Bloomberg BNA

The United States
Law Week

Case Alert & Legal News™

Use of Predictive Coding in Regulatory Enforcement Proceedings

BY JENNIFER KENNEDY PARK AND SCOTT REENTS

In February of this year, Magistrate Judge Andrew Peck issued a landmark ruling in *Da Silva Moore v. Publicis Groupe*,[1] which provided the first judicial approval of the use of computer-assisted review for identifying relevant electronically stored information for production. In that opinion, which was upheld by District Court Judge Andrew Carter Jr.,[2] Judge Peck found that computer-assisted review can significantly reduce the costs of e-discovery and that it could be used in civil litigations consistent with the parties' obligations under the Federal Rules of Civil Procedure. Since February, much has been written about the decision's impact and the impact of computer-assisted review generally on civil litigation. This article analyzes *Da Silva Moore* and computer-assisted review in the context of regulatory enforcement proceedings, such as investigations or proceedings brought by the Securities and Exchange Commission, Department of Justice, Financial Industry Regulatory Authority, Commodities Futures Trading Commission, and various state authorities, including state attorneys general. These proceedings differ from civil litigation in important ways, including the frequently more expedited timelines for production, the absence of a neutral arbitrator for the initial phases of the proceedings, and less protective legal limits on the scope of review. This article assesses how these differences should weigh on decisions about whether, when, and how to use computer-assisted review in enforcement proceedings.

[1] No. 11 Civ. 1279 (ALC) (AJP), 2012 BL 44145 (S.D.N.Y. Feb. 24, 2012).

[2] *Da Silva Moore v. Publicis Groupe*, No. 11 Civ. 1279 (ALC) (AJP), 2012 BL 10197 (S.D.N.Y. Apr. 26, 2012).

Jennifer Kennedy Park is a partner based in the New York office of Cleary Gottlieb Steen & Hamilton LLP, where she focuses on white collar defense and corporate investigations, as well as litigation, particularly related to capital markets transactions.

Scott Reents is the ediscovery attorney in the New York office of Cleary Gottlieb Steen & Hamilton LLP, where he advises clients and the firm on electronic discovery law, technology, and best practices.

The authors would like to thank Peter H. Fielding for his invaluable assistance in preparing this article.

Da Silva Moore and Predictive Coding

Judge Peck's order in *Da Silva Moore* was the first judicial opinion to ''recognize[] that computer-assisted review is an acceptable way to search for relevant ESI in appropriate cases.''[3] In *Da Silva Moore*, Judge Peck approved a protocol for the use of computer-assisted review—also known as predictive coding or predictive review—for a discovery request that involved approximately three million electronic documents.[4] There are different variations of predictive coding, but it is generally a multi-step process in which human reviewers train the computer to distinguish relevant from irrelevant documents by providing the computer with examples of each type of document. Reviewers start by coding a ''seed'' set of documents, which the computer uses to make an initial prediction as to which unreviewed documents are relevant and which are irrelevant. Reviewers code small samples of these unreviewed documents in subsequent rounds of training to help the computer refine its predictions, until the com-

[3] *Da Silva Moore*, 2012 BL 44145, slip op. at *2.

[4] *Id.*, slip op. at *5. Although Judge Peck approved the protocol in February, at the time this article was written, the protocol had not yet been implemented, pending resolution of various disputes and with the entry of a stay ''pending Judge Carter's decision on plaintiffs' motions for collective action certification and to amend their complaint.'' *Da Silva Moore v. Publicis Groupe*, No. 11 Civ. 1279 (ALC) (AJP) (S.D.N.Y. May 14, 2012). Judge Carter granted plaintiffs' motions on June 29, 2012. *Da Silva Moore v. Publicis Groupe*, No. 11 Civ. 1279 (ALC) (AJP), 2012 BL 178774 (S.D.N.Y. June 29, 2012).

 ISSN 0148-8139

puter's coding accuracy reaches a desired level. At this point, 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.

The protocol Judge Peck approved generally followed the outline above, but it also included provisions to increase transparency and encourage cooperation between the parties. Under the terms of the protocol, counsel for the producing party—the defendant—is required to share with plaintiffs' counsel all nonprivileged documents (whether responsive or not) and the coding from the initial seed set of documents and the seven subsequent rounds of review (of 500 documents each) that will be used to train the computer.[5] Plaintiffs' counsel is permitted to review these documents and provide the defendant with its own evaluation of the coding.[6] The parties are expected to attempt to resolve any disputes related to the coding of documents;[7] however, Judge Peck is available to resolve any intractable conflicts.[8] Documents identified as responsive by the computer will be subject to further review by defendants prior to production,[9] but documents identified as not responsive by the computer will receive no human review and will not be produced.

