Richmond Journal of Law & Technology

Volume 27 | Issue 4 Article 3

8-30-2021

Humans Against the Machines: Still Reaffirming the Superiority of Human Attorneys in Legal Document Review

Robert Keeling
Sidley Austin, LLP

Rishi Chhatwal
AT&T Services, Inc.

Peter Gronvall
Ankura Consulitng Group

Nathaniel Huber-Fliflet
Ankura Consulting Group

Richmond Journal of Law & Technology Volume XXVII, Issue 4

HUMANS AGAINST THE MACHINES: *STILL* REAFFIRMING THE SUPERIORITY OF HUMAN ATTORNEYS IN LEGAL DOCUMENT REVIEW

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

Cite as: Robert Keeling et al., *Humans Against the Machines: Still Reaffirming the Superiority of Human Attorneys in Legal Document Review*, 27 RICH. J.L. & TECH., no. 4, (2021).

[*] 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 the founder and co-chair of Sidley's eDiscovery and Data Analytics Team.

[**] 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.

I. INTRODUCTION

[1] In September 2020, the Richmond Journal of Law & Technology published an article by the undersigned authors,[1] in which the authors examined the limitations and risks of relying solely on predictive coding in the discovery process and demonstrated, with empirical data, that human attorney review significantly increases the quality of a document production. In response to this publication, Maura R. Grossman and Gordon V. Cormack have submitted a comment, in which they claim the article contains a "material error" that "calls into question its results and conclusions."[2] This claim is simply not true, on either count.

[2] The purpose of our research was to challenge the increasingly pervasive notion that predictive coding—*standing alone*—is superior to and should altogether replace human attorney review. The primary focus of our research compared the precision[3] of predictive coding alone to predictive coding followed by human review, and concluded that higher precision was achieved when predictive coding and manual review were combined.[4] The authors acknowledge that the article contained an erroneous data point and thank Grossman and Cormack for identifying this issue. As explained more fully below, however, even after the small adjustment incorporating our

[1] 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.

[2] Maura R. Grossman and Gordon V. Cormack, *'Reaffirming the Superiority of Human Attorneys in Legal Document Review and Examining the Limitations of Algorithmic Approaches to Discovery': Not So Fast* (Aug. 2021).

[3] Precision measures the portion of documents predicted to be relevant that actually are relevant. *See* Keeling et al., *supra* note 1, at 14–16 (explaining recall and precision).

[4] *Id.* at 48 (the object of the study was to "evaluate[] 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"); *id.* at 50–51 (concluding that manual review combined with predictive coding allows legal teams to achieve extremely high precision levels).

corrected data point, our conclusions remain the same and they remain overwhelmingly supported. Moreover, the data adjustment does not change our observations or conclusions about the dangers of exclusive reliance on machines to 'get it right.' Nor does it change the fact that use of predictive coding alone underestimates the importance of attorney review to protect clients' sensitive and confidential information from unnecessary disclosure.

[3] The authors welcome interest in our research and hope to engage in more peer-to-peer dialogue about these important issues in the future. We view this kind of thoughtful exchange as one of the benefits of academic research, and we believe it is a productive way to bring advancement to the eDiscovery field. Our central goal is to encourage transparency about the effectiveness and limitations of advanced analytics in the legal field and to ensure its appropriate implementation.

II. THE FOUR TEAM AND TREC DATA STUDIES

[4] At the outset, the comment makes a passing reference to the article's "erroneous" examination of the "flaws" and "misunderstandings" of two prior studies related to predictive coding (one of which is a study the comment's authors themselves conducted), but states that the comment will not address these points. The two studies the authors analyzed in the original article were the Four Team Study[5] and the TREC Data Study.[6] Broadly speaking, the article explained that courts and others in the legal field have mistakenly extended the studies' conclusions beyond what the data actually proves.[7] Specifically with respect to the TREC Data Study, the article explained that courts and legal professionals advocating for *exclusive* use of

[5] 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").

[6] Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in EDiscovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, 17 RICH. J.L. & TECH., no. 3, 2011, at 11 (hereinafter the "TREC Data Study").

[7] Keeling et al., *supra* note 1, at 24–47.

predictive coding have pointed to the study as evidence that "predictive coding holds an unequivocal advantage over human review"—even though the study does not support that conclusion.[8]

[5] The authors maintain that many in the legal community continue to misunderstand—and thus misapply—the TREC Data Study. Namely, because human review assisted the predictive coding teams, it is incorrect to conclude, as many have, that the TREC Data Study proves that "predictive coding is always superior to manual review."[9] Nor would it be correct to say, as some have, that the Study proves it is neither necessary nor beneficial to combine predictive coding with subsequent human review.

