When is it Rational to Review for Privilege?

Douglas W. Oard,[1] Jyothi Vinjumur,[1] Fabrizio Sebastiani[2]

[1]University of Maryland, College Park, USA
[2]Italian National Council of Research, Pisa, Italy
oard@umd.edu, jyothikv@umd.edu, fabrizio.sebastiani@isti.cnr.it

Abstract. Legal professionals are naturally reluctant to trust fully automated techniques for privilege review. There are certainly cases in which fully manual review is a rational choice, but there are also cases in which reliance on some degree of automation is rational. This paper proposes a model that can help practitioners to make rational choices, and to explain those choices once they have been made. The model balances two factors: the cost of performing review, and the risk of incurring a settlement, either from revealing privileged content or from failing to produce nonprivileged content.

Keywords: E-discovery, Cost Matrix, Privilege review

1 Motivation

In present practice, at least in the USA, discovery is normally conducted in three stages. In the first, some (but not all) documents[1] are collected for processing. In the second, collected documents that are responsive to the production request are identified. In the third, responsive documents are reviewed for privilege.[2] The first stage, referred to in the E-Discovery Reference Model (EDRM) as "Identification," is a matter for professional judgment, balancing (among other things) cost, time, and necessity. When the document set resulting from the first stage is large, it is today becoming increasingly common to perform some form of Technology Assisted Review (TAR), by which we mean a of review in which some type of automated text classification is used for at least a part of the process. The use of TAR for the third stage, review for privilege, is less common.

There are at least two reasons for this difference in practice between review for responsiveness and review for privilege. First, the ordering of the three stages

[1] Although we refer here to documents, we mean to use that term inclusively to refer to all Electronically Stored Information (ESI) for which review is required, with our term "document" consistently intended to refer to the unit of information (document, family, recording, etc.) on which review decisions must be rendered.

[2] There are several bases on which documents might properly be withheld from a production, some of which are properly referred to as privilege and others of which go by other names. For convenience, we group all such bases together and refer to them collectively as privilege in this paper.

reduces the number of documents that must be reviewed for privilege, thus rendering a linear manual review for privilege more affordable. Second, the failure to detect and properly withhold a privileged document might incur greater consequences for the party performing the review than would the failure to detect and—if later determined not to be privileged—properly disclose a responsive document.

These factors are amenable to quantification, at least to some degree: we can estimate the cost of reviewing a document; we can estimate the chance of making an incorrect decision, either by manual or by automated review; and we can estimate the consequences of incorrect decisions. With such estimates in hand, we can then determine the degree of automation that would be rational to employ in a specific case. Because some of these factors differ in different settings, the rational choice can and will be different in different cases.

To illustrate this point, imagine that a manual review for privilege can correctly detect 99% of the privileged documents, that an automated classifier exists that can correctly detect 80% of the privileged documents, that there are 50,000 responsive documents to be reviewed for privilege, that the cost of reviewing each document for privilege is $5.00,[3] and that (compared to that cost) running the automated classifier is essentially free. Then the cost of a manual linear review for privilege would be $250,000, and that review might still miss about 500 (i.e., 1% of 50,000) privileged documents. By contrast, a fully automatic review might miss about 10,000 privileged documents (i.e., 20% of 50,000). If, as seems reasonable to expect—at least in large cases—the risk incurred by unnecessarily disclosing 9,500 privileged documents is worth more, in dollar terms, than $250,000, then manual linear review would be the rational choice.[4] Now to consider the opposite extreme, imagine that everything is as in the previous example, except that the wizards toiling in the basement have come up with an automated privilege classifier that can find 99% of the privileged documents, and that can still be run at very little cost. If that were possible, automated review would be a rational choice.

