PREDICTIVE CODING: ADOPTING AND ADAPTING ARTIFICIAL INTELLIGENCE IN CIVIL LITIGATION

Gideon Christian*

This paper examines how predictive coding, an artificial intelligence (AI) technology, can effectively and efficiently complement the work of lawyers in the area of electronic discovery document review in civil litigation. It begins with a general overview of AI, and how machine learning can be used to automate the document review process in civil litigation. It then proceeds to a comprehensive overview of predictive coding technology and a discussion of legal issues related to the use of predictive coding technology in civil litigation. The legal issues are whether the use of artificial intelligence technology (as opposed to human intelligence) in document review complies with the rules of the court relating to documentary disclosure; and whether litigation privilege applies to seed sets (or training sets) used in training the predictive coding algorithm. Adopting a comparative law methodology, the paper seeks to address these issues. The paper concludes with a brief consideration of legal professionalism issues arising from the adoption of predictive coding technology in civil litigation in the context of Rule 3.1 of the Model Code of Professional Conduct dealing with competency. The paper argues that successful adoption of AI technology in civil litigation will extend the lawyer's duty of competence to include knowledge of the relevant legal technology.

Le présent article examine les façons dont le codage prédictif, une technologie d'intelligence artificielle (IA), peut venir appuyer de manière efficiente et efficace le travail des avocats relatif à l'examen des documents dans le cadre de communications de preuves électroniques en matière de contentieux civil. L'auteur débute par un survol de l'IA et examine comment l'apprentissage machine peut servir à automatiser le processus d'examen des documents dans le contexte de litiges civils. L'auteur donne ensuite un aperçu exhaustif de la technologie du codage prédictif et discute des questions juridiques que soulève le recours à cette technologie dans le domaine du contentieux civil. Les questions juridiques qui se posent se résument ainsi : le recours aux technologies d'IA (plutôt qu'à l'intelligence humaine) dans l'examen des documents est-il conforme aux règles des tribunaux relatives à la communication des documents? Le privilège relatif au litige s'applique-t-il à

* Assistant Professor (AI and Law), Faculty of Law, University of Calgary. Email: gideon.christian@uCalgary.ca; Legal Counsel (On Leave), Evidence Management Team (EvMT), Department of Justice Canada. The views expressed in this paper are strictly those of the author. They do not represent the views or the positions of the Department of Justice or those of the Government of Canada. The author is grateful to anonymous reviewers who provided helpful comments on earlier drafts of the manuscript.

l'ensemble des données servant de base d'apprentissage (seed sets ou training sets) qui ont été utilisées pour entraîner l'algorithme de codage prédictif? L'auteur adopte une méthodologie de droit comparé afin de répondre à ces questions.

En guise de conclusion, l'auteur se penche brièvement sur les questions de professionnalisme juridique qui pourraient découler de l'adoption de technologies de codage prédictif en contentieux civil par rapport à la règle 3.1 du Code type de déontologie professionnelle portant sur la compétence. Il soutient que l'adoption réussie de technologies d'IA dans le domaine du litige civil élargira le devoir de compétence de l'avocat pour y inclure des connaissances relatives aux technologies juridiques pertinentes.

Contents

Introduction

Artificial Intelligence (AI) technology is gradually and surreptitiously permeating diverse aspects of human life. AI is now being utilized in performance of tasks once considered to be within the exclusive domain of human intelligence. From high-tech driverless cars and trucks to basic facial recognition technology in your Facebook profile, the tentacles of AI are expanding and even the practice of law has not been out of reach. While it is unlikely that human society will reach the point where the work of lawyers is completely and efficiently replaced by AI, this research examines how predictive coding (an off shoot of AI) could effectively complement the work of lawyers—especially in the area of electronic discovery document review in civil litigation.

This paper begins with a general overview of AI, and machine learning as an aspect of AI, that is useful for the automation of the documentary disclosure process in civil litigation. It will then proceed into a comprehensive overview of predictive coding technology and a discussion of various legal issues related to the use of this technology in civil litigation. Two main legal issues will be considered in this regard. First, whether the use of artificial intelligence technology (as opposed to human intelligence) in document review complies with the rules of the court relating to documentary disclosure. A positive determination will support judicial mandate for the use of predictive coding technology in electronic discovery (e-discovery) document review. Adopting a comparative law methodology, the paper will consider the judicial approach to mandating the use of predictive coding technology in e-discovery document review in various jurisdictions. Second, the paper seeks to examine one important issue that remains unresolved in predictive coding litigation, namely whether litigation privilege applies to seed sets (or training sets) used in training a predictive coding algorithm. Even in the absence of any clear judicial decision to the effect that litigation privilege apply to seed sets, this paper takes the position that litigation privilege may apply to seed set depending on the methodology utilized in developing the seed set.

The paper concludes by examining the likely impact of the successful adoption of AI in civil litigation in the context of legal professionalism, especially as it relates to knowledge of legal technology by lawyers. A successful adoption of AI technology in civil litigation will extend the lawyer's duty of competence to include knowledge of the relevant legal technology.

1. Meaning and Scope of AI

AI is a concept that has defied a universally accepted definition even among experts in the field of computing. As noted by Scherer,[1] this difficulty lies with the rather "conceptual ambiguity of intelligence", which is often associated with human intelligence.[2] John McCarthy, a pioneer in the field of AI, once stated that "AI does not have to confine itself to methods that are biologically observable."[3] He went on to define AI as "the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence."[4]

At the early stage of AI, ability to perform intellectual tasks seem to be the focal point of definitional approach to AI. Scherer rightly observed that the "concepts of what constitutes artificial intelligence have shifted over time as technological advances allow computers to perform tasks that previously were thought to be indelible hallmarks of intelligence."[5] He further noted that "[t]oday, it appears that the most widely-used current approaches to defining AI focus on the concept of machines that work to achieve goals."[6] The concept of intelligence has gradually moved from the sole emphasis on human cognitive ability to incorporate the rationale ability to achieve defined goals. According to Scherer, the idea of AI is now associated with "machines that are capable of performing tasks that, if performed by a human, would be said to require intelligence."[7]

A) Machine Learning

Machine learning is a branch of AI in which computers 'learn' to perform some tasks and improve in the performance of the task over time through training using 'seed sets'. Thus, machine learning enables computers to perform tasks for which they are not explicitly programmed by developing intelligence from analysing training data. This process makes it possible for researchers to design computer programs to perform tasks that were once considered to be only capable of performance using human cognitive intelligence.[8]

---

1 Matthew U Scherer, "Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies" (2016) 29:2 Harvard JL & Tech 353.

2 *Ibid* at 359.

3 See John McCarthy, "What is Artificial Intelligence?" (2 November 2007), online (pdf): *Formal Reasoning Group* <www-formal.stanford.edu/jmc/whatisai.pdf>.

4 *Ibid* at 2.

5 Scherer, *supra* note 1 at 360.

6 *Ibid* at 361.

7 *Ibid* at 362.

8 Harry Surden, "Machine Learning and Law" (2014) 89:1 Wash L Rev 87.

Machine learning has been associated with the ability of computers to 'learn' from experience, and subsequently improve their performance, as a result of executing the same task over a period of time.[9] Surden noted that the use of 'learning' with reference to machine learning does not imply that computers are capable of possessing human cognitive abilities.[10] Rather, the concept of 'learning' in machine learning is in a functional sense—machines develop the capacity to change and better perform a given task as a result of experience acquired in the performance of similar or related tasks. In highlighting the 'intelligence' associated with machine learning, Surden stated that "[i]f performing well, machine learning algorithms may produce automated results that approximate those that would have been made by a similarly situated person."[11] Machine learning algorithm has been employed in many modern technologies such as speech recognition, facial recognition, auto-correct/predictive texting, spam filters in emails etc.

1) The Spam Email Analogy

A good example that can be used to illustrate the basic features of machine learning is the spam email analogy used by Surden in "Machine Learning and Law".[12] Spam emails usually constitute a nuisance to their recipients. Hence email service providers avail their users with the option to flag an email as spam. This enables the service to track similar emails and send them directly to the spam email folder rather than the user's inbox. This process entails the use of machine-learning algorithms, and the training of such algorithms, to detect the unique characteristics of spam emails by feeding the algorithms with examples (seed set) of junk emails.

For example, if a user receives a junk email from an online pharmacy requesting the user to place an order for Viagra or Cialis, the user could simply flag the email as junk or spam by sending it to the spam mail folder. This process of flagging the email and sending it to the spam folder serves to 'train' the machine-learning algorithm by providing it with seed set of spam emails to analyze. In doing this, the machine-learning algorithm will identify certain characteristics in the email from which it will learn to identify subsequent emails possessing those characteristics as spam emails. Such characteristics could include the domain name from which the email originated and words or phrases in the body of the email, such as "Cialis","Viagra", or "online pharmacy". The algorithm can use these

---

9 Stuart J Russell & Peter Norvig, *Artificial intelligence: A Modern Approach*, 3rd ed (New Jersey: Pearson, 2010) at 693.

10 Surden, *supra* note 8 at 89.

11 *Ibid* at 90.

12 *Ibid*.

and other characteristics to determine whether an incoming email is spam or not.

According to Surden, "machine-learning algorithms are able to automatically build such heuristics by inferring information through pattern detection in data. If these heuristics are correct, they will allow the algorithm to make predictions or automated decisions involving future data."[13] The ability of the machine-learning algorithm in this case to identify spam emails from other sources (aside from the online pharmacy) will improve with spam emails from other sources being flagged and sent to the spam email folder for analysis.

Surden also noted that, apart from learning characteristics that will enable it to identify an email as spam, the algorithm may also learn other characteristics that will enable it to identify an email as *not* being spam. Thus, it could learn that emails from individuals that the user had previously communicated with are not spam even if the emails from such individuals contain phrases like "Viagra" or "Cialis". As a result, the rate of accuracy in identification of spam email by the machine-learning algorithm improves with more seed sets being fed to and analyzed by the algorithm. According to Surden:

> This capability to improve in performance over time by continually analyzing data to detect additional useful patterns is the key attribute that characterizes machine learning algorithms. Upon the basis of such an incrementally produced model, a well-performing machine learning algorithm may be able to automatically perform a task—such as classifying incoming emails as either spam or wanted emails—with a high degree of accuracy that approximates the classifications that a similarly situated human reviewer would have made.[14]

This is not to imply though that machine-learning algorithms possess perfect accuracy, it is possible to have cases of false positives and false negatives.[15] That notwithstanding, with adequate training, machine-learning algorithms can achieve accurate result that meets or exceeds the rate achieved by humans—doing so in much shorter time.

---

13 *Ibid* at 91.