[6] The authors would like to make clear that any criticism of the conclusions others have drawn from the TREC Data Study was not intended to de-value the important research conducted by Grossman and Cormack. In fact, the article recognizes that the Four Team Study and TREC Data Study were instrumental to advancing the dialogue and driving the acceptance of predictive coding in the discovery industry.[10] More generally, we appreciate the positive effect this scholarship has had on the legal community.[11] Grossman and Cormack have made significant contributions

[8] *Id.* at 30–31 (*others'* misunderstandings of the TREC Data Study resulted in a "misleading comparison [*by others*] between human review and predictive coding"); *see also id.* at 21 ("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.") (emphasis added); *id.* at 22 ("*proponents of predictive coding* have often exaggerated the results of the predictive-coding studies") (emphasis added).

[9] *Id.* at 23 (explaining that "[n]either 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"); *id.* at 36–37 ("A core issue with the TREC Data study is that manual human review artificially inflated the performance of the predictive-coding teams.").

[10] *Id.* at 18 ("These studies significantly increased the legal profession's confidence in the use of predictive coding.").

[11] *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.

that have advanced the eDiscovery process and resulted in time and cost efficiencies for courts, attorneys, and parties.[12] Indeed, as the article stated, the authors use predictive coding regularly in their practices, and have seen firsthand that "predictive coding offers significant benefits for our clients."[13]

III. OUR ADJUSTED DATA STILL SUPPORTS THE AUTHORS' CONCLUSIONS

[7] Grossman and Cormack's comment focuses its criticism on the validity of the new research presented in the article. From the new research that was presented, the authors reached three conclusions: (1) "**combining** manual review and predictive coding can yield significant benefits," (2)

OF THE 40TH INT'L ACM SIGIR CONF. ON RES. AND DEV. IN INFO. RETRIEVAL, 5 5–14 (2017), https://dl.acm.org/doi/10.1145/3077136.3080812 [https://perma.cc/YQ8D-LTNB]; Gordon V. Cormack & Maura R. Grossman, *Engineering quality and reliability in technology-assisted review*, SIGIR '16: PROC. OF THE 39TH INT'L ACM SIGIR CONF. ON RES. AND DEV. IN INFO. RETRIEVAL, 75–84 (2016), https://dl.acm.org/doi/10.1145/2911451.2911510 [https://perma.cc/4UA6-PWBS]; 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; 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).

[12] *See, e.g.*, *Sinclair Wyoming Refining Co. v. A&B Builders Ltd.*, No. 15-CV-91, 2016 WL 11494744, at *8 (D. Wyo. Nov. 24, 2016) (approving use of predictive coding in eDiscovery and citing TREC Data Study); *Aurora Coop. Elevator Co. v. Aventine Renewable Energy-Aurora West, LLC*, No. 412-CV-230, 2015 WL 10550240, at *1 (D. Neb. Jan. 6, 2015); *Da Silva Moore v. Publicis Groupe & MSL Group*, 287 F.R.D. 182, 190 (S.D.N.Y. 2012); *Technology Assisted Review: The Judicial Pioneers*, 15 SEDONA CONF. J. 35 (2014) (recognizing the substantial contributions of Grossman and Cormack to the advancement of technology-assisted review in discovery).

[13] Keeling et al., *supra* note 1, at 10; *see also* Keeling et al., *supra* note 1 at 13 ("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.").

"predictive coding, unlike humans, cannot be a reliable tool for identifying the most important documents—those used at depositions and trial," and (3) use of predictive coding alone comes with an "increased risk of disclosing sensitive and confidential information."[14]

[8] Specifically, as to the empirical data supporting the authors' first conclusion, the comment states that the data required to correctly calculate the precision and recall of the human review and the predictive model has been omitted from the article. It then claims that when the precision and recall estimates are calculated correctly, the research "supports the unremarkable conclusion that post-predictive-coding human review trades recall for precision."[15] We address these points in turn.

[9] The original article stated that human review **combined** with predictive coding improved precision by 15.75% at a small cost of 2.67% recall. These figures were inadvertently miscalculated. In the text and table below, the authors provide additional information to increase transparency into our research and correct this calculation. Significantly, however, when the corrected figures are applied, ***our initial conclusions still hold true***.

[10] The following points provide clarification about our research process and, more specifically, about how the random sample data was used to perform our calculations:

> (1) We created a random sample of data to assess the impact that incorrectly overturned responsive documents have on a review. This sample was generated from documents that contained a predictive coding score, were **above** the 75% recall cut-off score, and had been reviewed for responsiveness by an attorney. For purposes of this response, we will refer to this data as Sample 1.

[14] *Id.* at 47–48, 55–56. Because Grossman and Cormack do not dispute the authors' second and third conclusions, we do not re-address those conclusions here.

