Factors Affecting the Performance of Reviewers in a Large-Scale Technology-Assisted Review Project

Andrew Harbison,[b] Maura R. Grossman,[a1] Gordon V. Cormack,[a] Bronagh McManus,[b] and Tom O'Halloran[b]

[a] *University of Waterloo, Waterloo, Ontario, Canada*
[b] *Information Retrieval Services, Grant Thornton Ireland, Dublin, Ireland.*

Abstract

As part of a programme of research arising from a major litigation in the Irish Courts, the authors were able to make observations as to the behaviour of a group of trained, experienced, professional barristers carrying out a series of reviews under real-world conditions. From these observations, we were able to discern several anomalous behaviour patterns which, if not identified and controlled, might have had a significant effect on the outcomes of the reviews. These behaviours were, as far as could be seen, not any fault of the reviewers, nor were they due to the specific conditions of the review, and therefore, would appear to be likely to occur in any similar review process. We describe the behaviours seen during the research programme, propose reasons why these behaviours may have arisen, and propose measures by which they can be controlled-for in other similar projects.

Keywords

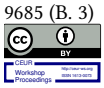
Information Retrieval, electronic discovery, technology-assisted review, TAR, Continuous Active Learning, CAL, recall, elusion, CEUR-WS

1. Introduction

While considerable work has been done in recent decades on measuring the performance of technology-assisted information-retrieval techniques and technologies (for example [1],[2] & [3] etc.), less attention has been paid to the performance of perhaps the most important aspect of any information-retrieval exercise: the reviewers relied upon to "train" the information-retrieval models in the first place. Some work was done on this topic early last decade, most notably by Grossman & Cormack [4] and [5], and Roitblat et al. [6], but in recent years there has been little attention paid to this issue. The authors recently addressed the problem of assessing reviewer influence on the performance of information-retrieval systems in another publication [7] but, in general, it seems the discipline still tends to rely on the assumption that the human review component of technology-assisted reviews is invariably accurate and that relevance can always be assessed according to some "gold standard" of truth, when it has been long understood that neither assumption can be relied upon [8].

Between 2017 and 2022, the authors carried out a complex, large-scale electronic discovery project comprising over 300 million unique documents drawn from the information systems and archives of a large, defunct insurance company, based in Ireland. The electronic discovery was carried out as part of legal proceedings taken by the insurance company's legal administrators against the company's former auditors for negligence. As part of the proceedings, the Defendant demanded extraordinarily broad discovery, covering over 15 years of documents, across an array of information systems, most of which had either been archived or retired. The project also entailed the cataloging and retrieval of data from over 1,000 backup tapes. The Plaintiff's claim was of the order of $1 billion.

The electronic discovery project posed a significant number of challenges which could only be effectively met using modern information-retrieval technology. For example, the Defendant required

[1] *3rd Workshop on Augmented Intelligence for Technology-Assisted Review Systems (ALTARS 2024): Evaluation Metrics and Protocols for eDiscovery and Systematic Review Systems. March 28, 2024, Glasgow, UK.*
EMAIL: gvcormac@uwaterloo.ca (A. 1); maura.grossman@uwaterloo.ca (A. 2); AJH_Work@hotmail.com (B. 1); bronagh.mcmanus@ie.gt.com (B. 2); ohalloranthomas1@gmail.com (B. 3)
ORCID: 0000-0002-5890-0293 (A. 1); 0000-0003-2279-4262 (A. 2); 0009-0006-7562-5306 (B. 1); 0009-0009-4535-7584 (B. 2); 0009-0007-3169-9685 (B. 3)


the Plaintiff to produce documents under 74 separate discovery criteria (Requests for Production, or RFPs) of which 70 required relevance assessment. Complicating matters further, Irish discovery rules require that documents be identified along with the specific discovery criteria (i.e., the RFP) to which they are responsive.

To respond to the challenges of this project, the administrators decided to employ Continuous Active Learning® (CAL®) tools developed by Maura R. Grossman and Gordon V. Cormack, as it was found that commercially available technology-assisted review (TAR) tools were unlikely to be able to cope with the review burden without the unreasonable expenditure of time and resources. The Grossman and Cormack CAL tools had to be validated, however, to ensure that they would be accepted as being of a standard equivalent to the leading commercially available tools, which would render them acceptable for use in an Irish Court. Accordingly, a multifaceted testing program was run between January 2021 and September 2022 to evaluate the Grossman and Cormack CAL® method against that of a leading commercial TAR tool and also to evaluate the logistics necessary for such an exceptionally large, complex document review. This testing program inevitably led to detailed assessment of reviewer behaviour and performance in the context of the two TAR systems under examination.