14 *Ibid* at 93.

15 Examples include situations where spam emails fail to be identified and flagged as such, and situations where emails that are not spam are wrongly flagged as spam and sent to the spam folder.

2) Predictive Coding

Predictive coding is a machine-learning process that relies on analysis of a sample data set or "seed set" to predict or classify documents in a larger dataset. The process involves "machine learning and a combination of different algorithmic tools."[16] Common tools employed in predictive coding include metadata searching, contextual searching, and concept searching. Concept searching includes:

> controlled vocabulary indexing (manual or automatic, with or without thesauri), multi-word phrase formation (by statistical and/or linguistic means), statistical query expansion methods, knowledge representation languages and inference systems from artificial intelligence, unsupervised learning approaches (including term clustering, document clustering, and factor analytic methods such as latent semantic indexing), as well as simple stemming, wildcards, spelling correction and string similarity measures.[17]

Unlike keyword search, which focuses on specific search term irrespective of the context, concept searching relies on the context in which a specific term is used. Thus, predictive coding is now increasingly used in the review of large document sets as well as sorting the documents into predetermined categories. The program also has the capacity to organize the documents using probability ranking, relevancy ranking, clustering or sorting the document by issue.[18] Charles Yablon and Nick Landsman-Roos have described predictive coding as "a process whereby computers are programmed to search large quantities of documents using complex algorithms to mimic the document selection process of knowledgeable, human document review."[19]

At the early stage of its use in document review, predictive coding attracted little interest until sometime in 2010 when results from two pilot research projects, which compared predictive coding to manual document review, were published.[20] Both studies concluded that the use of predictive coding in document review achieved a higher level of result than manual

---

16 Willis M Hampton, "Predictive Coding: It's Here To Stay" (2014), online (pdf): *Thompson Reuters Practical Law* <online.fliphtml5.com/yhnd/lqey/#p=1> at 29.

17 Douglas W Oard et al, "Evaluation of Information Retrieval for E-Discovery" (2010) 18:4 AI & L 347 at 360.

18 *Ibid*.

19 Charles Yablon & Nick Landsman-Roos, "Predictive Coding: Emerging Questions and Concerns" (2013) 64:3 SCL Rev 633 at 633.

20 Maura R Grossman & Gordon V Cormack, "Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review" (2011) 17:3 Rich JL & Tech 1 [Grossman & Cormack, "Technology-Assisted Review"]; Herbert L Roitblat et al, "Document Categorization in Legal Electronic Discovery:

review. Since then, support for use of predictive coding in document review has continued to grow. Today, the process has come to be variously referred to as technology-assisted review (TAR), or computer-assisted review (CAR).

Before going into details of how predictive coding, an aspect of AI, is now being applied in civil litigation, it is important to first provide an overview of the aspect of civil litigation where predictive coding (and hence AI) finds applicability.

2. Documentary Discovery in Civil Litigation

In Canada, discovery in civil litigation is governed by the rules of Court or rules of civil procedure. At the Federal Courts, the *Federal Courts Rules*[21] govern the commencement and conduct of proceedings, while in the province of Ontario, the *Rules of Civil Procedure*[22] apply. These rules also contain provisions relating to documentary disclosure. Rules 222–33 of the *Federal Court Rules* relates to documentary disclosure, and Rule 30 of the Ontario rules applies in the province.

In both sets of rules, "document" was defined to include electronically-stored information or information readable by means of a computer system or similar device.[23] The rules provide for the disclosure and production to the adverse party of copies of relevant documents in the party's possession and control. It imposes a dual discovery obligation on the parties— a duty to *disclose* the existence of all relevant documents in the party's possession and control, and the duty to *produce* for inspection to the other party(ies) all relevant documents for which the party (in possession and control) does not assert any claim of privilege.[24]

The disclosure obligation under the rules is discharged through the preparation and service on the other party of affidavit of documents, which lists and describes all relevant documents in the party's possession. The duty to produce relevant documents (except for those which the party asserts privilege) is discharged if upon request the party in possession or control avails the requesting party a reasonable opportunity to inspect the document and/or deliver to the other party copies of the relevant documents.[25]

Computer Classification vs. Manual Review" (2010) 61:1 J Am Society Information Science & Tech 70 at 74.

21 *Federal Courts Rules*, SOR/98-106 [*FCR*].

22 *Rules of Civil Procedure*, RRO 1990, Reg 194 [*RCP*].

23 *FCR*, *supra* note 21, s 222 (1); *Ibid*, s 30.01(1)(a).

24 *FCR*, *supra* note 21, ss 223(1), 228; *RCP*, *supra* note 22, s 30.02(1)(2).

25 *FCR*, *supra* note 21, s 228; *RCP*, *supra* note 22, s 30.04.

Thus, documentary disclosure under the rules of the courts would typically involve review by a party of documents in its possession or control (document review). The review is usually conducted by a review team composed of lawyers and other legal personnel such as paralegals. The document review will typically involve a two-stage process—review of documents for relevance and review of documents for privilege.

A) Relevance Review

This stage entails a review of documents in a party's possession or control to identify those relevant to the litigation. A document is relevant "if the party intends to rely on it or if the document tends to adversely affect the party's case or to support another party's case."[26] Relevance review thus requires a *bona fide* review by the party as it is obligated to disclose (as relevant) all documents related to the litigation which are beneficial or adverse to its interest. Relevance review will require a good understanding by the reviewers of the facts in issue and this could be gleaned from the pleadings filed by the parties. Documents which are identified as relevant at this stage of the review will then proceed to another stage of review.

B) Privilege Review

Documents that are identified as relevant are subject to further review for privilege. While parties are obligated to disclose all relevant documents, the obligation is subject to a valid claim of privilege. Hence, the privilege review stage involves identifying documents that contain privileged information for the purpose of excluding the documents from disclosure or redacting the privileged information before disclosure. Privilege review is usually one of the most critical and sensitive aspects of document review and requires detailed knowledge of the different types of privilege on the part of the reviewer(s).

Privileged documents are generally exempt from disclosure in litigation unless the privilege is expressly or impliedly waived by the party entitled. In the context of civil discovery, waiver could take the form of disclosure of a privileged document to the opposing party. By disclosing the document, the disclosing party is either expressly or impliedly indicating that it does not intend to assert the privilege. The issue becomes complicated where the disclosure is inadvertent. While privilege may be waived by a party entitled to rely on it and who *intends* to waive the privilege, inadvertent disclosure becomes problematic because of the absence of intention to waive. However, even in the absence of such intention, privilege may be lost as a result of inadvertent disclosure "based on considerations such as

26 *FCR*, *supra* note 21, s 222(2).

the manner of disclosure, the timing of disclosure, the timing of reassertion of privilege, who has seen the documents, prejudice to either party and the requirements of fairness, justice and search for truth."[27]

Inadvertent disclosure could arise from failure to identify privileged documents during the e-discovery document review process. It is also important for the disclosing party to timely reassert privilege over inadvertently disclosed documents. However, Master C MacLeod in *L'Abbe v Allen-Vanguard* cautioned that while "[i]nadvertence will not by itself amount to waiver but this does not mean the court will protect a party from reckless release of privileged documents. In any event notwithstanding the attempt to reassert privilege, the court may determine that privilege has been lost."[28]

The parties may seek to address the problem of inadvertent disclosure by way of a clawback agreement. While this clause may aid a party in recalling inadvertently disclosed privileged documents, and also prevent the opposing party from using the documents in court, the fact is that the clawback agreement provides very limited protection. A privileged document seen by the opposing party cannot be unseen. According to Wang:

> With lax privilege review, there is a greater probability that a privileged document will be revealed, and even if the opposing party cannot use the document in litigation, it will have seen it, likely be unable to forget it, and be able to use the information to strategize for its case.[29]

Thus, privilege review is a very delicate aspect of document review in the discovery process. It requires meticulous attention to details. This aspect of the review process ought to be handled with utmost caution and by very qualified legal professionals with sound knowledge of the various types of privilege as well as skills in detecting privileged documents in the review process.

3. Electronic Discovery in Civil Litigation

Traditional document review has often involved the review of sets of boxes of paper documents in a lawyer's office. However, the advent of digital technology has resulted in an exponentially increased volume of electronic (as opposed to paper) documents. While the advent of digital

---

27 *L'Abbe v Allen-Vanguard*, 2011 ONSC 7575 at para 37, [2011] OJ No 5982 (QL).

28 *Ibid* at para 38.

29 Jessica Wang, "Nonwaiver Agreements after Federal Rule of Evidence 502: A Glance at Quick-Peek and Clawback Agreements" (2009) 56:6 UCLA L Rev 1835 at 1846.

technology has come with a great deal of benefits, it has also compounded the burden of discovery as a result of the increased number of documents to review in a typical discovery process.[30] Thus, the traditional method of discovery and document review finds little utility in the new age of digital technology.

With the massive proliferation of electronic documents, document review in civil litigation has also taken on an electronic format. Legal technologies are now available to assist lawyers in review of documents for relevance and privilege before disclosure. This is an aspect of civil litigation where AI could be adopted and adapted to compliment the work of lawyers.

4. Predictive Coding: Adopting AI in Civil Litigation

Document review is a vital aspect of civil litigation practice. With an ever-increasing amount of electronically-stored information, keyword search has been (and still is) the primary method used for searching electronic documents in e-discovery document review. According to Oard:

> The term "keyword searching" … has been used in the IR [Information Retrieval] literature to refer to any or all of exact string matching, substring matching, Boolean search, or statistical ranked retrieval, applied to any or all of free text terms (e.g., space-delimited tokens or character n-grams), manually or automatically assigned controlled vocabulary terms, with or without augmentation by any combination of stemming, wildcards, multi-word phrase formation, proximity and/or word order restrictions, field restrictions, and/or a variety of other operators.[31]

Keyword search incorporates the use of connectors (e.g. "OR", "AND", etc.) and wildcard (e.g. "*") to locate documents containing relevant search terms. For example, a keyword search using "auto*" will locate terms such as auto, automatic, automobile, automate, autonomy, automatism etc. Hence the use of keyword search in litigation document review would require the lawyer to craft search terms that would result in locating relevant documents as well as narrow down the retrieval of irrelevant documents.

While keyword search is cost effective and efficient where the size of the document to be reviewed is not large, there are problems associated with keyword search. First is the problem of under or over-inclusiveness, which may result in the search capturing very few relevant documents or

---

30 Dana Remus, "The Uncertain Promise of Predictive Coding" (2014) 99 Iowa L Rev 101.

31 Oard, *supra* note 17 at 359–60.

a large set of non-relevant documents.[32] Keyword search is suitable where the search is for specific word(s) in a document irrespective of the context in which the word is used. Thus, the use of the keyword and wildcard "auto*" for search in the context of automobile will eventually capture words such as "autonomy" or "automatism" which are not necessarily related to automobile. This will result in over-inclusiveness. While the use of the keyword "car", for example, may not capture "automobile" or "vehicle" thus resulting in under-inclusiveness.