As Judge Peck recognized, predictive coding can be a valuable tool in civil litigation. First, in terms of production accuracy, predictive coding ''works better than most of the alternatives, if not all of the [present] alternatives.''[10] Indeed, predictive coding '' 'can (and does) yield more accurate results than exhaustive manual review, with much lower effort.' ''[11] Predictive coding has also been shown to be more accurate than keyword searches.[12] Related to its improved accuracy, predictive coding has the potential to reduce the costs of complying with discovery requests that involve the production and review of large amounts of ESI.[13]

Predictive coding promises similar advantages in regulatory enforcement matters. Document review and production is generally a significant—if not the predominant—component of such matters at the initial stages. Thus, the technology presents an opportunity to reduce the cost of responding to requests from regulatory authorities at least as much, if not more, than in civil litigation. However, regulatory enforcement proceedings are not civil litigation, and the significant differences between the two should be considered before deciding to use predictive review as an aid to responding to regulatory requests.

Key Differences Between Regulatory Enforcement Proceedings and Civil Litigation

Abbreviated Timelines

One difference between civil litigation and regulatory enforcement proceedings is that the timelines in the latter are often much more abbreviated, as regulators often insist on extremely expedited production schedules. Predictive coding can, at least in theory, speed the response to a regulatory request for documents or to a subpoena, as the technology can significantly reduce the number of documents that require time-consuming human review. However, predictive review entails some start-up time not required in linear review, such as the iterative rounds of attorney coding, running computer predictions, and attorney feedback. This is particularly true if, as with the protocol in *Da Silva Moore*, coding decisions must be shared with the requesting regulator. In addition to considering whether the time allowed for production is sufficient to allow for the computer training and iterative rounds of review, counsel should also weigh whether the costs of these up-front activities are worth the later time savings. The larger the number of documents that need to be reviewed, the more likely the time saved from reducing the amount of human review at later stages will outweigh the time spent setting up the predictive review.

Abbreviated timelines can also complicate predictive coding insofar as they require rolling collections and productions of documents. When production deadlines are short, parties will often need to begin review before the full collection of documents is complete. Many predictive coding platforms work less well when collections are loaded on a rolling basis because the early training of the computer is based on an incomplete set of data. When documents are later added to the collection, additional training must be undertaken to account for the new documents. Rolling productions, which are also a common way of responding to regulatory requests, present a similar problem. If a regulator wants certain custodians or time periods prioritized for review and production, the predictive coding process may need to be run separately for each prioritized set, increasing the amount of time spent training the computer and complicating the overall workflow.

All of that said, predictive coding can be useful in regulatory investigations even if it does not, at the end of the day, reduce the time or person-hours it takes to prepare productions because predictive review can shorten the amount of time needed for counsel to become familiar with the facts. The time it takes to become familiar with the underlying facts in a matter and identifies critical documents is particularly important in enforcement matters. Quickly learning the facts of a matter puts counsel in a better position to negotiate with a regulator about the size and scope of the production request, focusing its attention on the relevant time periods, custodians, and issues, and potentially limiting overly broad requests.

---

[5] *Da Silva Moore*, 2012 BL 44145, slip op. at *11-12, *23.
[6] *Id.*, slip op. at *11-12.
[7] *Id.*, slip op. at *5
[8] *See id.*
[9] *Id.*, slip op. at *23.
[10] *Id.*, slip op. at *11 (internal quotation mark omitted, alteration in original).
[11] *Id.*, slip op. at *19 (quoting Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, RICH. J.L. & TECH., Spring 2011, at 48); *see also* Herbert L. Roitblatt at al., *Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review*, 61 J. AM. SOC'Y FOR INFO. SCI. & TECH. 70, 79 (2010) (''On every measure, the performance of the two computer systems was at least as accurate (measured against the original review) as that of human re-review.'').
[12] *See Da Silva Moore*, 2012 BL 44145, slip op. at *18.
[13] *Id.* (quoting Grossman & Cormack, *supra* note 11, at 43) (' 'The technology-assisted reviews require, on average, human review of only 1.9% of the documents, a fifty-fold savings over exhaustive manual review.' '').

  LW ISSN 0148-8139

No Neutral Arbitrator

Another important difference between civil litigation and enforcement proceedings is the absence of a neutral arbitrator to resolve disputes in the latter. Generally speaking, in a regulatory enforcement proceeding, the regulator has the final say on the way in which a document review is conducted. 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. Thus, for example, any search terms used to cull the set of documents prior to review and production will generally be shared with the regulator for its comment and approval; however, regulators are generally careful to reserve their rights to seek additional documents outside the scope of the search terms should the need arise.

