Solving the ‘Transparency’ Issue in Technology Assisted Review

Continuous Active Learning Eliminates Concerns about Disclosure

John Tredennick, Esq.
Founder and CEO, Catalyst Repository Systems
Technology Assisted Review (TAR) Has a Transparency Problem

Notwithstanding TAR’s proven savings in both time and review costs, many attorneys hesitate to use it because courts require “transparency” in the TAR process.

Specifically, when courts approve requests to use TAR, they often set the condition that counsel disclose the TAR process they used and which documents they used for training. In some cases, the courts have gone so far as to allow opposing counsel to kibitz during the training process itself.

Attorneys fear that this kind of transparency will force them to reveal work product, thoughts about problem documents, or even case strategy. Although most attorneys accept the requirement to share keyword searches as a condition of using them, disclosing their TAR training documents in conjunction with a production seems a step too far.

Thus, instead of using TAR with its concomitant savings, they stick to keyword searches and linear review. For fear of disclosure, the better technology sits idle and its benefits are lost.

The new generation of TAR engines (TAR 2.0), particularly the continuous active learning protocol (CAL), however, enables you to avoid the transparency issue altogether.

Transparency: A TAR 1.0 Problem

Putting aside the issue of whether attorneys are justified in their reluctance, which seems fruitless to debate, consider why this form of transparency is required by the courts.

Limitations of Early TAR. The simple answer is that it arose as an outgrowth of early TAR protocols (TAR 1.0), which used one-time training against a limited set of reference documents (the “control set”). The argument seemed to be that every training call (plus or minus) had a disproportionate impact on the algorithm in that training mistakes could be amplified when the algorithm ran against the full document set. That fear, whether founded or not, led courts to conclude that opposing counsel should be able to oversee the training to ensure costly mistakes weren’t made.
Solutions in New TAR. This is not an issue for TAR 2.0, which eschews the one-time training limits of early systems in favor of a continuous active learning protocol (CAL) that continues through to the end of the review. This approach minimizes the impact of reviewer “mistakes” because the rankings are based on tens and sometimes hundreds of thousands of judgments, rather than just a few.

CAL also puts to rest the importance of initial training seeds because training continues throughout the process. The training seeds are in fact all of the relevant documents produced at the end of the process. At most you can debate about the documents that are not produced, whether reviewed or simply discarded as likely irrelevant, but that is a different debate.


Discussing the broader issue of transparency with respect to the training sets used in TAR, Judge Peck observed that CAL minimizes the need for transparency.

“If the TAR methodology uses “continuous active learning” (CAL) [as opposed to simple passive learning (SPL) or simple active learning (SAL)], the contents of the seed set is much less significant.”

Let’s see how this works, comparing a typical TAR 1.0 to a TAR 2.0 process.

TAR 1.0: One-Time Training Against a Control Set
A Typical TAR 1.0 Review Process. As depicted in the immediately preceding illustration, a TAR 1.0 review is built around the following steps:

1. A subject matter expert (SME), often a senior lawyer, reviews and tags a sample of randomly selected documents (usually about 500) to use as a “control set” for training.

2. The SME then begins a training process often starting with a seed set based on hot documents found through keyword searches.

3. The TAR engine uses these judgments to build a classification/ranking algorithm that will find other relevant documents. It tests the algorithm against the 500-document control set to gauge its accuracy.

4. Depending on the testing results, the SME may be asked to do more training to help improve the classification/ranking algorithm. This may be through a review of random or computer-selected documents.

5. This training and testing process continues until the classifier is “stable.” That means its search algorithm is no longer getting better at identifying relevant documents in the control set.

Once the classifier has stabilized, training is complete. At that point your TAR engine has learned as much about the control set as it can. The next step is to turn it loose to rank the larger document population (which can take hours to complete) and then divide the documents into categories to review or not.

Importance of Control Set. In such a process, you can see why emphasis is placed on the development of the 500-document control set and the subsequent training. The control set documents are meant to represent a much larger set of review documents, with every document in the control set standing in for what may be a thousand review documents. If one of the control documents is improperly tagged, the algorithm might suffer as a result.
Of equal importance is how the subject matter expert tags the training documents. If the training is based on only a few thousand documents, every decision could have an important impact on the outcome. Tag the documents improperly and you might end up reviewing a lot of highly ranked irrelevant documents or missing a lot of lowly ranked relevant ones.