The testing programme was unique in that it was carried out on a huge document corpus in a high-value, real-world litigation. The review criteria were determined by the counterparties in the litigation and the research team was not involved in their content or design. The reviewer group was made up of highly qualified and trained barristers all of whom had previous experience in large-scale document review projects and all of whom were paid commercial rates for their work. The review was supported by professional litigators from an international law firm (the Maples Group). The testing was well funded because of its importance to the overall litigation. Our testing budget was of the order of €3 million. Finally, an unusually heavy focus was placed on quality assurance, data collection, and performance measurement in the conduct of the review. It was essential for the purposes of the case that the results of the testing be highly defensible.

This paper summarises the findings of these assessments of reviewer performance. It does not delve in detail into specific aspects of reviewer activity but instead identifies different instances of anomalous behaviour or performance observed and measured during the broader testing and attempts to explain why these behaviour patterns arose. Any of the observed behaviors would be worthy of deeper analysis and some have, indeed, been examined in more detail already by the authors in other publications e.g., [9],[10].

2. Initial Steps

It was understood in advance that reviewers were unlikely to be able to cope with 70 RFPs simultaneously [9], so the 70 RFPs were repartitioned into 10 broader "composite categories" (CCs), each covering a subset of the 70. Our plan was to have the reviewers carry out 20 separate CAL reviews – one for each for the CCs on both the Grossman and Cormack CAL® system and on the leading commercial TAR tool. It was then intended that the reviewers carry out sub-reviews for documents found responsive to each of the CCs to assign them to the specific RFPs contained within the CC.

In carrying out reviews of individual CCs, reviewers were organised in pairs, or groups of four (in two pairs). Reviewers were issued documents for review in batches, usually 100 or 200 documents in size. Reviewers were also assigned an additional 10% of the documents also assigned to their partner reviewer for quality assurance purposes. This "crossover" set of documents allowed us to quickly detect when reviewer pairs were diverging in their assessment of documents. We attempted, as far as practical, to ensure that reviewers in pairs carried out their reviews in parallel, reviewing similar numbers of documents per day as one another. Reviewer pairs were briefed on the same material prior to each review, by the same briefing team, at the same time.

Tests were performed to see if this was a practical approach to the problem. Testing covered multiple different streams, which will be discussed below. However, in general it was found that while both technologies worked adequately, the Grossman and Cormack CAL® system significantly outperforming the commercial TAR tool. And while the reviewers, all trained and experienced barristers, were intelligent and capable, the review of the documents continually posed problems in

maintaining reviewer performance, consistency, and accuracy across both test systems. In essence, the reviewers and the technologies often failed to "gel".

3. Fundamental Problems in Manual Review of Documents in Litigation.

Order 31 Rule 12 of the Rules of the Superior Courts in Ireland, as enacted in Statutory Instrument 93 of 2990, requires that:

> *"Any party may apply to the Court by way of notice of motion for an order directing any other party to any cause or matter to make discovery on oath of the documents which are or have been in his possession, power or procurement relating to any matter in question therein. Every such notice of motion shall specify the precise categories of documents in respect of which discovery is sought and shall be grounded upon the affidavit of the party seeking such an order of discovery..."*[11]

This Order requires that documents be produced subject to the ill-defined criterion that they be "related to any matter in question." This criterion is supposedly compensated for by the stipulation that the categories of documents requiring discovery be precisely specified. However, in practice, categories are often defined in broad or even ambiguous terms.

This practice is not unique to the Irish jurisdiction. Discovery under the U.S. Federal Rules of Civil Procedure, the U.K. Civil Procedure Rules (specifically Practice Direction 31A [12]) and in other Common Law jurisdictions reflect similar ambiguity about the concept of document relevance. All assume that reviewers are always capable of correctly discerning relevant from non-relevant documents, even though reviews are typically conducted in circumstances where the documents themselves are ambiguous, the criteria against which they are being reviewed are imperfectly defined, and where the reviewers' own knowledge of the specific matters under contention is limited. The infallibility of reviewers has long been disproven in both the general information-retrieval literature and in research directly related to electronic discovery processes [7],[13],[14],[15]. Different reviewers working on the same document sets rarely achieve positive agreement of more than 70% in their assessment of document relevance. [5],[6],[16]

Reviewer assessments are also affected by the conditions of the review, such as the prior expectations of reviewers and the density or "richness" of responsive documents in the review set. Considerable emphasis is placed on the concept of Recall (i.e., the proportion of relevant documents in a document set returned by a specific information-retrieval process) even though, in the almost inevitable absence of a reliable "gold standard" of relevance, such a concept is of limited value. Recall is of some worth in comparing the relative performance of different systems across the same data set, but as a measure of absolute review performance it is considerably flawed. The Recall value alone (even based on independent blind assessments) tells us little and should not be used as an absolute acceptance standard in and of itself [7].

