PROTOCOL REGARDING COMPUTER ASSISTED REVIEW

Pursuant to the Parties' agreement, the following protocol shall govern the use of predictive coding to identify potentially relevant electronically stored information (ESI):

I. DEFINITIONS

A. Predictive Coding. A review strategy (also known as "Technology Assisted Review" or "TAR") that involves technology that enables a computer platform to automatically predict how documents should be classified based on a limited, but significant, level of human input. To be more specific, a small but statistically reliable subset of all potentially relevant documents is examined manually, and the documents are classified as responsive or nonresponsive. The algorithms and heuristics of the predictive coding platform extrapolate the human coding decisions made on the small subset to the remaining documents. The decisions of the platform are then verified using a sampling protocol that evaluates the result of the computer classification.

B. Predictive Coding Set. The documents identified to be subjected to Predictive Coding process.

C. Final Review Set. The documents identified as responsive by the Predictive Coding process (exclusive of any of their non-responsive family members).

D. Null Set. The documents identified by the Predictive Coding process to be excluded from review and production.

E. Confidence Level. The statistical reliability of a result. For example, "95% Confidence" means that if one were to draw 100 independent Random Samples of the same size, and compute the Confidence Interval from each Sample, about 95 of the 100 Confidence Intervals would contain the true value.

F. Confidence Interval. The statistical error rate of a measured Confidence Level.

G. Recall. A measure of how successful the Predictive Coding process has been in finding responsive documents out of the total number of estimated responsive documents believed to exist in the Predictive Coding Set.

H. Precision. A measure of how successful the Predictive Coding process has been in avoiding the retrieval of non-responsive documents in the Predictive Coding Set. It is a measurement of the number of true responsive documents identified in the Final Review Set.

I. Estimated Richness. The estimated total number of responsive documents in the Predictive Coding Set.

II. GOAL

The goal for use of predictive coding is to achieve the just, speedy and inexpensive resolution of the discovery pursuant to Rule 1 of the Federal Rules of Civil Procedure. The parties agree that, no matter what method is used to identify potentially relevant documents, the process should be reasonable and proportionate, and that perfection is neither required nor possible to achieve. The parties agree that predictive coding should be held to the same standard as keyword searching with respect to cooperation and engaging in a reasonable and proportionate process.

III. PREDICTIVE CODING PROCESS

The parties agree to the Predictive Coding process as follows:

A. Identification of Predictive Coding Set

A party using this predictive coding process will identify the criteria used to establish the Predictive Coding Set. The criteria consists of the following:

1. **Custodians or Sources to Be Searched**. [The producing party] will disclose the custodians or sources of data that will be subjected to the Predictive Coding process.
2. **Date Range Restrictions**. Within the custodian and sources selected, [the producing party] will disclose if it is restricting the population by using any date filters.
3. **Search Terms**. [The producing party] will disclose any search terms it plans on using to further cull documents from the set it will subject to Predictive Coding process. Within 10 days of disclosure, [opposition] will provide any reasonable and proportionate modifications to the terms. Within 10 days after receiving any proposed modifications, the parties will meet and confer to attempt to resolve any disagreements. If disagreements cannot be resolved within 30 days after the initial disclosure of terms, the parties agree to stop discussions and seek resolution by the court. The parties understand that the search terms should be developed in good faith such that they will attempt to avoid exclusion of potentially relevant documents from the Predictive Coding process. The parties agree that if there is a change to the discovery process such that it would require search terms (in lieu of Predictive Coding) to be run to identify the Final Review Set, then the parties are not bound by the search terms used to identify the Predictive Coding Set, as they may require further revision to achieve a reasonable and proportionate result for review. The parties further understand and

agree that the use of Predictive Coding in conjunction with search terms may result in some search term hits being identified as Non-Relevant by the Predictive Coding process, and that [The producing party] is under no obligation to review all such search hits in addition to the Final Review Set.

4. **Global De-duplication Across Custodians / Sources**. Global de-duplication across all custodians and sources will be applied to the set of documents to be used for Predictive Coding.

5. **File Exclusions**. Certain files may be excluded from the Predictive Coding process because they would be bad candidates to train the system. Any user-created files that are excluded from the Predictive Coding process but are otherwise identified as potentially relevant will be reviewed for responsiveness. A party is not required to review excluded system files.