Similarly, regulators are almost certain to want to sign off on the use of predictive coding before permitting its use to narrow the set of documents reviewed and produced. Thus, the use of predictive coding for production is essentially impossible without a regulator who understands and trusts the technology and is informed enough to vet a proposed methodology. As the technology is still in its infancy, many regulators may be hesitant to authorize use of a technology with which they do not have significant experience. On the other hand, some regulators are already agreeing to the use of predictive coding in specific matters, and one government agency has publicly acknowledged the importance of this and other new technologies in modern review. In proposed revisions to its rules, the Federal Trade Commission states that the parties responding to its requests may potentially ''utilize one or more search tools such as advanced key word searches, Boolean connectors, Bayesian logic, concept searches, predictive coding, and other advanced analytics.''[14]

Regulators who are open to the use of predictive coding may require assent not only to the decision to use predictive coding, but also to the 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. Regulators may go further and seek involvement similar to that permitted of the plaintiffs in the *Da Silva Moore* protocol—the right to review and challenge the producing party's coding of specific documents. This level of transparency could make regulators more comfortable with the review process because it exposes the criteria counsel uses to distinguish responsive from non-responsive documents. That said, this level of transparency, which is not typical in a linear review, comes with risks for producing parties, including 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.[15]

Finally, a producing party must consider the requirement by many regulators that a party certify the completion of its document production. In civil litigation, an agreement by the opposing party to a particular search methodology is effectively an acknowledgement that such a methodology satisfies the obligations imposed by the relevant rules of civil procedure, which typically boil down to a reasonableness standard. By contrast, even if a regulator has agreed up front to permit the use of predictive review, a regulator is unlikely to concede the sufficiency of the methodology for purposes of a producing party certifying that the production is complete. Regulators have substantial discretion over whether to certify a production, and even a preliminary decision not to certify completion could cause significant delay in the resolution of an investigation.[16] Even where counsel procures a regulator's prior agreement to the use of the technology and agreement that certification will be accepted using such technology, counsel (and the client) too must have sufficient understanding of and trust in predictive coding to be comfortable certifying the completeness of its productions.

Fewer Legal Limits on Scope of Review

Civil law suits in federal courts are governed by the Federal Rules of Civil Procedure, including Rule 26(b)(1), which limits discovery to documents relevant to a party's claims or defenses.[17] Proceedings in state courts have similar limitations.[18] Such limits do not apply in regulatory enforcement proceedings, where the limits on permissible discovery are much more expansive. Indeed, the incentive to cooperate discussed above, as a practical matter, significantly inhibits an investigated party's ability to contest the scope of a regulatory investigation. It is increasingly common for regulators to permit no, or only very light, relevance screens, and instead to demand production of all of each custodian's documents for a particular date range. Indeed, predictive review potentially makes this practice more attractive to regulators to the extent they can use the technology themselves to identify relevant documents without needing to undertake a costly manual review of the entire production.

Where a regulator does not permit a relevance screen on a production, predictive coding clearly has no role to

[14] FTC Rules of Practice, 77 Fed. Reg. 3191 (Jan. 23, 2012) (proposed rule amendments) (''Document discovery today is markedly different than it was only a decade ago. The growing prevalence of business files in electronic form . . . require[s] special skills and, if done properly, may utilize one or more search tools such as advanced key word searches, Boolean connectors, Bayesian logic, concept searches, predictive coding, and other advanced analytics.''), *available at* http://www.ftc.gov/os/2012/01/120113part2and4frn.pdf; *see also* 80 U.S.L.W. 982; and Craig A. Waldman et al., *Will Recent Court Approval of Computer-Assisted Document Review Spur Acceptance in Antitrust Investigations*?, Jones Day Publications (Mar. 2012), http://www.jonesday.com/will-recent-court-approval-of-computer-assisted-document-review-spur-acceptance-in-antitrust-investigations-03-14-2012/ (last visited Aug. 6, 2012).

[15] *See* Waldman et al., *supra* note 14 (noting the potential for document sharing to lead to broader document requests).

[16] *See id.* (raising the possibility that regulators may not view parties that use predictive coding as having ''substantially complied'' with production requests).

[17] *See* Fed. R. Civ. P. 26(b)(1) (''Parties may obtain discovery regarding any nonprivileged matter that is relevant to any party's claim or defense.'').

[18] *See, e.g.*, Del. Ch. Ct. R. 26(b)(1) (''Parties may obtain discovery regarding any matter, not privileged, which is relevant to the subject matter involved in the pending action, whether it relates to the claim or defense of the party seeking discovery or to the claim or defense of any other party.''); N.Y.C.P.L.R. § 3101(a) (''There shall be full disclosure of all matter material and necessary in the prosecution or defense of an action.'').

play in determining whether a document should be produced. However, it can still be useful as an efficient way to understand substantively what is being produced. Counsel could, for example, use predictive coding to highlight documents that are likely to be relevant, while still reviewing each and every document that is going to be produced. Or counsel could go a step further and use predictive coding to limit its review only to the documents the computer predicts are relevant, even while producing the larger universe of documents to the regulator. Of course, this means that counsel would be producing documents to a regulator that no human being had actually laid eyes on.