The assessment of electronic discovery reviews is beset with miscalculations of Recall by combining estimates taken from different samples, assessed by different reviewers, under different conditions. What is typically seen in modern litigation reviews is Recall measured either by:

- comparing the number of documents coded responsive during the review to the number estimated in advance from a random sample of the entire collection, and stopping when the former is 70% of the latter, or
- comparing the number of documents coded responsive during the review to the number estimated from a random sample of the as-yet-unreviewed documents (dubbed "Elusion" by Roitblat[1].) Recall is estimated to be the former divided by the sum.

Both these methods are flawed. The former because, all other issues aside, the basis for determining Recall is usually an assessment of the "richness" of the document collection based review of a modest sample of documents before the review begins. This makes any decision on the end-point of the review largely arbitrary (per Goodhart's Law). The latter method fails because typically the reviewers reviewing the "Elusion" sample are incentivised to mark as few documents relevant as

possible to maximise the Recall result, and therefore (consciously or unconsciously) assess the sample according to much stricter criteria than those used in the document review proper. The consequence of this is that the results of the document review proper and that of the Elusion test are likely to be barely related.

4. Examples of Observed Anomalous Reviewer Behaviour

As stated above, this paper is not intended to provide quantitative evidence of specific anomalous behaviours observed about reviewers working on the litigation, but instead to set out a typology of those behaviours, to provide real-life examples of each and, as far as is practical, to propose possible underlying causes for these behaviours. In general terms, we observed the following anomalous behaviours on multiple occasions throughout the 18 month testing programme:

- Substantial differences between reviewers' assessments of similar document sets as reflected in the proportion of documents found relevant in specific document review batches. In some cases, individual reviewers would continue to find large proportions of the documents provided to them relevant when other reviewers working on similar batches of documents, exported from the TAR system at the same stage of the review process, were finding much fewer. One would expect that, in general, document batches exported from any TAR system at a given point in a review would tend to have the same or similar proportions of relevant and non-relevant documents – after all, at a given point in any TAR process, a certain proportion of the relevant documents in the system would already have been found, and a certain proportion would remain to be found. While this is what was usually observed in reviews, it often was not.
- Significant disagreements between reviewers on the relevance of documents, and also disagreements between the reviewers and other team members assigned to provide quality assurance on the review. Indeed, often paired reviewers would agree with one another more that with the quality assurance (QA) team, raising the question as to whether there is much benefit in doing review QA at all where such a pairing approach is employed.
- The extent to which reviewers disagreed as to the relevance of documents was also far greater than might have been expected. We had assumed that, where one reviewer found a document relevant and another not relevant, that the difference would typically be minor – a question of context and interpretation. Instead, we often found that the disagreements were fundamental, with reviewers holding strongly opposed views as to the relevance of certain documents.
- Calculation of Recall using "Elusion testing" methods was found to be highly unreliable. Other validation techniques relying on confusion-matrix testing (see below), while still flawed, were nevertheless considerably more reliable.
- The review platforms themselves and the way the different TAR methodologies were put into practice appeared to have a substantial bearing on the results of the reviews.
- Categorisation of documents proved to be of very limited value – a finding of considerable relevance to Irish legal cases where (as noted above), categorisation of documents is mandatory in electronic discovery. These findings have been discussed in a separate publication [9] and, unfortunately, there is not space to replicate them here.

We will now discuss each of these issues in more detail.

4.1. Different Reviewer Assessments of Similar / Identical Documents

Despite the QA measures in place (as described above in Section 2), it was regularly observed that, after some time, individual reviewers would develop very different views of what constituted

relevance than their review partner (or group.) This either manifested as the reviewer being far more conservative than their partner(s) about what they considered a relevant or far more open.

It was observed that in CAL-based TAR reviews, when most reviewers began to "run-out" of relevant documents as the CAL process reached its end-point, certain reviewers would go on marking documents relevant at much the same rate as they had before. For example, in the four-person review of composite category 5 (ref. Table 1), despite the four reviewers reviewing at approximately similar rates, a single reviewer, J, developed a much broader interpretation of what constituted a relevant document than his three partners. He therefore continued finding "relevant" documents in the data set long after the others had finished.

Table 1

Percentage of relevant documents found in each document batch issued to each reviewer.
Composite Category #5, 200-500 document batches.