Another problem with keyword search is that it is ineffective where the document set is very large. In this regard, it has been noted that: "Data volumes are quickly becoming such that even with the best keyword search terms and an army of reviewers, it could still take months or years to sift through all the data and there would still be no guarantee of satisfactory results."[33] These problems with keyword search have prompted the push for the application of predictive coding, an AI technology, in electronic discovery document review. Predictive coding is based on concept searching and has been described as the "next generation of technology for electronic discovery."[34]

The first step in the use of predictive coding for document review would require developing a "seed set" or "training set". This refers to a set of documents that is randomly or judgmentally selected as sample from the entire document set to be reviewed. A person very knowledgeable with the litigation (usually a senior lawyer) would then review each of the documents in the seed set and code them accordingly. The coded documents from the seed set are then fed into the predictive coding software to "train" the software. The software analyzes the seed set for common concepts. From this analysis, it develops an internal formula for future prediction.

The software is then made to apply the algorithm in coding documents from the universal set. Samples from the computer coded documents are then reviewed by the lawyer(s), corrected and fed back into the system. The "training" of the software continues with further coding and feeding of documents until the software "learns" sufficiently to achieve a desired or acceptable rate of accuracy. The software is then made to apply the algorithm to the entire document set, coding documents and classifying them accordingly.

---

[32] Tonia Hap Murphy, "Mandating Use of Predictive Coding in Electronic Discovery: An Ill-Advised Judicial Intrusion" (2013) 50:3 Am Bus LJ 609.

[33] Victoria L Lemieux & Jason R Baron, "Overcoming the Digital Tsunami in e-Discovery: Is Visual Analysis the Answer?" (2011) 9:(1-2) CJLT 33 at 36.

[34] Murphy, *supra* note 32 at 616.

Predictive coding is now being touted as a veritable tool for advancing electronic discovery reform. One advantage of predictive coding lies in its ability to "filter out large swathes of documents that are likely to be irrelevant so that the attorney does not have to waste limited cognitive resources analyzing them."[35] Hence, predictive coding is more likely to return consistent and accurate results than keyword searching and manual linear review. It has been asserted that the use of predictive coding allows for significant cost reduction in the document review process especially where the document size is extremely large.[36] This assertion though has been countered by the fact that there is scanty empirical evidence to prove it.[37]

In the face of technology's changing landscape, as well as increasing numbers of electronic documents in civil litigation discovery process, it is now important for the legal profession to embrace this aspect of AI technology in legal practice, especially in the area of electronic-discovery document review. While this new technology holds a great deal of promise for civil discovery reform, its adoption in civil litigation discovery has been so slow that sometimes it has to be figuratively forced down the throat of unwilling litigants.[38] Many factors are individually and cumulatively responsible for the slow adoption of this AI technology in e-discovery. These factors include "lack of adequate technical understanding by lawyers, lack of transparency of the process, concern about accuracy of results and … 'the uncertainty of judicial acceptance.'"[39]

For predictive coding technology to find acceptability in civil litigation, there are important issues that need to be resolved. First, whether the court can rightly mandate the use of predictive coding in document review against the will of either party, and even when the use of technology-assisted review such as predictive coding is not expressly provided for in the rules of court or rules of civil procedure. Secondly, whether the seed sets or training sets used in training predictive coding algorithm in document review are entitled to litigation privilege. These issues will be examined below.

---

35 Surden, *supra* note 8 at 101.

36 See Master Mathews in *Pyrrho Investments Limited v MWB Property Limited*, [2016] EWHC 256 (Ch), [2016] 2 WLUK 413 [*Pyrrho*].

37 Murphy, *supra* note 32 at 620.

38 *Da Silva Moore et al v Publicis Groupe* (2012), 287 FRD 182, 868 F Supp 2d 137 [*Da Silva Moore*]; *Brown v BCA Trading Limited*, [2016] EWHC 1464, [2016] 5 WLUK 371 [*Brown*]; *Irish Bank Resolution Corporation Ltd & Ors v Quinn & Ors*, [2015] IEHC 175, [2015] 3 WLUK 71 [*Irish Bank*, 2015].

39 Murphy, *supra* note 32 at 624.

5. To Use or Not: Predictive Coding in the Courtroom

In most cases where parties to litigation have opposed the responding party's use of predictive coding in document review, one common reason for the opposition is based on the argument that the use of the technology does not comply with the rules. Thus, courts in various jurisdictions have had to decide whether to mandate the use of predictive coding in document review in cases before them. In deciding this novel and controversial issue, the courts have had to review the technology and its compliance with discovery rules in their respective jurisdictions. Some of the approaches taken by courts in the resolution of this issue in various jurisdictions will be discussed below.

A) United States

For the first time in its judicial history, a US court was in 2012 faced with a serious question of whether to mandate the use of predictive coding in e-discovery. This was the case of *Da Silva Moore*, *et al v Publicis Groupe*.[40] The case involved gender discrimination litigation against an advertising conglomerate and its US subsidiary. For proponents of predictive coding, the fact that this issue came before Magistrate Andrew J Peck, a judge very knowledgeable of the predictive coding technology, was like icing on the cake.[41] Magistrate Peck had to decide whether to approve the defendants' proposal to use predictive coding to cull down some three million electronic documents from various custodians.

The Magistrate started by reciting portions from his article on predictive coding wherein he expressed an opinion that "computer-assisted coding should be used in those cases where it will help 'secure the just, speedy, and inexpensive' determination of cases in our e-Discovery world."[42] He went on to review various alternatives to predictive coding as well as plaintiffs' concerns with the defendants' predictive coding protocol. Other alternatives reviewed by the learned Magistrate included manual linear reviews and keyword searches. With respect to manual review, Magistrate Peck noted that, even though it is considered by lawyers to be the "gold standard", it would be too expensive in the present case which

---

40 *Da Silva Moore*, *supra* note 38.

41 Sensing the defence counsel's delight that the case was referred to him, Magistrate Peck told the counsel, "You must have thought you died and went to Heaven when this was referred to me". An obviously excited defence counsel responded: "Yes, your Honor. Well, I'm just thankful that, you know, we have a person familiar with the predictive coding concept": *Da Silva Moore*, *supra* note 38 at 184, n 3.

42 Andrew Peck, "Search Forward: Will Manual Document Review and Keyword Searches be Replaced by Computer Assisted Coding?" (2011) 206:2 New Jersey LJ 30.

involved more than three million documents.[43] Referencing statistics from research studies, he concluded that technology-assisted reviews yield more accurate results than manual review.[44]

With respect to a keyword search, the Magistrate noted, among others, the problem of over-inclusiveness. Responding to the plaintiffs' reservation with predictive coding, which they referred to as a new technology that had to be "proven out", Magistrate Peck reminded the parties that predictive coding: "works better than most of the alternatives, if not all of the [present] alternatives. So, the idea is not to make this perfect, it's not going to be perfect. The idea is to make it significantly better than the alternatives without nearly as much cost."[45] The Magistrate thus ruled that the use of predictive coding in the e-discovery process was appropriate. He based his decision on five factors: (1) the agreement of the parties;[46] (2) the amount of documents involved; (3) the finding that predicting coding works better than other alternatives; (4) cost effectiveness and proportionality; and (5) defendants' transparency in the discovery process.[47] The Magistrate's assertion that predictive coding was agreed upon by the parties resulted in even more controversy than resolution of the problem.[48] The plaintiff asserted that they never consented to any agreement to use predictive coding. They appealed the ruling to the District Judge Andrew Carter Jr. on the ground that the use of predictive coding was unreliable and unacceptable under the *Federal Rules of Civil Procedure*. Judge Carter Jr., while noting the discrepancy about the plaintiffs' consent to the use of predictive coding, affirmed the decision of Magistrate Peck and stated that the confusion relating to the plaintiffs' consent was immaterial. What is material is the fact that there is no evidence to conclude that the use of predictive coding will deny the plaintiffs access to discovery as required

---

43 *Da Silva Moore*, *supra* note 38 at 190.

44 *Ibid* at 18–19. See Herbert L Roitblat, Anne Kershaw & Patrick Oot, "Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review" (2010) 61:1 J Am Society Information Science & Tech 70; Grossman & Cormack, "Technology-Assisted Review", *supra* note 20 at 48.

45 *Da Silva Moore*, *supra* note 38 at 187.

46 A review of the predictive coding protocol proposed by the defendants and attached as an exhibit to the court's decision shows that the plaintiffs were not clearly in agreement with the use of predictive coding in the discovery process. See para A(1) and J(1) of Exhibit to the Order.

47 *Da Silva Moore*, *supra* note 38 at 192.

48 The plaintiffs rather asserted that they were "steamrolled" by Magistrate Peck. This resulted in their application for recusal by Magistrate Peck. They referenced Magistrate Peck's out of court advocacy for predictive coding, his ties to the predictive coding vendors evident from teaching fees for promoting predictive coding, his participation in e-discovery panels with the defendants' e-discovery counsel among others. Magistrate Peck refused to recuse himself. The challenge was overruled by the District Court Judge and a petition to the Supreme Court was unsuccessful.

under the rules, and if on production the plaintiffs determine that relevant documents are missing, the issue may be revisited.[49]

Shortly after the decision in *Da Silva Moore*, the issue came up again before the Virginia Court in *Global Aerospace Inc. v Landow Aviation*.[50] The size of the documents involved in the review was some 250 gigabytes of documents (representing more than 200 million documents).[51] The defendants made a strong case for the use of predictive coding over manual review or keyword search citing efficiency, lower costs and better results. They also asserted that their request to use predictive coding falls within the Rules of the Supreme Court of Virginia and leading jurisprudence.

The plaintiffs opposed the use of predictive coding, characterising it as "a radical departure from the standard practice of human review of documents."[52] They insisted on the use of the traditional human review method. In a very short decision, County Judge Chamblin allowed the defendants to proceed with the use of predictive coding. Thus, *Global Aerospace* was added to the list of cases where a US court had sanctioned the use of predictive coding. Unfortunately, the decision is so brief that it fails to provide any detailed analysis of the parties' arguments or the rationale for Judge Chamblin's decision.