As these oversimplified examples illustrate, the rational decision depends in part on the accuracy of manual review, in part on the accuracy of automated review, in part on the costs of each, and in part on the consequences—in dollar terms—of errors. As it also illustrates, there are situations in which linear manual review is the rational choice, other situations in which fully automated review would be the rational choice, and (as we show below) situations in which some mixture of manual and automated review would be rational. The reason that this third case—what is, in essence, TAR—is interesting is that automated classifiers actually have different error rates on different subsets of documents. Indeed, the wizards in the basement may already have built automated privilege classifiers that are nearly as accurate as human reviewers would be on some subsets of the documents, and for those documents it is the relative cost of reviews and

[3] For convenience, we express all costs in US Dollars.

[4] Of course, concerns might also arise from incorrectly classifying an unprivileged document as privileged; we account for that possibility in the next section.

mistakes that will determine what is rational. Formalizing that idea is our goal in this paper.

2 Modeling Privilege Review

George Box famously observed that "all models are wrong, but some are useful" [5]. Models are abstractions of reality that are created for some purpose, and the operative question is not whether they capture every aspect of reality—they never do—but rather whether the model reflects the salient aspects of reality sufficiently well for the results of the model to be useful. Let us thus start by saying what we would like our model to tell us, which is which documents should be subjected to manual review for privilege and which should be handled automatically. To formalize the factors that we need to consider, we can write the simple equation

$$E[c] = p(P|N) * c(P|N) + p(N|P) * c(N|P) \quad (1)$$

Reading that equation from left to right, we say that the expected cost $E[c]$ of making a mistake on any one decision can be computed as the probability $p(P|N)$ of misclassifying a document as privileged (P) given ($|$) that it is actually not privileged (N), times the cost $c(P|N)$ of that mistake, plus the probability times the cost for making the opposite mistake. There are several things to note here. First, we assume that there is no cost to getting things right, only to getting things wrong. This is a bit of an oversimplification, since it actually costs us something to build and run the classifier. But simplification is precisely what we want when we build a model—we want to get small and inconsequential distractions out of the way so that we can more easily see the key issues. If this factor is not actually small and inconsequential, we can add it back later. Second, the actual cost of a mistake may be large or small, depending on the nature of the mistake and perhaps also on the amounts at stake in the case, but if the probability of making that mistake is small then the *expected* cost from any one decision will be small. Third, the total expected cost for the review is simply the sum of the expected costs of each different decision that we might take. We could think of ways to make this equation more complex, but this is enough for us to get started.

Now with that little bit of math out of the way we can state the simplest version of our first key point: if $E[c]$, the expected cost of our automatic classifier making a mistake, is higher than the cost of manually reviewing the document in question, then it would be rational to review the document. (This is another oversimplification, since here we are assuming that human reviewers do not make mistakes; we will relax this assumption in Section 4.) And conversely, if the magnitude of the cost of reviewing the document exceeds the magnitude of the expected benefit (from the expected reduction in mistakes), then it would be rational not to review that document. Establishing that then leads to our next key point: deciding whether to review a specific document depends on five factors: two probabilities (i.e., the probability of each type of mistake) and three

costs (i.e., the cost for each type of mistake, and the cost for reviewing the document).

Taking these in turn, the easiest one to nail down is the cost of reviewing a document for privilege. That cost will be known (at least on average) for any workflow. For example, one source estimated that cost at $5.00 per document [1]. The other two costs are less commonly quantified, but we can certainly think of ways of making reasonable estimates. For example, in some specific case attorneys might quantify $c(N|P)$ by estimating that, on average, there is a 2% chance that incorrectly disclosing one privileged document might result in a $100,000 settlement. Such an estimate would result in an expected cost for an incorrectly disclosed privileged document being $2,000 (i.e, 2% of $100,000). For $c(P|N)$, attorneys might adopt a different estimation process, since the review is required only to be reasonable, not to be perfect. Thus they might estimate that $c(P|N)$ is zero until 4% more documents have been withheld than are actually privileged, after which their privilege review might be challenged and they might be compelled to turn over all documents on which they had claimed privilege, thus leading them to choose instead to pay the same $100,000 settlement. If there are 500 privileged documents, we can amortize the cost of that unpleasant surprise after 20 (4% of 500) mistakes by assuming that on average $c(P|N) = \$5,000$ for each of the first 20 mistakes. Of course, we don't initially know how many privileged documents actually exist, but we can estimate that value using sampling (and, of course, a limited amount of manual review). One clear implication of these examples is that the cost of a mistake depends on both the type of the mistake, and how we choose to quantify the costs. Another clear implication is that all such costs are estimates of averages. Averages will clearly suffice for our purpose (since in the end we are going to add up all the costs, and the sum of average costs is the same as the sum of costs). As for how the estimates should be made, that's a matter for the parties to discuss. But as we have illustrated, estimating expected costs is not rocket science – it is simply a matter of formalizing some factors that are easily articulated, but that might otherwise have gone unquantified.