Review	Bat27	28	29	30	31	32	33	34	35	36	37	38	39
A	92%	96%	30%	25%	14%	18%	4%	7%	3%				
J	100%	100%	97%	100%	99%	100%	99%	98%	100%	100%	91%	86%	97%
D	92%	73%	83%	76%	69%	53%	62%	62%	59%	37%	34%	24%	15%
L	82%	91%	86%	42%	46%	22%	41%	32%	25%	15%	10%	10%	3%

Review	Bat40	41	42	43	44	45	46	47	48	49	50	51	52
A													
J	94%	90%	87%	88%	86%	86%	86%	86%	63%	53%	27%	11%	10%
D	14%	8%	3%										
L													

Similarly in a two-person review (ref. Table 2), this time of a different composite category with different reviewers, Reviewer L began "running out" of documents long before Reviewer A, despite both reviewers reviewing broadly in parallel. It was found, again, that A had developed a much broader definition of relevance than reviewer L.

Table 2

Percentage of relevant documents found in each document batch issued to each reviewer.
Composite Category #9, 100-200 document batches.

Review	Bat9	10	11	12	13	14	15	16	17	18	19	20	21
A	91%	80%	62%	33%	33%	40%	75%	26%	23%	9%	14%	9%	7%
L	47%	33%	24%	7%	0%								

The same behaviour was observed on several other occasions. It did not seem to be restricted to particular reviewers, review tools, or CCs (review criteria). Instead, a single reviewer in a single review under a specific CC would unilaterally develop their own view of what constituted relevance, often based on their understanding of documents previously reviewed or because they had learned (usually incorrectly) to correlate certain vocabulary or phraseology in documents with relevance. They would then use this altered version of the relevance criterion to continue the review, despite the fact that the correct review criterion remained readily available to them in their briefing notes.

This tendency was also visible when the progress of reviews was plotted on a graph. For example, in Figure 1, it can be observed that in a review of CC#3, certain reviewers perceived a downward trend in the number of relevant documents being seen well before others did. Reviewer A reviewed a batch on 13 December with only 30% of documents deemed relevant whereas Reviewer C was still finding more than 30% in batches from the same CAL process almost a month later, on 10 January.

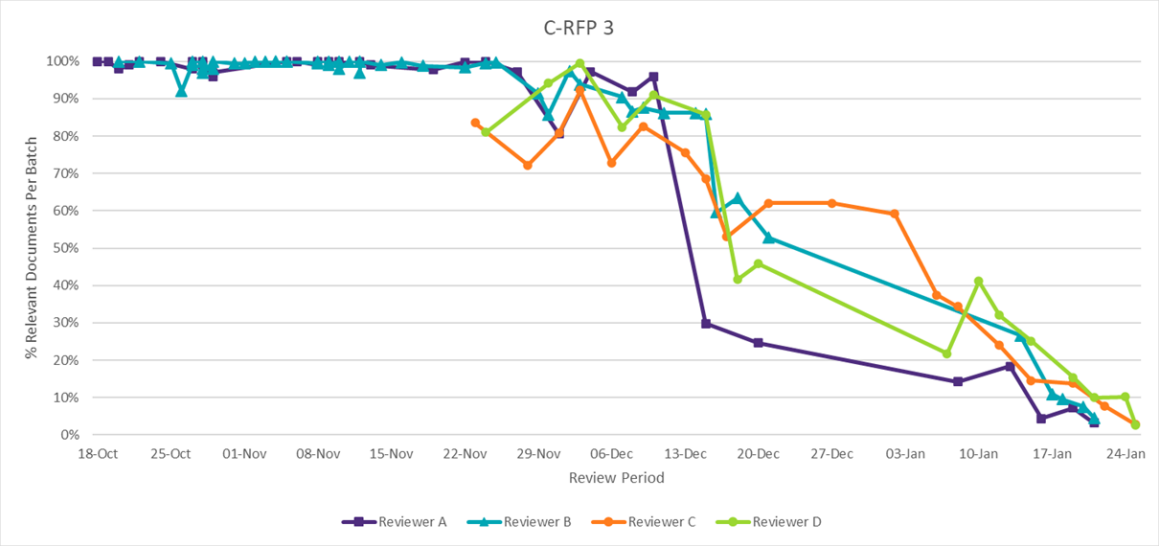


Figure 1: Composite Category #3 (4 Reviewers)
Batches 200-500 docs. Each line indicates an individual reviewer. Data points indicate the issuance of a new batch.

Similarly, in CC#5, a two-reviewer process (Figure 2), we see Reviewer B assessing a downturn in relevant documents present to them days before their partner, Reviewer A.