The importance of transparency when seeking the mandate of the court to use predictive coding—especially where the parties are in dispute as to such use—was evident in *Progressive Casualty Insurance Company v Delaney*.[53] In *Progressive*, the parties agreed on a joint proposed ESI protocol that was approved by the court. Under the protocol, the parties agreed to the use of search terms to scan through entire documents for responsive documents. Progressive was to either produce all responsive non-privileged documents (subject to clawback agreement), or manually review "hit" documents before production. Progressive opted for the latter. The agreed upon search terms generated some 565,000 documents. Progressive began manual review of the documents for relevance. Six to eight months later, having concluded that manual review presented "an unacceptable high cost", Progressive began to explore other alternatives.

---

49 *Da Silva Moore et al v Publicis Groupe*, 2012 WL 1446534 (US Dist Ct) [*Da Silva Moore* Appeal] at 2.

50 *Global Aerospace Inc v Landow Aviation LP*, 2012 WL 1431215 (Va Cir Ct) [*Global Aerospace*].

51 *Global Aerospace Inc v Landow Aviation LP* (Memorandum, Landow Aviation Limited Partnership) at 2 [*Global Aerospace* Memo].

52 *Ibid* at 2 (Brief in Opposition to Motion for Protective Order Regarding Electronic Documents and "Predictive Coding").

53 *Progressive Casualty Insurance Company v Delaney*, 2014 WL 12785311 (WL US) (D Nev Dist Ct 2014) [*Progressive Casualty Insurance*].

Thereafter, Progressive unilaterally opted to use predictive coding to review the "hit" documents. Failed attempts to obtain detailed information from Progressive about its alternative proposal resulted in a Motion to Compel. Judge Leen noted the courts disposition to the use of predictive coding in document review. According to him "[h]ad the parties … agreed at the onset of this case to a predictive coding-based ESI protocol, the court would not hesitate to approve a transparent, mutually agreed upon ESI protocol."[54]

Having considered the unique facts of the case, Judge Leen refused Progressive's request to use predictive coding in the document review. Many reasons where given for the refusal. First, Progressive abandoned its own protocol, not because it was not responsive to documents, but because it was costly and time consuming. Second, Progressive was unwilling to engage a cooperative and transparent protocol for a reasonable search acceptable to the court and the parties. Judge Leen noted Progressive's lack of "transparency and cooperation regarding the search methodologies applied."[55] Third, Progressive was only willing to apply predictive coding to a limited set of documents as opposed to the entire set of documents. Finally, Progressive unilaterally abandoned the court approved protocol without consultation with the opposing party or the approval of the court.[56] It is important to note that the decision in *Progressive* was not a judicial indictment or distrust of the predictive coding technology. The negative decision in this case was more of an indictment on the party's conduct than on the technology the party sought to use. Judge Leen left no doubt in her ruling that a court would readily order the use of predictive coding where there is "an unprecedented degree of transparency and cooperation among counsel in the review and production of ESI responsive to discovery requests."[57] Thus, lack of transparency by Progressive was the reason for the court's refusal to order the use of predictive coding in this case.

---

54 *Ibid* at 13.

55 *Ibid*.

56 *Ibid* at 16.

57 *Ibid* at 15. In *Dynamo Holdings Limited Partnership et al v Commissioner of Internal Revenue*, 143 TC 183 (Wash 2014), Memo 2018-61, the United Stated Tax Court sanctioned the use of predictive coding. In response to the Respondent's assertion that predictive coding is an "unproven technology", the Court stated at 15:

> [w]e disagree. Although predictive coding is a relatively new technique, and a technique that has yet to be sanctioned (let alone mentioned) by this Court in a published Opinion, the understanding of e-discovery and electronic media has advanced significantly in the last few years, thus making predictive coding more acceptable in the technology industry than it may have previously been. In fact, we understand that the technology industry now considers predictive coding to be widely accepted for limiting e-discovery to relevant documents and effecting discovery of ESI without an undue burden.

The decision in *Progressive* can be contrasted with *Bridgestone Americas Inc. v International Business Machines Corp.*[58] In *Bridgestone*, just as in *Progressive*, the parties had previously agreed to a protocol that was subsequently sanctioned by the court. Using search terms provided by the defendant, the plaintiff was faced with a review of 2 million documents. Rather than proceeding with the manual review as previously agreed, the plaintiff sought for permission to use predictive coding instead. The defendant opposed the move as an unfair and unwarranted change to the original order of the court. Upon a review of the extensive pleadings filed by the parties, the learned Magistrate Judge noted that the matter boiled down to a "judgment call" about efficiency and cost-efficacy. Taking into consideration the volume of documents involved (about 2 million) and the transparency and openness of the plaintiff evident from its promise to share seed set documents with the defendant, the Judge allowed the plaintiff's switch from manual review to predictive coding.

Thus, a plethora of cases from the US demonstrate that the use of predictive coding in discovery complies with discovery obligation under the various rules of civil procedure, and is now judicially sanctioned by various courts in that jurisdiction. There is no reason to believe or speculate that this trend will change. Rather, decisions in various US courts have continued to blaze judicial trends in other jurisdictions, as can be seen from an examination of selected cases in the United Kingdom and Ireland.

B) United Kingdom

The case of *Pyrrho Investments Limited v MWB Property Limited*[59] was historically the first case in which an English court was asked to make an order mandating the use of predictive coding in e-disclosure. In the substantive litigation, which was an action for breach of fiduciary duty, the original number of documents involved in the discovery exercise was in the range of 17.6 million. Through a process of de-duplication, the number was ultimately reduced to some 3.1 million documents, and even then the latter number was considered large and costly to search.

The presiding Magistrate, Master Mathews, was faced with a very novel application regarding the electronic disclosure obligation of the parties. But his task was made somewhat easier by the fact that the parties to the litigation had all consented to the use of predictive coding. Thus, Master Mathews had the task of making the order for the use of predictive

---

[58] *Bridgestone Americas Inc v International Business Machines Corp*, 2014 WL 4923014, 172 F Supp (3d) 1007 (MD Tenn Dist Ct 2014) [*Bridgestone Americas*].

[59] *Pyrrho*, *supra* note 36.

coding and stating the reasons for his decision. Considering the novelty of the application, Master Mathews unsurprisingly embarked on a detailed analysis of rules governing e-disclosure. He noted that disclosure in the court is governed by Part 31 of the *Civil Procedure Rules*[60] and its Practice Directions as well as directions given by the Court.[61] The rules and provisions governing disclosure show that what is vital in the discovery process is the scope and quality of the search as opposed to listing and production. Hence, any defect in search cannot be remedied at the listing and production stage.

Notwithstanding the importance of search in e-disclosure, the Rules lacked a detailed guide on how the search is to be conducted. The Practice Direction, however, contemplated the use of "automated methods of searching", a term that could reasonably accommodate the use of predictive-coding software.[62] In justifying his decision to order the use of predictive coding in the discovery process, Master Mathews reviewed various authorities in England,[63] the United States,[64] and Ireland.[65] He listed various factors justifying his order for the use of predictive coding. These include the experience in other jurisdictions such as the United States and Ireland, where the courts have variously endorsed the use of predictive coding even in contentious cases, and the fact that the parties in this case have consented to the use of predictive coding. Master Mathews

---

60 *Civil Procedure Rules* (UK), 31B PD 25, Rule 31 [*CPR* (UK)].

61 *Ibid* at para 6.

62 *CPR* (UK), *supra* note 60, Rules 25 & 26 [emphasis added] of the Practice Direction provides as follows:

> 25. It may be reasonable to search for Electronic Documents by means of Keyword Searches *or other automated methods of searching* if a full review of each and every document would be unreasonable.
>
> 26. However, it will often be insufficient to use simple Keyword Searches or other automated methods of searching alone. The injudicious use of Keyword Searches *and other automated search techniques*—
>
> (1) may result in failure to find important documents which ought to be disclosed, and/or
>
> (2) may find excessive quantities of irrelevant documents, which if disclosed would place an excessive burden in time and cost on the party to whom disclosure is given.

63 *Goodale and Ors v the Ministry of Justice and Ors*, [2010] EWHC B40, 2009 WL 6326685 (QB); Celina McGregor, "Keywords aren't Enough Anymore" (2014) Solicitors J 27; Technology and Construction Solicitors' Association, *Guide to eDisclosure*, Version 0.1, 1 November 2013; Charles Hollander QC, *Documentary Evidence*, 12th ed (London: Sweet & Maxwell, 2015).

64 *Da Silva Moore*, *supra* note 38.

65 *Irish Bank Resolution Corporation Ltd v Quinn*, [2012] NICh 1, [2012] BCC 608 (I HC) [*Irish Bank*, 2012].

also noted that there is no provision in the *Civil Procedure Rules* or Practice Directions which prohibits the use of the technology.

Other factors cited by the learned Master in the justification of his order included the volume of documents to be reviewed (over 3 million) and the cost of manually searching the document (which was considered unreasonable) compared to predictive coding, which was held to be far less expensive and a viable alternative. This cost was held to be proportionate to the value of the claim in the substantive litigation. Suffice it to say though that predictive coding's cost-saving postulation has often been met with criticism relating to the absence of empirical evidence to substantiate the claim. It is noteworthy that notwithstanding the dearth of such empirical evidence in this regard, the Master accepted the cost-saving postulation. Though the non-contentious nature of the case may explain this.

Master Matthews also noted the absence of any evidence to show that the use of predictive coding would produce less accurate result compared to manual review alone, or keyword searches and manual review combined.[66] It is important to note that although Master Mathews was all in favour of mandating the use of predictive coding in this case, he made it clear that the applicability of predictive coding in other cases would depend on the facts of each case.[67]

Shortly after the decision in *Pyrrho Investments Limited v MWB Property Limited*, the case of *Brown v BCA Trading Limited*,[68] another predictive coding litigation, came before the English court. Unlike *Pyrrho*, which was non-contentious, the parties in *Brown* could not agree on the use of predictive coding. The respondents at the case management hearing presented evidence to show why they believe predictive coding was "the most reasonable and proportionate method of disclosure."[69] This evidence included cost implications in favour of predictive coding.[70] Surprisingly, the evidence was not contradicted by the opposing party.

Although the cost postulation was accepted without contradiction, Mr. Registrar Jones noted that the cost benefit would only be relevant and persuasive if it was shown that predictive coding would be effective to achieve the disclosure required.[71] In the end, the decision to order the use of predictive coding in this case was based on two factors. Firstly, the fact

---

66 *Pyrrho*, *supra* note 36 at para 33.

67 *Ibid* at para 34.

68 *Brown*, *supra* note 38.

69 *Ibid* at para 2.

70 *Ibid* at para 3. While use of predictive coding would incur cost in the range of £132,000, keyword searching was estimated to costs at least £250,000 or more.