B. Disclosure of Predictive Coding Technology To Use

[The producing party] will use [insert name of engine] predictive coding engine.

C. Use of Subject Matter Experts to Train the Machine

[The producing party] will use one or more attorneys who are highly familiar with the claims and defenses in the matter to engage in the Predictive Coding process (the “expert(s)”).

D. Description of Responsive Review Criteria During Predictive Coding

[The producing party] will draft a document detailing what it will consider to be responsive information to be used as a guide in coding documents for relevancy during the Predictive Coding process, and provide that document to [opposition]. [Opposition] will have 7 days to request modifications to the guide. After providing any requested modifications within the 7 day period, the parties will meet and confer within 7 days to resolve the issue. If the issue is not resolved after 21 days, the parties will seek final court determination of what information will be considered responsive during the Predictive Coding process.

E. Description of the Predictive Coding Training Process To Follow

The general process of using [insert name of engine] to conduct predictive coding is described below. [The producing party] will use [insert name of engine] "out-of-the-box" default percentages for Confidence Level (95%) and Confidence Interval (5%) during this process. If at any point during the process [The producing party] identifies potential issues, [The producing party] will communicate with [engine provider] on those potential issues to seek guidance on resolution, and then meet and confer with [opposition] on the potential issues and proposed resolutions.

1. Preparation

The first step in the process is the "Preparation" stage. The Predictive Coding Set is imported into the [insert name of engine] and set up with the appropriate review issue tag. With respect to the issue tag used, the expert(s) will review only for relevancy of each document presented throughout this entire process. The reviewer will have three choices: Relevant, Not Relevant, and Skipped. Documents will only be coded Skipped if the expert(s) could not tell what a document was about, or if they ultimately could not otherwise make a relevancy determination on the document.

2. Assessment

The second step in the workflow is the "Assessment" stage. Assessment is the first of two stages (the other is the Training stage) in which expert review of the Predictive Coding Set documents occurs. Assessment begins with the creation of the "control set", a randomly selected set of 500 documents used to evaluate system performance during later stages, which is reviewed by the expert(s). For any matter, the "classifier" – the system component that calculates the relevance score for documents in the Predictive Coding Set – is created by analysis of the training documents. The performance of the classifier is measured against control set documents. These control documents represent the "gold standard" against which the progress of training is monitored, the performance of the classifier is checked, and results are quantified. In terms of statistics generated, the system uses the control documents to estimate the richness of the collection, and (in later stages) the recall and precision achieved by the classifier.

3. Training

After Assessment is complete, the expert(s) will proceed to a third, iterative "Training" stage. Training begins with the system selection of an initial training sample of 40 documents that are presented to the expert(s) for coding. Although this is considered an initial "seed set", [insert name of engine provider] recommends that this initial training sample be generated randomly by the machine. Thus, no seeding of specific examples of relevant or non-relevant documents will take place at this time. However, the parties may meet and confer if it is determined that seeding may be applicable at a later date.

Once the first sample is coded, the system begins an iterative process of selecting additional sample sets of 40 documents for expert(s) review. While the first training sample was a random sample, all subsequent sample sets are selected by the machine using an Active Learning approach. Under Active Learning, each training sample is selected based on what has been learned from the expert(s)'s coding of previous samples. In selecting sample documents, the system's objective is to maximize the sample's contribution to the training process—in other words, to choose a sample that will teach the system as much as possible about the population of documents. Based on this criterion, the system selects samples that provide comprehensive coverage of the population (reducing under-inclusiveness), while also fine-tuning and nuancing the concept of relevance that the classifier has developed to date (reducing over-inclusiveness). This multi-layered approach ensures that the classifier is exposed to a cross-section of relevant documents that is as broad as possible, and in so doing, broadens its concept of relevance to capture more of the relevant documents, while, on the other hand, refining the classifier to eliminate false positives.

During the training process, the system issues alerts if an expert provides inconsistent relevancy input. For example, if an expert marks one training document as Not Relevant and a near-duplicate as Relevant, the system will ask the expert to verify these tags. After each round of 40 documents are completed, the system will calculate a training status. Three training status states are possible: (i) not stable, (ii) nearly stable or (iii) stable. Until stability is reached, the experts continue on to the next sample.