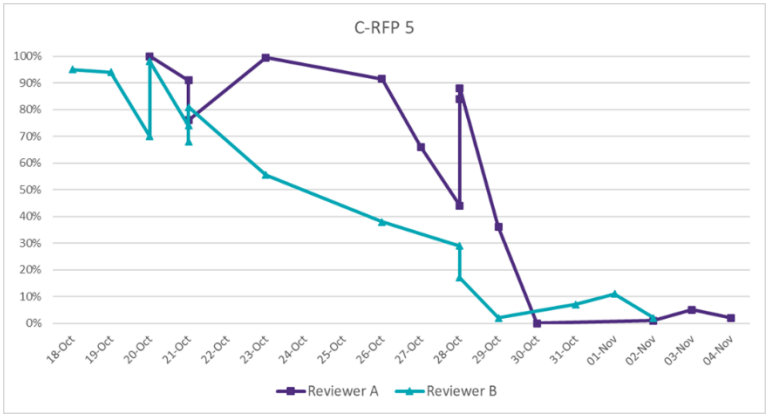


Figure 2: Composite Category 5 (2 Reviewers)
Batches 100-300 docs. Each line indicates an individual reviewer. Data points indicate the issuance of a new batch.

We conclude that there is not a great deal that can be done to correct this phenomenon. At the core of the issue is the simple and long-understood fact that different people often interpret the same documents differently, and this interpretation is influenced by multifarious factors arising both from review conditions and from the individuals' prior training, knowledge, and life experiences. In many cases, the reviewer with the anomalous review results was not strictly-speaking "wrong" in their relevance assessments, but rather had simply acquired a different understanding of what relevance truly was. In others, it turned out that the reviewer had simply "lost track" of what the criterion for relevance was. This was not a reflection of the reviewers' lack of intelligence, experience, or attention to detail, but instead something that could happen to anyone. It is the nature of humans that they identify patterns in data, and sometimes they fix on an incorrect pattern, leading them to go astray.

4.2. Disagreement on Relevance.

As discussed in Section 2 above, reviewers were assigned to specific reviews in pairs. They were also each assigned 10% of their review partner's documents to allow us to quickly identify when pairs of reviewers were diverging in their understanding of the relevance criterion (CC) in each review. Two of the composite categories underwent two separate reviews for reasons relating to the proceedings. In five of the reviews, the 10% of shared documents were also reviews independently by our QA team, made up of solicitors from the Maples Group who were deeply familiar with the issues in the legal proceedings, but resource limitations prevented us from doing this for all reviews. We recorded the levels of agreement between paired reviewers in respect of a review of large subset of the overall data set, and also between each reviewer in a pair and the QA reviewer and obtained the following results (ref. Table 3.)

Table 3
Documents reviewed in common between reviewers, by CC, percentage agreement on relevance.

Review, QILUA Dataset	Rev 1 vs Rev2	Rev 1 vs QA	Rev 2 v QA
Composite Category 1A	89%	89%	88%
Composite Category 1B	56%		
Composite Category 2A	83%		
Composite Category 2B	89%		
Composite Category 3	93%	93%	92%
Composite Category 4	92%		
Composite Category 5	74%	72%	69%
Composite Category 6	92%		
Composite Category 7	66%	74%	74%
Composite Category 8	69%		
Composite Category 9	86%	84%	81%
Composite Category 10	97%		

As can be seen, the method of using 10% of documents as "crossovers" for ensuring consistency between reviewers worked well in most cases and reviewers achieved levels of agreement greater than 80% in most reviews. There were, however, four cases where we were unable to achieve agreement as high as 80%. In particular, in the second review of CC1 we observed an agreement level of only 56% despite having the reviewers working in parallel and conferring on what documents should be considered relevant or not. Remember that the reviewers in question were similarly experienced, qualified barristers who had been briefed simultaneously using the same briefing material, yet still they could not consistently agree on relevance on a CC where previously another reviewer pair had achieved close agreement.

Similar characteristics were observed in the case of CCs 5, 7, and 8. In the case of CCs 5 and 7, bringing in QA did not make a substantial difference because the QA reviewers tended to disagree roughly equally with the assessments of both primary reviewers. These findings raise the question as to how much value a third, independent QA reviewer adds to the process, as it appears that QA reviewers' assessments are likely to be as subjective as those of the reviewers they are overseeing.

4.3. Extent of Reviewer Disagreement

As part of our assessment of aspects of the Grossman and Cormack CAL® tool, we decided that it would be helpful to determine the extent to which reviewers were disagreeing on the relevance of certain documents. We therefore selected three CCs and, instead of having the reviewer pairs review Relevant / Non-Relevant as usual, we had them review according to four degrees of relevance:

Strongly Agree / Agree / Unsure / Not-Relevant, and then assessed the differences between the paired reviewers' relevance assessment. Table 4 sets out our findings.