71 *Ibid* at para 4.

that predictive coding would at least be able to identify documents which could otherwise be identified using other search methods such as keyword searching. Secondly, the court accepted the cost saving argument that predictive coding would be considerably cheaper than keyword search.

In reaching a decision, Mr. Registrar Jones examined the ten factors outlined by Master Mathews in *Pyrrho*, finding that nine of the ten factors (to some extent) applied to the case before him.[72] He noted that predictive coding was new hence there may be concern about its effectiveness compared to other traditional but expensive search methods. This concern was mitigated by the fact that the decision to order predictive coding will "cause the parties to sit down before the predictive coding begins in order to discuss the criteria to adopt and the general process of disclosure."[73] At the end, Mr. Registrar Jones left the door open for the parties to return to the court for further direction should any new problem arise during the discovery process.

C) Ireland

The case of *Irish Bank Resolution Corporation Ltd & Ors v Quinn & Ors*[74] was the first before an Irish court in which a party sought an order from the court for the use of predictive coding in document discovery. Initial use of keyword search in scoping relevant documents in the case yielded some 1.7 million documents. The number was substantially reduced to some 670,809 documents following de-duplication. At the onset of the discovery process, the defendants partially consented to the use of predictive coding. However, their position changed following a meeting in which the plaintiffs' experts provided details regarding the search terms used, issues relating to transparency, as well as the probability of documents being overlooked. The defendants thereafter reversed their position, asserting rather that predictive coding does not comply with the rules of the courts. The court thus had to determine for the first time whether the use of predictive coding in discovery complied with the *Rules of the Superior Court*,[75] and the appropriate methodology to be used in this case.

What is noteworthy about this case is the fact that the plaintiffs came to the court with substantive expert and empirical evidence to persuade the court to grant the request to use predictive coding in the review process.

---

72 The tenth factor did not apply because the parties in the present case (unlike those in *Pyrrho*) did not consent to the use of predictive coding.

73 *Brown*, *supra* note 38 at para 13.

74 *Irish Bank*, 2015, *supra* note 38.

75 *Rules of the Superior Court* (I), SI 1986/15, Order 31, r 12.

The plaintiffs estimated that with a team of 10 experienced reviewers using a traditional linear review, it would take about nine months and cost €2 million (in addition to supervision and technology costs) to complete the review. The plaintiffs' expert suggested that using predictive coding would require "a fraction of the cost" and could be completed within a shorter timeframe.

The plaintiffs presented their proposed ten-stage protocol before the court. The protocol addressed issues relating to transparency, keyword searches, disclosure procedure for documents to be used in the training set, and the process for conducting a linear review for relevance and privilege in respect of applicable set of documents. The protocol also outlines quality control measures as well as disclosure procedure on the completion of the review. With regards to the latter, the protocol states: "The plaintiffs will, when making discovery, produce an expert certificate confirming that the TAR [Technology Assisted Review] was statistically valid and providing a detailed basis for drawing that conclusion."[76]

The court then went on to address the objections raised by the defendants. One of the grounds for the defendants' objection was that predictive coding would not identify all relevant documents and, as a result, was incompatible with the obligation of the disclosing party. This objection seems to be at odds with reality as no discovery methodology can identify all relevant documents in discovery process. Thus, the court rightly noted that all review methods would inevitably result in omission of some relevant document, and compared to manual review, predictive coding results in fewer omissions of relevant documents.[77]

In addressing the defendants' objection that predictive coding does not comply with the rules of the court, Justice Fullam noted that the rules of the court contained no provision for the adoption of predictive coding or manual review in discovery process. In addressing this procedural vacuum, Justice Fullam examined various authorities from within and outside the jurisdiction to develop a principle to guide the court in choice of review methodology.[78] A party making discovery has a duty to undertake a reasonable search for and disclosure of documents.[79] In examining this

---

76 *Irish Bank*, 2015, *supra* note 38 at para 46.

77 *Ibid* at para 47.

78 *Atlantic Shellfish Ltd v Cork County Council*, [2007] IEHC 215 [*Atlantic Shellfish*]; *Thema International Fund Plc v HSBC Institutional Trust Services (Ireland)*, [2011] IEHC 496 [*Thema*]; *Ryanair Plc v Aer Rianta CPT*, [2003] IESC 62; *Dome Telecom Ltd v Eircom Ltd*, [2007] IESC 59 [*Dome*]; Grossman & Cormack, "Technology-Assisted Review", *supra* note 20; Andrew Peck, *supra* note 42; *Dynamo Holdings v CIR*, 143 TC 183, Tax Ct Rep (CCH) 60, 021 (US 2014).

79 *Atlantic Shellfish*, *supra* note 78.

duty, the court takes into consideration the fact that, where the document is large, there might be innocent failure to identify and disclose relevant documents.[80]

Moreso, in the absence of a specific rule relating to the use of predictive coding, Justice Fullam took the position that the court has an inherent power to develop its own procedure to fill the vacuum, doing so in a manner that is equitable and taking into consideration the "objectives of expedition and economy".[81] The learned judge agreed that in reviewing large sets of documents "using predictive coding is at least as accurate as, and, probably more accurate than, the manual or linear method in identifying relevant documents."[82] He also agreed that as long as sufficient transparency is shown, predictive coding complies with the parties' discovery obligation under the rules. He then went on to lay down a general principle to guide the court in making an order for use of predictive coding: "Provided a party seeking to make discovery using predictive coding acts *bona fide* and the proposed system is transparent, opposition or non-cooperation by the requesting party should not deter the Court from making an appropriate order."[83]

Having examined the plaintiffs' proposed protocol and their conducts during the course of the discovery process, Justice Fullam was satisfied that they acted *bona fide* and their proposal was transparent. He then went on to endorse the use of predictive coding in the discovery process.

D) Canada

Recent judicial decisions in Canada (though sparse) seem to suggest a willingness by Canadian courts to accept predictive coding technology in litigation document review. This was evident in the decision of Canada's Competition Tribunal in *The Commissioner of Competition v Live Nation Entertainment Inc. et al*[84] In this motion before the Tribunal, the Commissioner sought an order from the Tribunal compelling the Respondents to produce further and better affidavits of documents. The Respondents had previously stated in their Affidavit of Documents (AODs) that "[t]he documents listed herein, if any, were located through

---

80 *Thema*, *supra* note 78.

81 *Irish Bank*, 2015, *supra* note 38 at para 65, citing Justice Geoghegan in *Dome*, *supra* note 78.

82 *Ibid* at para 66.

83 *Ibid* at para 69.

84 *The Commissioner of Competition v Live Nation Entertainment Inc. et al*, 2018 CACT 17. The Competition Tribunal is a specialized restrictive trade practices administrative tribunal composed of judges of the Federal Court and expert lay people.

the use of technology-assisted review …".[85] The Commissioner objected to the Respondent's AOD on the ground that the search for the documents was clearly inadequate as it produced fewer documents than expected. The respondents' document collection process had initially resulted in the collection of some 2.5 million documents. Counsel reviewed some 8,287 sample documents (seed set) which was then used to train and validate the predictive model.[86] Using predictive coding, the potentially relevant documents were culled down to 55,000 documents and the task was completed within a relatively short period of time.

Unlike the American case of *Da Silva Moore*, the requesting party in *Live Nation Entertainment* did not oppose the use of predictive coding technology in the review of the documents. Noting this, the Tribunal then went further to provide what appears to be the most explicit endorsement of predictive coding technology by any Canadian court or tribunal, stating: "[t]he Tribunal encourages the use of modern tools to assist in these document-heavy cases where they are as or more effective and efficient than the usual method of document collection and review."[87] Prior to *Live Nation Entertainment*, the most relevant case in Canada in which the issue of predictive coding arose was the Ontario Superior Court case of *Bennett v Bennett Environmental Inc*.[88] The issue before the Ontario court related to (among other things) the reasonableness of the fees charged by counsel who used predictive coding to conduct a first level review of a massive "document dump", thus substantially reducing the number of documents that were subsequently subject to human review by paralegals. In his ruling, Justice Mesbur upheld the reasonableness of the document review fees. According to the Judge: "Given the use of predictive coding for the first level review of massive document disclosure, I do not find it unreasonable for the lawyer to then use paralegals to conduct the next level or levels of review. I make no adjustment on this account."[89]

Unlike the other jurisdictions discussed above, there is paucity of Canadian judicial decisions endorsing predictive coding technology in litigation document review. This should not be misconstrued to mean that the technology is uncommon in Canadian litigation practice. Rather,

---

85 *Ibid* at para 5. Technology Assisted Review was used to refer to predictive coding.

86 *The Commissioner of Competition v Live Nation Entertainment Inc et al*, "Respondents Motion Record" (1 October 2018), online (pdf): *Competition Tribunal* <www.ct-tc.gc.ca/CMFiles/CT-2018-005_Respondents Motion Record in response_25_66_10-1-2018_6545.pdf> at paras 11, 14.

87 *Ibid* at para 15.

88 *Bennett v Bennett Environmental Inc*, 2016 ONSC 503, 2016 CarswellOnt 670 (WL Can).

89 *Ibid* at para 44.

when it comes to the use of this new technology in litigation document review, Canadian litigants are less contentious than their counterparts in other jurisdictions. Litigants seem to be more concerned with the adequacy or thoroughness of documentary disclosure by the opposing party than the technology used in the document review process. Another explanation may lie in the fact that litigants who use the technology in document review do not initially seek the consent of the opposing party or the court prior to using the technology. This reduces the likelihood of the technology becoming a contentious issue in litigation before the court. It would be gratifying to see more Canadian judicial endorsements of a technology that has the capacity to revolutionize and modernize electronic document discovery process in civil litigation.

6. Predictive Coding: Seed Sets and Privilege

One issue that remains unresolved in predictive coding jurisprudence is whether litigation privilege applies to seed sets used in training the predictive coding algorithm. To date, no court seems to have ruled on this issue. Although in some cases, the courts have made pronouncements indicating the applicability of litigation privilege to seed sets,[90] in some other cases, courts have ordered parties to disclose their seed sets to opposing parties without any consideration of the likelihood of privilege applying to the set.[91]

While the use of predictive coding technology in litigation document review has been welcomed by courts in different jurisdictions, the fact remains that the acceptability of this technology could be impeded by legal uncertainties surrounding the obligation of a party to disclose its seed sets in an effort to meet the so-called 'cooperation and transparency' requirements by the courts. Resolution of this issue would require a determination of the legal status of seed sets, i.e. whether they are subject to privilege or not. If seed sets are subject to privilege, then compelling or coercing a party to disclose privileged documents to the opposing party (in an effort to convince the court that the party seeking to use predictive coding is 'cooperative and transparent') would not only amount to an affront on the bedrock of our adversarial litigation system, but could also impede the acceptability of predictive coding technology in civil litigation.