[15] Grossman & Cormack, *supra* note 2.

(2) Earlier in the review process, a different random sample was used (the control set) to assess the performance of the predictive model and to establish the review population. The control set was drawn from the population of documents that met the required standard for a predictive coding workflow. For purposes of this response, we will refer to this data as Sample 2.

(3) Sample 1 provided an additional estimate of the recall and precision of both the model and the human review performed on the model's resulting document review population, in an effort to confirm whether human review can improve the precision of a predictive coding process while minimizing its impact on recall. This sample was used only for the purpose of answering this research question.

(4) Sample 1 was generated from documents that contained a predictive coding score, were above the 75% recall cut-off score, and were reviewed by attorneys. Sample 1, therefore, does not provide a measurement of the recall of the complete production population (the production the requesting party actually received) because the production contained documents with scores **below** the 75% recall cut-off score and/or that were not reviewed by attorneys. Inherently, then, some of the produced documents could not be part of Sample 1, which was pulled only from the population of documents **above** the 75% recall cut-off score and that were reviewed by attorneys.

(5) A null set sample was used to estimate the recall of the complete production population. The null set sample is a statistical sample that is pulled from the documents that will be excluded from production to confirm that the responsive rate within the null set is not higher than expected. In other words, the null set verifies that documents with predictive coding responsiveness scores below the cut-off score do not

> include an unexpectedly high proportion of responsive documents. For purposes of this response, we will refer to this data as Sample 3.

Table 1 contains the additional review coding numbers for the blind Subject-Matter Expert ("SME") review applied to the random sample that was created from the review population above the 75% recall cut-off score (Sample 1).

	First Level Review	**SME Responsive**	**SME Non-Responsive**
Responsive	1,384	1,190	194
Non-responsive	218	57	161
Total	1,602	1,247	355

Table 1: Additional Results of Blind Subject-Matter Expert Review

[11] Of the 1,602 documents in Sample 1, first-level reviewers coded 1,384 documents as responsive and 218 documents as non-responsive. Subject-matter experts confirmed that 1,190 of the 1,384 responsive documents were correctly coded responsive by the attorneys, confirming that the human review judgements applied to the review population achieved 85.98% precision—in other words, the human reviewers correctly coded responsive documents nearly 86% of the time. Among the 1,602 documents in Sample 1, 1,247 are true responsive documents, which makes the precision of the predictive coding model 77.84% (1,247 / 1,602 = 77.84%). For our calculations, we assumed the recall of Sample 1 is 75% since it was selected from the population of documents with scores above the 75% recall cutoff score. The actual human recall estimated from Sample 1 was 71.57% (0.75 * 1,190 / 1,247 = 71.57%).

[12] The context from which we derived these statistics is important and demonstrates the real-life, positive impact human review can have on a production. Project A was not a small, limited review. Instead, the production was massive, even by modern standards. Millions of documents were ultimately produced in Project A. Accordingly, even incremental improvements to the production create significant impact on the quality of

the review and substantial benefits to the client. Specifically, the use of predictive coding followed by human review improved the overall precision by 8.14%, meaning attorneys performing the manual review correctly identified (and removed) hundreds of thousands of non-responsive documents from the production. Without this manual human review, 22.16% of the documents that would have been produced by the predictive coding process would have been non-responsive. Additionally, subject-matter experts coded 57 of the manually reviewed non-responsive documents in the sample as responsive. Thus, this improved precision result comes at a small cost of 3.43% recall.

[13] To bring additional perspective to the results of the research, we also calculated the F1-score for each review process, predictive coding alone and predictive coding combined with human review, using the Sample 1 data. The F1-score is the harmonic mean of the precision and recall and can be used to evaluate the overall performance of a predictive model—not just precision or recall. This measure provides further clarity on the impact the lost recall has on the overall results. The F1-score for predictive coding alone was 0.764, while the F1-score for the predictive coding/human combined process was 0.782. The combined process's F1-score was 0.018 better than predictive coding alone, which means that even with a small reduction in recall, the overall performance of the human review was better than predictive coding alone.

[14] Precision improvements help manage the risk of producing non-responsive documents, which improves the overall quality of a document production and is critically important to our clients. Going forward, the authors believe that precision will become more important for judging the effectiveness of predictive coding. The authors acknowledge, however, that some receiving parties disfavor increased precision at the cost of achieving the desired recall rate. With this in mind, we also measured Project A's final recall rate at the time of production – the recall of the complete production that was received by the requesting party.