Table 4

Extent of Disagreement on Assessments of Documents against 3 CCs (document numbers)

Reviewer Designations	CC6	CC8	CC4
Agree / Not Relevant (2 Degs)	137	207	418
Agree / Unsure (1 Deg)	51	19	24
Not Relevant / Unsure (1 Deg)	207	52	46
Same Designation (0 Deg)	1,586 (66%)	1,945 (81%)	1,656 (69%)
Strongly Agree / Agree (1 Deg)	238	107	189
Strongly Agree / Not Relevant (3 Degs)	113	66	62
Strongly Agree / Unsure (2 Degs)	67	2	1
Total	2,399	2,398	2,396

As might be expected, in the case of all three CCs reviewed, most documents were assigned the same designation by both the paired reviewers, although not quite as high as might have been expected from the results in Table 3. However, where there was disagreement, it was often substantial. In CC6, 496 documents were reviewed with one degree of disagreement between the paired reviewers, 204 with two degrees, and 113 with three degrees (i.e., one reviewer designating a document "Strongly Agree," while the other designating it "Not Relevant"). In CC8, 178 documents were assessed with one degree of disagreement, 209 with two degrees, and 66 with three. In CC4, 259 documents were assessed with one degree of disagreement, 419 with 2 degrees, and 62 with three.

These findings indicate that it is unsafe to assume that, where reviewers disagree as to the relevance of a document that that disagreement is most probably minor. Instead, disagreement can be quite fundamental. For example, in the case of CC4, the reviewers disagreed by two degrees or more on over 20% of the documents reviewed. It was not merely a matter of one reviewer thinking that a document was probably relevant and the other that it was probably not. There was often fundamental disagreement as to the correct way to assess a document against the criterion provided.

4.4. Reviewer Disagreement Makes Recall Measures Unreliable.

In carrying out the comparison of the Grossman and Cormack CAL® system against the leading commercial discovery TAR tool, we relied upon confusion-matrix tests[2] to assess the relative Recall of both platforms for most of the ten CC criteria for which documents were reviewed. In the case of the commercial discovery TAR tool, we also carried out Elusion tests in the manner set out by the tool's manufacturer in their support documents and training courses. Each Elusion test was carried out twice for each CC, once with the test being completed by the same reviewer pair who carried out the CAL review, and a second time, but with the Elusion test review being completed by a different reviewer pair to those who had carried out the CAL review. The results are set out in Table 5 (overleaf).

Several results tend to stand out. The Recall results for the leading commercial TAR tool are usually substantially lower under confusion-matrix testing than in Elusion tests carried out both by the review team and by independent reviewers (note that time and resources were only available to carry out confusion-matrix tests for six of the ten CC reviews under both the Grossman and Cormack

[2] The formula for calculating Recall for the leading commercial TAR system under the confusion-matrix method discussed here was as follows: the number of relevant documents in the sample for commercial TAR tool / the total number of relevant documents in the sample for both the commercial and Grossman & Cormack CAL system + the number of mis-labelled relevant documents in the sample + the number relevant documents in the unreviewed population.
The formula for calculating Recall for the Grossman and Cormack system under the confusion-matrix method was the number of relevant documents in the sample for CAL® / the total number of relevant documents in the sample for both systems + the number of mislabelled relevant documents in the sample + the number relevant documents in the unreviewed population.
A detailed description of the method can be found in [10]

CAL® system and the leading commercial TAR one.) It appears that Recall results tend to be overstated in Elusion testing for reasons described in Section 3 above.

Table 5
Recall results under Confusion Testing, Grossman and Cormack CAL® System and under Confusion and Elusion Testing, Leading Commercial TAR System.

Recall Test	CC1	CC2	CC3	CC4	CC5	CC6	CC7	CC8	CC9	CC10
G&C CAL- Confusion	.83	N/A	.89	.58	.77	N/A	.58	N/A	.94	N/A
Leading TAR - Confusion	.73	N/A	.88	.54	.25	N/A	.43	N/A	.59	N/A
LeadingTAR - Elusion, Non-Independent Review	.94	.92	.89	.40	.40	.98	.67	.51	.91	.89
LeadingTAR- Elusion, Independent Review	.94	.96	.98	.54	.32	1.0	.93	.53	.4	.15

Recall results produced by Elusion testing on the leading commercial TAR system were quite inconsistent. It was by no means certain that using an independent Elusion test reviewer would produce lower Recall figures. Some results produced were highly problematic. CC6 was, in fact, a somewhat ambiguous review criterion which produced issues of review consistency throughout the testing process. Nevertheless, both Elusion tests supposedly proved that the TAR results had achieved almost perfect Recall. The real reason for the high Recall score was that, because the CC was ambiguously defined, Elusion reviewers could justify marking practically every document in the Elusion-testing sample as non-relevant, resulting in "perfect" Recall. This demonstrates a fundamental and insurmountable problem in the use of Recall as a validation method for TAR-based reviews.