That leaves us with the matter of how to estimate the probability of error. Many classifiers naturally produce some kind of a confidence value for each decision (i.e., they can tell us which of their decisions they are most confident are correct). Converting those confidence estimates into probabilities calls for some additional art, however. The key idea here is simple—we simply draw a small sample of the classifier's decisions, stratified by confidence (e.g., a few high-confidence decisions, a few medium-confidence decisions, and a few low-confidence decisions) and manually review the sample to see what the actual error rate is for each of those cases. Since the confidence estimates are typically numbers rather than categories, we can then interpolate between these measured

error rates to estimate the error rate for any confidence value that the classifier can produce.[5]

With this mathematical machinery at hand, the approach is then simple. All that is needed is to run the classifier on every document, sort the documents in decreasing order of $E[c]$, and then work down from the top of that list, manually reviewing documents for privilege as we go, until we reach the first document for which $E[c]$ (which is the cost we would expect to incur if we didn't review the document) is less than the cost of reviewing that document. This is a similar approach to that described by the third author of the present paper at DESI V [4], which has already been applied at least once to the review of sensitive content (in that case, for identifying documents that qualify for exemptions in government transparency laws) [3]. The approach we have described differs in two important ways from that prior work. First, in those cases the costs $c(P|N)$ and $c(N|P)$ that appear in Equation 1 had a particular form that is commonly used in information retrieval research, whereas we here allow them to be explicitly specified so as to give attorneys the flexibility to estimate them in whatever way makes sense in their case. Second, bringing into the framework both annotation costs and misclassification costs allows us to suggest a stopping point, where costs are minimized. But the general approach is otherwise identical.

3 Answering the Question

The question that we started with was "When is it rational to review for privilege?" We now have the means to answer that question using our model. To illustrate the approach, we consider the following simple way of estimating the cost of each type of error:

$$c(N|P) = VAR * x$$
$$c(P|N) = VAR/(PRIV * g)$$

where VAR is the value at risk (i.e., the expected cost of settling the case on disadvantageous terms), x (for "exposure") is the degree of risk (i.e., the probability) that inadvertently disclosing any one privileged document would result in choosing to settle, $PRIV$ is the estimated number of priviliged documents, and g (for "challenge") is the amortized risk that incorrectly classifying one more non-privileged document as privileged would risk presumptively waiving privilege, and thus choosing to settle. This is the same example that we introduced in the previous section; we have simply formalized it here. We emphasize that this model (and specifically the formula for $c(P|N)$) is oversimplified, but it makes for a useful example.

Choosing some specific values just for the purpose of an example, we could set x to 0.05% (i.e., you would need to inadvertently release about 2,000 privileged documents to risk a settlement) and g to 20% (i.e., you would need to

[5] This is a form of *probability calibration*, the usual approach to which is somewhat more complex than what we have described here in order to get the greatest benefit from the smallest sample [4,6].

incorrectly withhold 20% more documents than you properly should before you would likely be considered to have done a review so poor as to have waived privilege), and if we assume that both probabilities ($p(N|P)$ and $p(P|N)$) are never smaller than 0.05 and never higher than 0.95, then we can check to see if $E[c]$ is always below \$5.00 (the assumed cost of reviewing a document), always above \$5.00, or sometimes above and sometimes below \$5.00 (depending on those probabilities for specific documents). Those situations correspond, respectively, to it being rational (according to the model) to never, always, or sometimes choose to manually review a responsive[6] document for privilege.