**TAR 2.0 Continuous Learning; Ranking All Documents Every Time**

TAR 2.0 systems don’t use a control set for training. Rather, they rank all of the review documents every time on a continuous basis. Modern TAR engines don’t require hours or days to rank a million documents. They can do it in minutes, which is what fuels the continuous learning process.

A TAR 2.0 engine can continuously integrate new judgments by the review team into the analysis as their work progresses. This allows the review team to do the training rather than depending on an SME for this purpose.

It also means training is based on tens or hundreds of thousands of documents, rather than rely on a few thousand seen by the expert before review begins.

**TAR 2.0: Continuous Active Learning Model**

Modern TAR engines **don’t require hours or days** to rank a million documents. **They can do it in minutes**, which is what fuels the continuous learning process.

As the infographic demonstrates, the CAL process is easy to understand and simple to administer. In effect, the reviewers
become the trainers and the trainers become reviewers. Training is review, we say. And review is training.

**CAL Steps**

1. Start with as many seed documents as you can find. These are primarily relevant documents which can be used to start training the system. This is an optional step to get the ranking started but is not required. It may involve as few as a single relevant document or many thousands of them.

2. Let the algorithm rank the documents based on your initial seeds.

3. Start reviewing documents based in part on the initial ranking.

4. As the reviewers tag and release their batches, the algorithm continually takes into account their feedback. With increasing numbers of training seeds, the system gets smarter and feeds more relevant documents to the review team.

5. Ultimately, the team runs out of relevant documents and stops the process. We confirm that most of the relevant documents have been found through a systematic sample of the entire population. This allows us to create a yield curve as well so you can see how effective the system was in ranking the documents.

With CAL the entire process is transparent. There is no worry that the control set has been mis-tagged because there is no control set.

**What Does This Have to Do With Transparency?**

In a TAR 2.0 process, transparency is no longer a problem because training is integrated with review. If opposing counsel asks to see the "seed" set, the answer is simple: “You already have it.”

Every document tagged as relevant is a seed in a continuous review process. And every relevant non-privileged document will be or has been produced.
Likewise, there is no basis to look over the expert’s shoulder during the initial training because there is no expert doing the training. Rather, the review team does the training and continues training until the process is complete. Did they mis-tag the training documents? Take a look and see for yourself.

This eliminates the concern that you will disclose work product through transparency. With CAL, all of the relevant documents are produced, as they must be with no undue emphasis placed on an initial training set. The documents, both good and bad, are there in the production for opposing counsel to see. But work product or individual judgments by the producing party are hidden. Voila, the transparency problem goes away.

Postscript: What About Other Disclosure Issues?

Discard Pile Issues. In addressing the transparency concern, I don’t mean to suggest there are no other disclosure issues to be addressed. For example, with any TAR process that uses a review cutoff, there is always concern about the discard pile. How does one prove that the discard pile can be ignored as not having a significant number of relevant documents?

That is an important issue, one that I discuss in these two articles:

- Measuring Recall in E-Discovery Review, Part One: A Tougher Problem Than You Might Realize
- Measuring Recall in E-Discovery Review, Part Two: No Easy Answers

The problem doesn't go away with TAR 2.0 and CAL, but it is the same issue advocates have to address in a TAR 1.0 process as well. This requires sampling as I explain in my two measuring recall articles referenced above.

Tagging for Non-Relevance; Collection. What about documents tagged as non-relevant by the review team? How do I know that was done properly?

This too is a separate issue that exists in both TAR 1.0 and 2.0 processes. Indeed, it also exists with linear review.
And, last but not least, did I properly collect the documents submitted to the TAR process? Good question but, again, a question that applies whether you use TAR or not. Unless you collect the right documents, no process will be reliable.

Again, these are issues to be addressed in a different article. They exist in whichever TAR process you choose to use. My point here is that with a TAR 2.0 process, a number of the “transparency” issues that bug people and have hindered the use of this amazing process simply go away.

**About the Author**

**John Tredennick** is the founder and CEO of Catalyst. John is passionate about the role of search in e-discovery. Before founding Catalyst in 2000, he was a trial lawyer and litigation partner. John has been a frequent speaker on legal and technology issues for more than 30 years. He’s also written and edited five best-selling books and countless articles on litigation and technology issues.