4.5. Potential Influence of Review Platform on Reviewer Behaviour

As can be observed in Table 5, the Recall results produced using the Grossman and Cormack CAL® system were substantially better than those obtained by the same review team using the market leading TAR platform when both systems were assessed in the same way. The Grossman and Cormack CAL® system also obtained consistently higher review precision and numbers of responsive documents (on average over 30% more responsive documents identified) despite the commercial TAR systems' often high Elusion-testing-based Recall figures. This demonstrates that the CAL model employed in carrying out a legal review can have a substantial impact on the results of that review.

Another issue observed was that the commercial TAR tool, in training its CAL model made use of "uncertainty sampling.[3]" We are advised that the commercial TAR tool requires that around 30% of the documents reviewed in training to be non-relevant. We observed, however, that the use of uncertainty sampling seemed to confuse reviewers and increase the chances of them losing track of what constituted relevance. We had intended following-up on this finding further but have since been informed that in future releases of the commercial TAR tool, uncertainty sampling will be able to be manually switched off by the user at any point, usually when they predict stabilisation[4] has occurred. The impact this will have on the commercial TAR system's ability to continue to learn and accurately predict for new, yet, unfound classes of documents remains unclear.

Conversely, we observed that the Grossman and Cormack CAL® tool, which tends to provide reviewers with document sets much richer in relevant material, seemed to increase the chances that reviewers would mark equivocally relevant documents relevant.

Finally, the commercial TAR tool seemed to score ambiguous documents lower than the Grossman and Cormack CAL® tool. Many documents identified and marked relevant in the Grossman and

[3] Uncertainty Sampling, at least as it is defined in terms of the commercial tool tested in this research, involves the inclusion of low-ranked documents in the TAR review set to allow the TAR model to accurately model the cut-off between relevant and non-relevant documents. We believe that uncertainty sampling is required by the commercial tool because it bases its TAR model on Support Vector Machines, which requires the inclusion of relevant and non-relevant documents in training its model.

[4] In this context, stabilisation is reached at a point in training the model where the addition of further training data has little or no effect on the the rankings of documents within the model. It is a rather ill-defined concept.[ ]

Cormack CAL® review were never even seen in the review of the same documents carried out using the leading commercial TAR tool. This to some degree explains the better precision and responsive document retrieval rate observed using the Grossman and Cormack CAL® system.

5. Conclusions

There remains a problem with the fundamental concept of "relevance" certainly as it applies to information retrieval in the legal sphere and probably more generally in the discipline. However much one might be desired, there is unlikely to be any "gold standard" against which the relevance of documents to specific criteria can be assessed. Human language is often imprecise and it is the nature of legal proceedings that the documents involved are often equivocal in meaning and ambiguous in content. If there was no uncertainty in the nature of the documents involved, there might not, after all, be any legal questions to be decided in the first place.

In practice, documents can be fully relevant, probably relevant, or possibly relevant, and it is by no means certain that even the best reviewer will review them in a manner consistent with another competent reviewer. Reviewers draw the line between relevance and non-relevance in different places and in different circumstances, and often disagree with one another far more fundamentally than might be expected. The only occasion, it seems, when reviewers will keep a consistently conservative view of what constitutes relevance is in completing Elusion tests, where there is normally an incentive to find as few relevant documents as possible as this will maximise the Recall figure.

We have observed that reviewers:

- often lose track of what constitutes relevance while reviewing a document set. This error seems to occur more often the longer a review continues. We found that having reviewers review in pairs with 10% of crossover documents between them allowed us quickly to identify reviewers losing track of relevance, but it could not stop those reviewers from drifting off the relevance criteria.
- will also occasionally keep marking documents relevant in CAL reviews even when reviewers on the same project have begun coding most of the documents provided to them as non-relevant. This phenomenon seems to arise either from the reviewers losing track of relevance or because they have, during their review, developed a much broader sense of what is relevant than their counterparts may have.
- often fundamentally disagree about relevance. There may be an assumption that when reviewers disagree about how well a document aligns to a review criterion that such judgments are quite subtle. We found that, in fact, for a significant number of decisions disagreements between reviewers are considerable and fundamental.
- are influenced by how their review platform works and is set-up. In particular, it appears that feeding low-scoring documents into a CAL review can significantly affect reviewers' ability to "stay on track" in the review.
- disagree with QA reviewers quite as much as they do with one another. This raises the question of whether QA reviewers provide much additional value, particularly where other QA measures, such as paired reviewers are already in place.
- And, while this is considered in much more detail in [9], reviewers are, in practice, extremely poor at categorizing documents. The more categories there are, the worse and slower they get.