To the extent that counsel is producing documents without human review, there is the risk that the regulator will find documents and facts about which counsel is not fully informed. Counsel taking this approach must therefore have substantial confidence that its technology and process are sound, because the consequences of a mistake are not the typical consequences of a production that contains some irrelevant documents or is missing some relevant ones, but the arguably more damaging consequence of a production that contains relevant (and potentially important) documents about which counsel is entirely unaware. On the other hand, there is always a risk that human reviewers will miss or misunderstand the significance of important documents. Indeed, as *Da Silva Moore* notes, there is evidence that predictive review is actually more accurate than traditional human review, so the use of predictive coding may not necessarily increase the risk of missing important documents.[19] Nevertheless, counsel should proceed cautiously when pursuing a strategy of producing documents without any human review.

While a regulator may not permit a relevance screen on a production, it generally does (and must) permit a screen for privileged material. Predictive coding has not been as rigorously tested as a device for identifying potentially privileged documents. Since the technology works by identifying documents that are topically similar to one another, it may not be as effective at identifying privileged documents, where determinations often turn not on topical similarity of documents, but rather on very specific (and subtle) contextual differences between documents, such as whether a lawyer is included on a distribution list for purposes of seeking that lawyer's legal advice or for some other, non-legal purpose. While predictive coding is arguably valuable as a means of identifying potentially privileged material, it should probably be paired with more traditional methods of identifying potentially privileged material, such as the use of search terms.

Where documents are being produced without human review, counsel should also consider its client's risk tolerance for inadvertent disclosure of privileged documents. Federal Rule of Evidence 502 limits the risk that an inadvertent disclosure to a federal regulator will result in a subject-matter waiver, removing the threat of perhaps the most damaging consequence of an inadvertent disclosure.[20] Nevertheless, inadvertent disclosure could waive privilege with respect to the documents disclosed and, regardless, could reveal sensitive information that would not have otherwise been shared with the regulator. The risk of inadvertent production can be mitigated to some extent if the producing party has the ability to claw back privileged documents after production. However, regulators do not typically enter into claw-back agreements prior to production and some may resist claw-back of inadvertently produced documents entirely. Given these risks, counsel should investigate the legal and ethical duties related to the return or destruction of privileged documents in the pertinent jurisdiction before using predictive coding to conduct a privilege review.

Recommendations

In deciding whether predictive coding is appropriate in a regulatory enforcement matter, counsel should take into account the following considerations:

- **Consider whether the document volume, timing, and collection logistics will allow for predictive coding**. Extremely expedited schedules, especially when combined with rolling collections and productions, may not be ideal situations for the use of predictive review.

- **Consider whether the regulator has experience and comfort with predictive coding.** A less sophisticated regulator may be less likely to agree to predictive coding and/or more likely to refuse to or delay accepting a certification of completeness.

- **Make sure you have enough comfort with predictive coding to be able to make a certification of completeness.** Even if the regulator has agreed to predictive coding, you may still be required to certify independently to the efficacy of the methodology.

- **Consider what aspects of the methodology will require agreement with the regulator.** A highly transparent protocol such as that used in *Da Silva Moore* could complicate the review and open the door for an expanded inquiry. An alternative protocol might provide agreement on other details of the methodology—numbers, confidence intervals, or general relevance guidelines—to ease any concerns about the technology being a ''black box,'' while not being as intrusive as the *Da Silva Moore* protocol.

- **When the regulator insists on productions without relevance review, consider other methods in addition to predictive coding to identify privileged documents.** Also consider your risk tolerance for inadvertent disclosure and your ability to claw back any inadvertently produced documents.

[19] *See Da Silva Moore*, 2012 BL 44145, slip op. at *19 (quoting Grossman & Cormack, *supra* note 11, at 48) (' '[T]he myth that exhaustive manual review is the most effective—and therefore the most defensible—approach to document review is strongly refuted. Technology-assisted review can (and does) yield more accurate results than exhaustive manual review, with much lower effort.' ').

[20] *See* Fed. R. Evid. 502(a) (a disclosure of privileged materials made in a ''federal proceeding or to a federal office or agency'' that waives privilege with respect to those materials only waives privilege with respect to undisclosed materials if (1) the waiver was ''intentional''; (2) the disclosed and undisclosed materials ''concern the same subject matter''; and (3) ''they ought in fairness be considered together.''). For matters involving state regulators, consider whether there are any equivalent state law protections.

  LW ISSN 0148-8139