Once the system is stable, [the producing party] will initiate the calculation of relevance scores for the remainder of the Predictive Coding Set population. Each document receives a relevance score in the range of 0 through 100, with higher scores indicating a greater degree of relevance. A relevance score of 0 indicates a document was designated by the expert(s) during assessment or training as Not Relevant. A score of 100 indicates that the experts have designated the document as Relevant. All other scores between these two extremes are system generated.

4. Decision

The Decision stage is the point at which a determination is made as to the minimum relevance score (the "cut-off") used to designate inclusion or exclusion of documents in the set of documents selected for review (the "Final Review Set"). [The producing party] will set a cut-off relevance score that it believes is reasonable and proportionate to the needs of this matter.[1] That relevance cut-off score corresponds to a specific percentage of the population that will be subject to review, or excluded. Documents with relevance scores below the cut-off mark (the Null Set) are presumed non-responsive and will not be subject to discovery.[2] Documents with relevance scores at or above the cut-off mark (the Final Review Set) will be subject to human review for verification of responsiveness and potential production.

5. Verification / Quality Control

The final phase in the workflow is the "Verification" stage. The [insert name of engine] system has a built-in process called "Test-the-Rest" that is used to validate the training. Test-the-Rest involves [insert name of engine] generating a random sample of documents from the proposed Null Set to measure the estimated prevalence of relevant documents in the Null Set. The purpose of this phase is to verify that the Null Set contains a low prevalence of relevant documents and that the reasonable and proportionate assumptions underlying the cut-off decision are valid.

[Insert name of engine.] will generate a random sample based upon a worst-case-scenario assumed 50% richness, a 95% Confidence Level, and a __% Confidence Interval. The expert(s)

[1] As provided in Section III.F. below, this information will be disclosed to [opposing party].

[2] Please note, a sample of the set of documents with relevance scores below the cut-off would be subjected to review as a quality control measure. In addition, documents with relevance scores below the cut-off may ultimately be a part of the Final Review Set if they are family members of a document that has a relevance score above the cut-off.

will review the sample, and the results will be analyzed to confirm the training was successful and the desired level of recall will be met. Any Responsive, not-privileged documents identified as a result of the Test-the-Rest process will be produced.

F. Disclosure of the Predictive Coding Process Results

[The producing party] will disclose the following information after stability of training has been reached, or if stability cannot be reached, at a point in time where training should stop, as any additional training of the machine would be considered unreasonable and not proportionate to the needs of the case:

1. The total number of documents in the Predictive Coding Set;
2. The total number of documents reviewed as part of the Assessment Phase, with a breakdown of the total number of Responsive and Non-Responsive documents identified;
3. The estimated richness of the Predictive Coding Set after the Assessment Phase, and the final estimated richness after training is complete, with corresponding Confidence Intervals;
4. The total number of documents reviewed as part of the Training Phase, with a breakdown of the total number of Responsive and Non-Responsive documents identified;
5. The estimated recall to be achieved for the Final Review Set, with corresponding Confidence Interval;
6. The total number of documents in the Final Review Set (not family complete);
7. The total number of documents in the Null Set (not family complete);

8. The total number of documents reviewed as a part of the Test-the-Rest process, with a breakdown of the total number of Responsive and Non-Responsive documents identified;
9. The estimated richness of the Null Set; and
10. The identification in []'s production of the Responsive, non-privileged documents that were used in the Assessment and Training of the Predictive Coding Set.

IV. REVIEW OF FINAL REVIEW SET

As a default, [the producing party] intends to use additional human review on the Final Review Set, as not all documents identified by the Predictive Coding machine will be responsive to the scope of discovery in the matter. However, if, based upon the interests of time and cost, [the producing party] determines to not engage in the human review of any or all of the Final Review Set, the parties agree that [the producing party] may bulk designate all documents as Confidential, and [the producing party] will not be penalized or otherwise found in violation of any agreements, stipulations or orders relating to confidentiality designations of documents for this action. Any disputes regarding confidentiality designations of the documents produced as such will be resolved on a document-by-document basis.