The findings described here are necessarily summary in nature. These are observations that were collected in research primarily focused in other areas. Nevertheless, all of these observed behaviours must necessarily have a significant bearing on the effectiveness and accuracy of any legal review using TAR or, indeed, any information-retrieval process requiring the human review of documents. There is an old saying in computer science "garbage in, garbage out." These findings suggest that perhaps more attention should be paid to the training of information-retrieval systems so that garbage does not creep in simply as a consequence of normal, unavoidable human behaviour.

6. Acknowledgements

The authors would like to acknowledge the assistance of Martin Elliot, the Partners and staff of Grant Thornton Ireland, and the partners and staff of the Maples Group in the research that led to this paper.

7. References

[1] H. L. Roitblat, "Search and information retrieval science." In Sedona Conf. J., vol. 8, p. 225. 2007.
[2] D.W. Oard, J.R. Baron, B. Hedin, et al. Evaluation of information retrieval for E-discovery. Artif Intell Law 18, 347–386 (2010). https://doi.org/10.1007/s10506-010-9093-9
[3] B. Zhou, Y. Yao, Evaluating information retrieval system performance based on user preference. J Intell Inf Syst 34, 227–248 (2010). https://doi.org/10.1007/s10844-009-0096-5
[4] G. V. Cormack and M. R. Grossman, "Navigating Imprecision in Relevance Assessments on the Road to Total Recall: Roger and Me," In Proc of. SIGIR '17, Aug. 2017, doi: 10.1145/3077136.3080812.
[5] M. R. Grossman and G. V. Cormack, "Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review," Richmond Journal of Law and Technology, vol. 17, no. 3, p. 11, Jan. 2011, [Online]. Available at: https://scholarship.richmond.edu/cgi/viewcontent.cgi?article=1344&context=jolt
[6] H. L. Roitblat, A. Kershaw, and P. Oot, "Document categorization in legal electronic discovery: computer classification vs. manual review," Journal of the Association for Information Science and Technology, vol. 61, no. 1, pp. 70–80, Oct. 2009, doi: 10.1002/asi.21233.
[7] M.R. Grossman, G.V. Cormack, A. Harbison, T. O'Halloran, B. McManus, (2024) "Unbiased Validation of Technology-Assisted Review for eDiscovery", In Proc of. SIGIR '24, July. 2024 doi: 10.1145/3626772.3657903
[8] A. Roegiest, G. V. Cormack, C. L. A. Clarke, and M. R. Grossman. Impact of surrogate assessments on high-recall retrieval. In Proc. SIGIR '15, 2015.
[9] B. McManus, T. O'Halloran, A. Harbison, M.R. Grossman, G.V. Cormack, (2024).. Limitations of the Utility of Categorization in eDiscovery Review Efforts. In: Li, S. (eds) Information Management. ICIM 2024. Communications in Computer and Information Science, vol 2102. Springer, Cham. https://doi.org/10.1007/978-3-031-64359-0_24
[10] T. O'Halloran, B. McManus, A. Harbison, M.R. Grossman, G.V. Cormack, (2024). Comparison of Tools and Methods for Technology-Assisted Review. In: Li, S. (eds) Information Management. ICIM 2024. Communications in Computer and Information Science, vol 2102. Springer, Cham. https://doi.org/10.1007/978-3-031-64359-0_9
[11] Government of Ireland, Statutory Instrument. No. 93/2009 - Rules of the Superior Courts (Discovery) 2009, https://www.irishstatutebook.ie/eli/2009/si/93/made/en/print
[12] U.K. Department of Justice, Civil Procedure Rules, Practice Direction 31A – Disclosure and Inspection, https://www.justice.gov.uk/courts/procedure-rules/civil/rules/part31/pd_part31a
[13] W. Webber, D. W. Oard, F. Scholer, and B. Hedin, "Assessor error in stratified evaluation," In Proc of. CIKM '10, Oct. 2010, doi: 10.1145/1871437.1871508.
[14] T. Saracevic, "Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance," Journal of the Association for Information Science and Technology, vol. 58, no. 13, pp. 1915–1933, Jan. 2007, doi: 10.1002/asi.20682.
[15] T. Saracevic, "Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behaviour and effects of relevance," Journal of the Association for Information Science and Technology, vol. 58, no. 13, pp. 2126–2144, Jan. 2007, doi: 10.1002/asi.20681.
[16] E. M. Voorhees, "Variations in relevance judgments and the measurement of retrieval effectiveness," Information Processing and Management, vol. 36, no. 5, pp. 697-716, Sep 2000, doi: 10.1016/s0306-4573(00)00010-8.

[17] James Waldron, John Rabiej, “Technology Assisted Review (TAR) Guidelines”, January 2019, EDRM, https://edrm.net/wp-content/uploads/2019/02/TAR-Guidelines-Final.pdf