---

[90] While the court in *Re: Biomet M2a Magnum Hip Implant Products Liability Litigation*, 357 F Supp (3d) 1389 (ND Ind 2013) seems to have taken that position, it refused the disclosure of the seed sets to the requesting party on the basis that the request for the seed set amounts to discovery within discovery. The court did not provide any legal justification or basis for treating seed set as litigation privileged.

[91] *Federal Housing Finance Agency v JPMorgan Chase & Co*, 902 F Supp (2d) 476, Fed Sec L Rep P 97075 (SDNY 2012) [*Federal Housing Finance*].

Going further, this research will examine the applicability of privilege to seed sets.

A) Developing seed set in predictive coding

A seed set (also referred to as a 'training set') is a sample of documents selected from a larger document set and then used to train the predictive coding algorithm to enable it to identify documents with similar information or characteristics in the larger database. The ability of the predictive coding algorithm to properly identify documents in the larger database is subsistent on properly selecting the seed set. As such, the "garbage in, garbage out" (GIGO) principle will apply in the case of a poorly crafted seed set. There are three different approaches to development of seed sets used in training a predictive-coding algorithm.

The random sampling approach to the development of a seed set involves using a statistical method that is equally likely to select any document from the database for inclusion in the sample set.[92] This process generates a seed set by "taking a statistically valid sample from the universe of potentially responsive information."[93] The random samples could be generated using a function of the predictive coding technology.[94] One basic feature of the random sampling technique is that it involves little to no human intellectual exercise in the selection of the sample documents. Hence, no preference is given to any document in the selection process, implying that each document has an equal chance of being selected as sample irrespective of the information (relevant or not relevant) it contains. Once the documents have been selected by the computer software, or any automated process, the selected document will then be reviewed and coded by a lawyer as relevant or not relevant before being used to train the predictive coding algorithm.[95]

The judgmental sampling approach involves the extensive use of human intelligence to select documents that meet or conform to the selector's bias or predetermined criteria. In judgmental sampling, counsel relies on their opinion of the case before them and independent legal judgment in the selection of documents for use as seed set. Unlike random sampling where all the documents in the database have an equal probability of selection, in judgmental sampling, the documents selected

---

92 See Maura R Grossman & Gordon V Cormack, "The Grossman-Cormack Glossary of Technology-Assisted Review" (2013) 7:1 Fed Cts L Rev 1 at 27.

93 John M Facciola & Philip J Favro, "Safeguarding the Seed Set: Why Seed Set Documents May be Entitled to Work Product Protection" (2015) 8:3 Fed Cts L Rev 1 at 15.

94 *Ibid*.

95 Charles Yablon & Nick Landsman-Roos, "Predictive Coding: Emerging Questions and Concerns" (2013) 64 SCL Rev 633 at 639.

are not typically representative of the entire database. Rather, the choice of document is skewed in favour of the lawyer's predetermined criteria based on knowledge of the case and litigation strategy. Therefore, in judgmental sampling, the choice of documents for use as a seed set is based on a lawyer's exercise of skill, judgment, and reasoning. Facciola and Favro have noted that the distinguishing characteristics of seed sets developed from judgmental sampling is that they are meticulously selected by a lawyer "based on the exercise of legal judgment."[96] Thus, in selecting documents used to develop the seed set, the lawyer may utilise various search techniques such as using keywords search to locate particular documents of interest, linear review of documents from specific custodians, or review of documents based on specific dates and timelines.[97] The documents selected by the lawyer are used to train the predictive coding algorithm to identify similar documents in the larger database.

In addition to the two main approaches to development of seed sets, some authors have identified the hybrid approach, which entails a combination of the two main approaches.[98] Here, the extent to which the two approaches are combined will depend on the discretion of the lawyer. Where the hybrid approach is adopted in development of seed sets, it is important to note the extent to which the two approaches are combined, taking particular note of the dominant mix. As we will see later, the approach used in developing seed sets is vital to a determination of whether privilege is applicable to the seed set.

B) The purpose and scope of litigation privilege

The primary purpose of litigation privilege (also referred to in some jurisdictions as solicitor-work product privilege) is to create a 'zone of privacy' for counsel to prepare and present their case to the court.[99] It is a vital privilege in a modern adversarial system and its objective is to create the professional space needed by counsel to prepare their case without the risk of exposing their litigation strategy to opposing counsel. In the US, this privilege was formally endorsed by the United States Supreme Court in *Hickman v Taylor*.[100] The Court in *Hickman* noted the vital importance of the privilege in affording counsel "a certain degree of privacy, free from unnecessary intrusion by opposing parties and their counsel."[101] The Supreme Court went further to observe the need for counsel to be able to

96 Facciola & Favro, *supra* note 93 at 16.

97 Ralph Losey, "Predictive Coding and The Proportionality Doctrine: A Marriage Made In Big Data" (2013) 26 Regent UL Rev 7 at 22.

98 *Ibid*; Facciola & Favro, *supra* note 93 at 13.

99 *Blank v Canada (DOJ)*, 2006 SCC 39, [2006] 2 SCR 319 [*Blank*].

100 *Hickman v Taylor*, 329 US 495, 67 S Ct 385 (SC 1947) [*Hickman*].

101 *Ibid* at 510.

"assemble information, sift what he considers to be the relevant from the irrelevant facts, prepare his legal theories and plan his strategy without undue and needless interference."[102]

While the primary purpose of discovery in litigation is to avail the parties access to information relevant to the litigation, such access is neither limitless nor without boundary. Litigation privilege tends to provide a justifiable limit to such access by preventing the disclosure of information prepared by counsel for the dominant purpose of litigation. Such limitation is necessary to enable counsel (in the preparation of their case) to gather information, dissect relevant from irrelevant information, prepare their legal theory and strategy independently rather than free riding on the effort of opposing counsel. In an adversarial system of justice, parties are not allowed to litigate their case on "wits borrowed from the adversary."[103] Hence a private space for the lawyer's preparation of client's case is vital to the litigation system. Without such space, the court in *Hickman* noted that:

> much of what is now put down in writing would remain unwritten. An attorney's thoughts, heretofore inviolate, would not be his own. Inefficiency, unfairness, and sharp practices would inevitably develop in the giving of legal advice and in the preparation of cases for trial. The effect on the legal profession would be demoralizing. And the interests of the clients and the cause of justice would be poorly served.[104]

Thus, if counsel's litigation strategy or legal theory about a case is deprived of privacy and protection, it is doubtful that counsel would want to devote the effort and resources necessary in developing one or putting it down in written form. Such situations will not be in the best interest of the legal profession and would not work in the interest of justice.

The Supreme Court of Canada adopted a similar view in *Blank v Canada* wherein the court observed that the object of litigation privilege is "to ensure the efficacy of the adversarial process."[105] To achieve this objective, the Supreme Court stated that parties to litigation "must be left to prepare their contending positions in private, without adversarial interference and without fear of premature disclosure."[106] Prior to the decision in *Blank v Canada*, Jackett P of the former Exchequer Court of Canada explained the purpose of litigation privilege (once referred to as lawyer's brief rule) as follows:

---

102 *Ibid* at 511.

103 *Ibid* at 516.

104 *Ibid* at 393–94.

105 *Blank*, *supra* note 99 at para 27.

106 *Ibid*.

> Turning to the 'lawyer's brief' rule, the reason for the rule is, obviously, that, under our adversary system of litigation, a lawyer's preparation of his client's case must not be inhibited by the possibility that the materials that he prepares can be taken out of his file and presented to the court in a manner other than that contemplated when they were prepared ... If lawyers were entitled to dip into each other's briefs by means of the discovery process, the straightforward preparation of cases for trial would develop into a most unsatisfactory travesty of our present system.[107]

Further, Sharpe has noted that the rationale for litigation privilege is based on "the need for a protected area to facilitate … preparation of a case for trial by the adversarial advocate."[108] Thus, Canadian case law and jurisprudence clearly acknowledge the fact that litigation privilege plays an important role in creating a conducive environment that facilitates the adversarial litigation process.

The scope of litigation privilege is very specific. It covers all documents prepared by counsel for the dominant purpose of litigation either during the course of litigation or when litigation was reasonably contemplated or anticipated. The Supreme Court of Canada in *Lizotte v Aviva* noted that the privilege "protects against the *compulsory disclosure* of … documents whose *dominant purpose* is preparation for litigation."[109] Adopting the Supreme Court's view on the scope of litigation privilege, it might be reasonable to state that the privilege may, in some situations, apply to documents prepared by counsel for use in training predictive-coding algorithms. This will also apply to documents prepared by counsel that might provide insight into counsel's opinion of the case or its litigation strategy.

Facciola and Favro were of the view that, since the decision of the United States Supreme Court in *Hickman*, and its subsequent codification in Rule 26(b)(3) of the Federal *Rules of Civil Procedure*, parties to litigation in the US are precluded from discovery of documents that are prepared by opposing parties in anticipation of litigation.[110] Hence all documents that have the potential to reveal a lawyer's case strategy, assessment of the strength and weaknesses of the case, as well as the party's line of proof or defence (as the case may be) is almost absolutely immune from

---

[107] *Susan Hosiery Ltd v Minister of National Revenue*, [1969] 2 Ex CR 27 at para 9, [1969] CTC 353.

[108] RJ Sharpe, "Claiming Privilege in the Discovery Process" in *Special Lectures of the Law Society of Upper Canada* (Don Mills, Ont.: Richard De Boo Publishers, 1984) 163 at 164–65.

[109] *Lizotte v Aviva Insurance Company of Canada*, 2016 SCC 52 at para 1, [2016] 2 SCR 521.

[110] Facciola & Favro, *supra* note 93 at 11.

discovery.[111] Facciola and Favro also noted that this protection will apply to counsel's selection of document where such selection might provide an indication of the lawyer's mental processes.[112]

Thus, the judicial approach to determination of litigation privilege in the United States and Canada seems to suggest that litigation privilege could apply to seed sets developed by counsel for use in training predictive-coding software in e-discovery. Going further, this paper will examine circumstances in which litigation privilege may apply to seed sets.