[15] In this real-world matter, the Project A document review process, which included predictive coding, was designed to achieve at least 75%

recall. Sample 1 data, which was used to derive our performance metrics, was sampled from documents *above* the cutoff score and did not include any documents with scores below the cutoff score that had a family relationship[16] with a document above the cutoff score. However, the matter required production of any document that was not privileged, had a score below the cutoff score, and was a family member to a document above the cutoff score.

[16] What this means is that, while the estimated human recall of Sample 1 was 71.57%, the overall process for production, which combined predictive coding with human review, achieved much higher recall. The estimated recall of the complete production in this real legal matter was 92.7%. We attribute this recall improvement to the production of responsive family members that were below the cutoff score – documents that could not be measured using Sample 1 data – and also to the variations between the actual document distribution and the estimated distribution that was inferred using the control set and the null set.

[17] We confirmed the estimated recall of the complete production by examining the estimated number of responsive documents in Sample 3 (the null set) and comparing it to Sample 2's (the control set) estimated number of responsive documents in the predictive coding workflow population, as well as the number of documents in the final production population.

[18] Using Sample 2 (the control set), we estimated that there were 3,448,540 responsive documents in the document review population. Our review protocol required that we identify at least 75% of these documents to achieve 75% recall. Specifically, we estimated that we needed to identify at least 2,586,405 documents and could only exclude 862,135 documents from production (or 25% of the responsive documents).

[19] Using Sample 3 (the null set), the sample that was reviewed by attorneys to confirm how many responsive documents were excluded from

[16] A family member relationship occurs when an email contains an attachment, for example.

the review process, we estimated that there were 251,551 responsive documents in this population. According to Sample 2's (the control set) responsive population estimation, our review process captured 610,584 more responsive documents than required to achieve at least 75% recall (862,135 - 251,551 = 610,584).

[20] As a result, our estimated recall for the complete production that was received by the requesting party that used a predictive coding workflow **combined** with human review was: 92.7% or 3,196,989 out of 3,448,540 responsive documents.

[21] Project A's ultimate production size was 3,995,566 documents. In addition to identifying an estimated 3,196,989 (2,586,405 + 610,584) responsive documents for production using the combined predictive coding and human workflow, there were other documents produced over the course of the project. These included documents that could not go through the predictive coding process because they did not have text, had too little text, or were reviewed outside the predictive coding workflow for other reasons.

[22] As the above analysis demonstrates, this adjusted and additional data confirms and strengthens our original conclusion: human review, particularly when performed by or supervised by subject-matter experts, can significantly increase the precision of a production above and beyond that achieved by predictive coding alone, and can significantly improve the overall production quality. Human review reduces the intrinsic risk of the machine's error by withholding from production non-responsive documents that the machine labeled responsive. Further, the small reduction in recall identified in our experiment does not take into account the fact that, in a real-life matter, the true recall rate of the final production population will be higher due to the inclusion of additional responsive family members.

[23] Thus, contrary to Grossman and Cormack's characterization of our conclusion as "unremarkable," we believe that the ability to identify and withhold hundreds of thousands of non-responsive and potentially sensitive and confidential client documents represents a significant and noteworthy benefit of human attorney review, particularly when we are simultaneously

able to achieve this without any material loss to recall. To be clear, the authors never purported to prove the absolute "superiority of human review," as Grossman and Cormack claim. Rather, the authors have consistently maintained—and the data overwhelmingly supports—the conclusion that there are significant limitations and risks to using predictive coding alone, and that human attorney review can significantly improve the quality of a document production.[17] These conclusions remain true.

* * *

[24] The authors appreciate the opportunity to engage in a robust dialogue with other knowledgeable experts on this important topic and look forward to continuing the discussion to advance the use of emerging technologies in the eDiscovery process. We maintain that the adjusted and additional data presented here supports the original article's conclusion—that predictive coding alone does not provide the same benefits and quality that can be achieved by combining predictive coding with human review. Moreover, we disagree that our research and conclusions are "unremarkable," and believe they provide much needed clarity and perspective about the inherent limitations and risks of relying solely on machines to get it right. Undoubtedly, technological advancement will continue to march on. But as the empirical evidence confirms, human attorney review also has an essential role to play and adds significant value to the discovery process.

[17] Keeling et al., *supra* note 1, at 7 (cautioning that "the risks of *completely removing humans* from document-review projects are significant") (emphasis added); *see* Keeling et al., *supra* note 1, at 11 (explaining that the authors' data "challenges the idea that manual review hinders the quality of a document review *that incorporates the use of predictive coding*") (emphasis added); Keeling et al., *supra* note 1, at 64 ("The future looks more like a co-existence of humans and machines, not complete replacement of the former with the latter."); Keeling et al., *supra* note 1 ("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.") (emphasis added); Keeling et al., *supra* note 1 (advising of the risks of "document review *completely without* manual attorney review") (emphasis added).