding models by finding the documents that the models themselves recognize they probably incorrectly identified. Parties themselves should also be mindful of these pitfalls and recognize how and in what circumstances predictive coding can be useful and when the risks involve counsel strongly against document review completely without manual attorney review.

[276] *See* John Markoff, *The End of Lawyers? Not So Fast.*, N.Y. TIMES: BITS (Jan. 4, 2016, 12:33 AM), https://bits.blogs.nytimes.com/2016/01/04/the-end-of-work-not-so-fast/ [https://perma.cc/2SAR-C4HU].