C) Application of litigation privilege to seed sets

To date, the use of predictive coding, or issues relating to its use in civil litigation, has yet to be comprehensively litigated before any Canadian court. Therefore, we are left with a dearth of jurisprudence in the Canadian jurisdiction to provide any clue to judicial view or opinion on the applicability of litigation privilege to seed sets. In light of this fact, it may be proper to turn attention to the US jurisdiction which seems to be developing a robust body of case law on predictive coding. Addressing the legal basis for the applicability of litigation privilege to seed sets would require a consideration of whether selection of documents from a larger population of documents by counsel for use in the preparation of a case gives rise to privilege. This issue was addressed by the United States Court of Appeals for the Third Circuit in *Sporck v Peil.*[113]

The issue in *Sporck v Peil* arose from pre-trial discovery. In response to a discovery request, the defendant produced hundreds of thousands of documents from which the plaintiff's lawyers selected about one hundred thousand for copying. Prior to deposition, the defendants' lawyers selected some documents (from the larger population of documents already produced to the plaintiffs) to prepare the witness for deposition. The documents were said to represent the lawyers' opinion on relevancy and possible legal defences. The plaintiffs sought discovery and production of all the documents that were used to prepare the witness for the deposition. The defense counsel refused to identify the specific documents arguing that the documents had already been disclosed as part of the broader discovery process, and that the group of documents used to prepare the witness was subject to attorney-work product privilege and immune from disclosure. The trial judge, while acknowledging that the documents constituted attorney-work product, ordered production of the documents

111 *Ibid* at 12.

112 *Ibid*.

113 *Sporck v Peil*, 759 F (2d) 312, 84 ALR Fed 763 (US CA 3rd Cir 1985) [*Sporck*].

on the basis that "it was not 'opinion' work product entitled to absolute protection."[114]

On appeal to the United States Court of Appeals for the Third Circuit, the appellate court ruled that the selection and compilation of documents by counsel in this case for pre-trial discovery qualifies as opinion work product and was therefore immune from discovery. The court based the rationale for its decision on a quote from *James Julian Inc v Raytheon Co*[115] where it was stated:

> In selecting and ordering a few documents out of thousands counsel could not help but reveal important aspects of his understanding of the case. Indeed, in a case such as this, involving extensive document discovery, *the process of selection and distillation is often more critical than pure legal research*. There can be no doubt that at least in the first instance the binders were entitled to protection as work product.[116]

Facciola and Favro identified four principles[117] emerging from the decision in *Sporck v Peil*. First, assertion of privilege over a group of documents selected from a larger population requires proof that the identification of the selected documents would reveal counsel's mental impression. Second, the privilege will only apply to identification of the selected documents. It does not protect the documents from disclosure as part of a broader document disclosure process. In this way, the group of documents becomes like a needle in a haystack. While counsel is not obligated to identify the needle, counsel is obligated to turn it over with the haystack. Third, a compilation of documents by counsel may not be privileged where counsel has no reasonable expectation that the mental impression likely to be revealed by the selected documents would remain private. Fourth, the court may not be inclined to protect counsel's compilation of documents where the number of documents is voluminous.[118] This is based on the assumption that the more voluminous the selected documents, the more difficult it becomes to discern the lawyer's mental process from a review of the selected documents. This fourth principle, however, must be approached with caution in an era where the volume of documents involved in typical electronic discovery (and in the creation of seed sets) is becoming bigger in size than ever.

---

114 *Ibid* at para 1.

115 *James Julian Inc v Raytheon Co*, 93 FRD 138, 33 Fed R Serv 2d 509 (D Del 1982).

116 *Ibid* at para 11 [emphasis added].

117 Facciola & Favro, *supra* note 93 at 24–26.

118 See *Disability Rights Council of Greater Washington v Washington Metropolitan Transit Authority*, 242 FRD 139 at 142–44 (DDC Dist Ct 2007).

What is evident from the discussion so far is that litigation privilege may apply to a lawyer's selection of a group of documents from a larger population of documents. This is especially the case where such selection is done in the course of preparation for litigation and where the lawyer's mental impression about the case or legal strategy could be gleaned from an examination of the selected documents.

Applying this principle to seed sets, it is argued that litigation privilege may apply to seed sets used in training a predictive coding algorithm in litigation document review. The applicability though, will be dependent on the method or technique used by counsel in creating the seed set. Earlier on, this paper examined three possible ways in which a seed set could be created. Random sampling techniques entails the non-judgmental selection of documents using computer software or automated process that require very minimal to no human intellectual exercise. The absence of any exercise of skill, judgment, and reasoning by counsel in selecting or generating documents under the random sampling techniques makes the generated seed set unqualified for litigation privilege. A group of documents generated through this process is incapable of revealing a lawyer's legal theory, litigation strategy, or the lawyer's opinion on the strength and weaknesses of the case. Applying litigation privilege to this group of documents does not in any way advance the principle or policy objective of litigation privilege.

However, a seed set generated using judgmental sampling is deserving of litigation privilege. Judgmental sampling technique involves extensive use of human intelligence to select a group of documents that meet predetermined criteria set by the lawyer. Thus, development of seed sets using this technique requires exercise of independent legal judgment by the lawyer. The documents are meticulously selected by a lawyer based on their exercise of skill, judgment, and reasoning. Evident in this method of generating seed sets are the basic elements required for a valid assertion of litigation privilege. This group of documents should be entitled to litigation privilege for many reasons. First, the documents were selected by the lawyer in the preparation of litigation. Second, the documents were meticulously selected by the lawyer and hence potentially reflect the lawyer's understanding of the case, the lawyer's opinion, and legal theory. Third, disclosure of such documents may compromise or reveal the lawyer's mental impression of the case and litigation strategy.

With regards to seed sets developed using the hybrid technique, the application of litigation privilege will depend on the extent to which the two techniques mentioned above are combined in the development of the set. Where the dominant technique utilized is the judgmental technique,

it is reasonable to argue that litigation privilege should apply. The contrary would be the case where random sampling is the dominant mix.

D) Disclosure of seed set in predictive coding litigation

As much as it has been argued in this paper that case law jurisprudence from the US jurisdiction seems to support the application of litigation privilege to seed sets used in training predictive-coding algorithms, it appears that no US case law has explicitly applied litigation privilege to predictive coding seed sets. It is important to note that there are two ways documents in seed sets may become entitled to a claim of privilege.

The first is the individual privilege that may attach to a document as a separate and distinct document, e.g. solicitor-client privilege or settlement privilege. The second (as argued in this paper) is the privilege that may apply to all the documents in a seed set as a group, i.e. litigation privilege. The latter privilege will only apply when the documents are considered as members of the sample group, i.e. seed sets. The documents in the seed set lose any applicable privilege as a group once they fall outside the group, e.g. when they are considered individually or as part of the entire document population. On the other hand, the privilege that attaches to an individual document in a seed set remains even when the document falls outside the (seed set) group. A document that is solicitor-client privileged and forms part of a seed set will continue to be solicitor-client privileged (unless waived) even when it loses the litigation privilege that it may enjoy as a member of the seed set.

Cases on predictive coding (in the US jurisdiction) that dealt with issues relating to the disclosure of seed sets have always done so independent of the concept of privilege. These cases can be conveniently grouped into two main categories. The first are cases involving disclosure of seed sets, either voluntarily by the parties or mandatorily on the order of the court. The second group is comprised of cases where the court has declined to order disclosure of a seed set against the will of the responding party. These two groups are further discussed below.

1) Cases involving disclosure of seed set

A very good example of a predictive coding case involving voluntary disclosure of a seed set by a party is *Da Silva Moore.*[119] In *Da Silva Moore*, the responding party sought the approval of the court to use predictive coding in document review. The party voluntarily agreed to provide the defendant's counsel with all the *non-privileged documents* in the seed

119 *Da Silva Moore*, *supra* note 38.

set used to train the predictive coding software.[120] The court noted the importance of 'cooperation' in seeking judicial approval to use predictive coding in document review. The court linked 'cooperation' to the party's willingness to disclose, defining the term as "strategic proactive disclosure of information."[121] An important aspect of cooperation according to the court is transparency which the court deduced from the defendants' willingness to share the "seed set" used to train the predictive coding algorithm.

In the *Da Silva* case, there was no reference to the possibility of litigation privilege applying to the seed set in question. More so, it is not evident from the facts of the case whether the seed set was generated through random or judgmental sampling. If any privilege indeed applied to the seed set as a group of documents, such privilege was expressly waived by the responding party when it agreed to provide non-privileged documents in the seed set. The defendants' agreement to provide all *non-privileged* seed set documents implies that they did not waive any privilege that may have been attached to individual documents in the seed set. Thus, documents which are solicitor-client privileged and which form part of the seed set would still be subject to such privilege and not liable to production as part of the seed set.[122] Similarly, in *Bridgestone Americas Inc. v IBM Corporation*[123] and *Federal Housing Finance Agency v JPMorgan Chase & Co.*,[124] the court approved the responding party's request to use predictive coding following their agreement to share their seed sets with the opposing parties.

2) Cases against disclosure of seed sets

Biomet II[125] stands for the proposition that courts cannot compel parties in e-discovery (against their will) to disclose their predictive coding seed set to opposing parties. In *Biomet II*, the defendants having produced all responsive documents to the plaintiffs as part of the document disclosure

---

120 *Ibid* at 192.

121 *Ibid* at 193.

122 Hence in drawing up the predictive coding protocol, counsel should be careful about the choice of words used to consent to disclosure of seed set. Broad consent to disclose all documents in seed set may actually amount to not just waiver of litigation privilege that may attach to the documents as a group, but also a waiver of solicitor-client privilege and other privileges that may attach to individual documents in the seed set.

123 *Bridgestone Americas*, *supra* note 58.

124 *Federal Housing Finance*, *supra* note 91.

125 *Re: Biomet M2a Magnum Hip Implant Products Liability Litigation*, 357 F Supp (3d) 1389 (ND Ind 2013) (Memoranda and Order), online (pdf): *United States District Court Northern District of Indiana:* <www.innd.uscourts.gov/sites/innd/files/Disclosure of docs re predictive coding ord.pdf> [*Re: Biomet* Memo].

process, the plaintiffs further requested the defendants to specifically identify the seed set they used in training the predictive coding algorithm. Biomet refused to identify the documents, insisting rather that the discoverable documents used in the seed set were part of the greater population of responsive documents already disclosed to the plaintiffs.

In refusing the plaintiffs' request, Judge Robert Miller Jr. took the position that the request reached beyond the scope of permissible discovery. First, the court noted that complying with the plaintiffs' request would entail production of irrelevant and privileged documents used to train the predictive coding algorithm. It is self-evident that parties in civil litigation have no right to discovery of irrelevant and privileged documents. Secondly, given that Biomet had produced all producible documents, the plaintiffs' request amounted to 'discovery within discovery'. According to the court, the plaintiffs' request was not about "whether a document exists or where it is, but rather how Biomet used certain documents before disclosing them."[126] Although the court agreed that Biomet's cooperation in the matter falls below the standard endorsed by the Sedona Conference, and that such conduct may affect the court's exercise of its discretion, Judge Miller Jr., noted that the court discretion does not extend to compelling Biomet to disclose its seed set. According to the Judge:

> An unexplained lack of cooperation in discovery can lead a court to question why the uncooperative party is hiding something, and such questions can affect the exercise of discretion … But I don't have any discretion in this dispute. I won't order Biomet to reveal which of the documents it has disclosed were used in the seed set, but I urge Biomet to re-think its refusal.[127]

In the *Biomet II* decision, there was no reference by Judge Miller Jr. as to whether the documents in the seed set are litigation privileged. In fact, it is clear that litigation privilege was not expressly a basis for the decision to refuse access to the documents in the seed set. However, by indicating that the plaintiffs in *Biomet II* were precluded from seeking to know "how Biomet used certain documents before disclosing them"[128], the court seems to further the underlying idea which supports the application of litigation privilege to seed sets—providing a measure of confidentiality to litigants' preparation of their case, including their selection of documents.

A reason why parties to litigation may resist disclosure of their methodological decisions, including decisions relating to choice of documents for use in training predictive coding software, was noted by

126 *Ibid* at 3.
127 *Ibid*.
128 *Ibid*.

Judge Leen in *Progressive Casualty Insurance Company v Delaney*.[129] Judge Leen stated that "methodological decisions reveal work product".[130] The Judge further noted that, in cases where the court has approved the use of predictive coding, it has often required *full disclosure* from the requesting party.[131] She noted (among others) an exception to the requirement of full disclosure, such as privilege attached to lawyer's work product (litigation privilege). Therefore in determining the ambits of "full disclosure", the court is bound to take into consideration any valid objection raised by a party.

Seed sets developed substantially using judgmental sampling should be entitled to litigation privilege. Unless such privilege is waived by the party, the seed set should be immune from discovery by the opposing party, or from a compulsory disclosure order by the court against the will of the responding party. The applicability of litigation privilege to seed sets conforms with the legal principle guiding the application of litigation privilege in adversarial litigation. However, considering the novelty of predictive coding technology in civil litigation discovery, and the immense benefit of the technology to the profession, it is necessary to state that the application of litigation privilege to seed sets, and the responding party's insistence on asserting this privilege may further the fear of this novel technology and may stifle its acceptability in civil litigation practice.[132]

That notwithstanding it would be wrong to judicially whittle down the applicability of litigation privilege to seed sets in order to make predictive coding technology acceptable. Neither should the court insist on a party's agreement to waive this privilege (and hence disclosure of the seed set) as evidence of the 'cooperation and transparency' requirement essential to the court's exercise of discretion to approve the use of the technology. Parties to litigation should agree at the initial stage of the discovery process whether to use predictive coding technology for the document review. The parties should also strive to reach an agreement on disclosure of seed sets. It is important to note that failure by a party to waive its litigation privilege over seed set does not necessarily spell doom to the use of predictive coding in the document review. The parties can agree to jointly develop the seed sets for use in training the predictive-coding algorithm. Seed sets jointly developed by parties in litigation cannot be considered to be litigation privileged by any of the parties.

---

129 *Progressive Casualty Insurance*, *supra* note 53.

130 *Ibid* at 15.

131 In *Da Silva Moore*, *supra* note 38, the disclosure of seed set was made voluntarily.

132 *Re: Biomet* Memo, *supra* note 125.

Additionally, another way to overcome the problem associated with disclosure of seed sets in predictive coding would be to avoid the use of seed sets altogether. Seed sets are only relevant in predictive coding technology based on the passive learning technique. As a result, predictive coding technology based on active learning (as opposed to passive learning) does not require the use of seed sets. In active learning, the algorithm selects and prioritises responsive documents based on its analysis of coding decisions made by the document reviewer.

7. Impact of Predictive Coding Technology in Civil Litigation

Holloway identified technological revolution as part of the "forces driving the current wave of change in the legal profession."[133] Law firms are increasingly adopting technology to support increased efficiency in the provision of legal services to clients.[134] Predictive coding technology has been described (rightly so) as "the most groundbreaking and disruptive" civil discovery technology.[135] Successful adoption of predictive coding technology in civil litigation will definitely have an impact on civil litigation practice.

The increasing number of electronic documents available for review in modern litigation adds to the cost of litigation. Predictive coding technology assists in addressing this problem through cost-efficient and speedy review of electronic documents, resulting in speedy disposition of cases before the courts. Where manual review of large document sizes could take years, review of the same documents using predictive coding technology could take a few months. Since discovery is a pre-trial issue that must be resolved before cases go to trial, predictive coding technology can dramatically reduce pre-trial wait times. Cases are expeditiously disposed resulting in a decreased backlog in the court system. Also, predictive coding technology has the potential to substantially reduce the cost of document review in litigation involving the review of large sets of documents, and in effect, substantially reducing the overall cost of litigation. Thus, predictive coding technology is a useful tool in addressing problems relating to the cost and length of litigation.

Unfortunately, knowledge of predictive coding and legal technology is not common among litigating lawyers. This gives rise to a legal professionalism issue. Rule 3.1 of the *Model Code of Professional*

---

133 Ian Holloway, "A Canadian Law School Curriculum for this Age" (2014) 51:4 Alta L Rev 787 at 797.

134 *Ibid.*

135 Thomas Davey & Michael Legg, "Predictive Coding: Machine Learning Disrupts Discovery" (2017) 32 L Soc NSW J 82 at 82.

Conduct defines a competent lawyer as "a lawyer who has and applies relevant knowledge, skills and attributes in a manner appropriate to each matter undertaken on behalf of a client."[136] In the age of new and emerging technologies such as artificial intelligence, does this rule impose an obligation on lawyers engaged in litigation involving extensive documentary discoveries to acquire competence in the use of legal technologies such as predictive coding? Also, where a lawyer's lack of knowledge of legal technology (such as predictive coding) results in substantial cost to the client, would it be reasonable for the lawyer to bill such avoidable cost to the client?[137] Would a lawyer, whose ignorance of legal technology results in substantial additional legal cost to the client, be deemed a "competent lawyer" within the context of Rule 3.1?

The need for familiarity with legal technologies and the use of the same by lawyers in litigation is quickly becoming evident in judicial pronouncements from the court. In *Cass v 1410088 Ontario Inc.*,[138] an Ontario court capped the cost award recoverable by the successful party in the litigation because counsel should have used artificial intelligence technology to significantly reduce the cost incurred in the preparation for the litigation. Similarly, in *Drummond v The Cadillac Fairview Corp Ltd*, Justice Perell noted that artificial intelligence technology is "a necessity for the contemporary practice of law" which should be anticipated and encouraged.[139] Therefore, a lawyer's ignorance of legal technology should not occasion significantly avoidable costs to the client or the opposing party in litigation. The Federation of Law Societies of Canada (FLSC) has sought to address the problem of legal technology competence by proposing an amendment to the Model Code provision relating to competence by lawyers. The proposed amendment states that "[t]o maintain the required level of competence, a lawyer should develop and maintain a facility with technology relevant to the nature and area of the lawyer's practice and responsibilities."[140] Although this proposed amendment has not been adopted by FLSC, or any of the provincial law societies in Canada,

---

136 See "Model Code of Professional Conduct" (10 March 2016), online: *Federation of Law Societies of Canada* <flsc.ca/national-initiatives/model-code-of-professional-conduct/>.

137 *Ibid*, s 3.6-1. The *Model Code of Professional Conduct* imposes an obligation on lawyers to charge fair and reasonable fees.

138 *Cass v 1410088 Ontario Inc*, 2018 ONSC 6959 at para 34, 2018 CarswellOnt 19514 (WL Can).

139 *Drummond v The Cadillac Fairview Corp Ltd*, 2018 ONSC 5350, 2018 CarswellOnt 15158 (WL Can) at para 10.

140 See "Model Code of Professional Conduct Consultation Report" (31 January 2017), online (pdf): *Federation of Law Societies of Canada* <flsc.ca/wp-content/uploads/2014/10/Consultation-Report-Draft-Model-Code-Amendments-for-web-Jan2017-FINAL.pdf>.

it goes further to highlight the need for legal technology competence by lawyers.[141]

Conclusion

This research paper has sought to provide a basic insight into artificial intelligence technology and how predictive coding, an offshoot of AI technology, is now being used in e-discovery document review in civil litigation. The use of predictive coding in the e-discovery document review process has gained and will continue to gain wide acceptance in legal jurisdictions. With the ever-increasing volume of electronic documents in the discovery process in civil litigation, it is undoubtedly clear that manual linear review and keyword search are no longer sustainable or efficient methods of document review in cases involving large documentary discovery. Thus, the future of e-discovery document review seems from all indications to be headed the way of predictive coding technology. This is not to imply though that the adoption of predictive coding in e-discovery document review would result in the demise of linear review and keyword search. On the contrary, the author is of the view that, depending on the nature of the review, predictive coding could be used in combination with the other review methods. For example, where the document size is large, predictive coding could be used to conduct a first level review of the large document set and responding documents could be subjected to further review using a keyword word search, and linear review. Also, where different levels of review are conducted for relevancy and privilege, predictive coding could be used for the relevancy review stage, while adopting linear review for the privilege stage.

This paper has also examined the applicability of litigation privilege to seed sets used in training the predictive coding algorithm. The paper takes the position that litigation privilege should apply to seed sets developed by counsel for use in training predictive coding algorithms if the seed sets were substantially developed using judgmental sampling. Such seed sets qualify as a lawyer's work product and should be immune from discovery.

One important legal issue that would need further scrutiny is the requirement by the court that parties seeking an order to use predictive coding must be cooperative and transparent. While there are many factors the courts consider in a determination of cooperation and transparency,

---

[141] In 2012, the American Bar Association (ABA) formally approved a similar rule referred to as the duty of technology competence. A majority of the state bars in the United States have adopted this rule in their jurisdiction. For a full list see Robert J Ambrogi, "37 States Have Adopted the Duty of Technology Competence" (Accessed 19 September 2019), online (blog): *LawSites* <www.lawsitesblog.com/tech-competence>.

the trend in this regard has been to equate cooperation and transparency with the willingness of a party to disclose its seed set to the opposing party. While no court has yet expressly ruled on the applicability of litigation privilege to a seed set, it is important for the court to note that seed sets may be protected under litigation privilege where the necessary conditions are met. While parties should endeavour to exhibit good faith in the discovery process, failure by a party to waive its litigation privilege should not be interpreted by the court as a lack of transparency and cooperation. Otherwise, the court may indeed be lifting the very immunity that has historically been held to be fundamental to our adversarial litigation process.