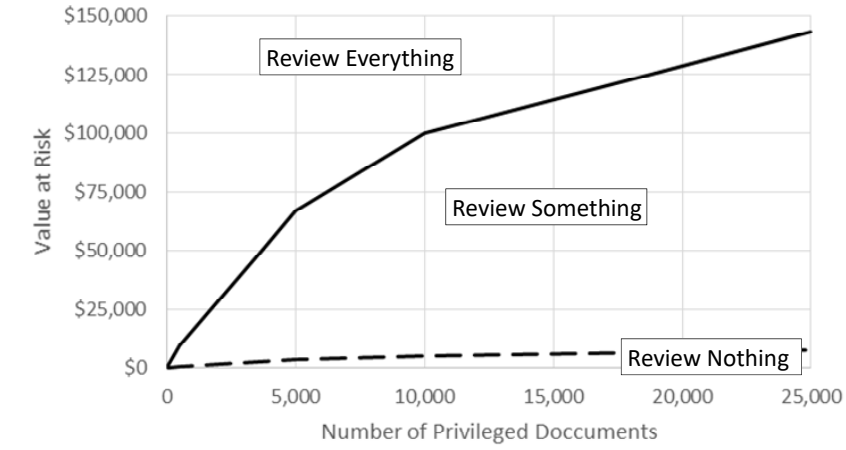

Fig. 1: The region between the dashed and solid lines is an example of a selective review region where some documents may require review and some may not.

Figure 1 depicts those three regions. To build this figure, we simply picked some values for VAR and PRIV and then used our model to compute E[c] for every possible value of $p(N|P)$ and $p(P|N)$.[7] Any values that are above the solid black line (e.g., VAR=\$125,000 and PRIV=10,000) correspond to a condition in which it would be rational to manually review everything. Any values that are below the dashed black line (e.g., VAR=\$3,000 and PRIV=5,000) corresponds to a condition in which it never makes sense to review anything. As can be seen, when the value at risk exceeds is very large it would never be rational to let the machine run unattended, and when the value at risk is very small it would not generally be rational to manually review anything. The key point here is not

[6] Note that we don't distinguish between truly responsive documents and those that are believed, after review, to be responsive. Rather, we model the situation as if review for responsiveness had been perfect. This is yet another of our simplifications.

[7] In practice, we only need to try this with values of 0.05 and 0.95 for those probabilities.

the specific numbers—which were chosen just to illustrate that the regions we have shown can exist—it is that defining a model of the type we have described can tell you whether or not it is rational to conduct a manual linear review. Moreover, although not illustrated here, when it is not rational to review every document, such a model also can offer guidance on which documents do and which do not rationally merit manual review. In other words, what is important here is not our model, but rather your model—the one you develop to more closely reflect the reality of the case you are actually working on.

4 The Rest of the Story

As George Box has reminded us, all models, including the one we have introduced above, are wrong. So if we are to assess the degree to which models of the structure we have presented in Section 3 are useful for their intended purpose (of telling us which documents to review) it is essential that we itemize the known limitations of that model. First, in comparing the expected cost $E[c]$ of automatic classification with the cost of manual review we have implicitly assumed that human review is perfect. We know that's not the case, however; the best we can say for manual review is that we know *a priori* (indeed, almost by definition) that if properly conducted by qualified individuals, manual review is reasonable. We can and should measure the probability that human reviewers will make the same type of errors, and with that information in hand we can compute $E[c]$ for human reviewers just as we have for machines. Second, we know that building and running a classifier is not free—the wizards in the basement expect to be paid. But estimating those costs is not hard, and for any existing automated workflow they would already be known. If we use an a to indicate automatic review and an m to indicate manual review, then we can improve our model by computing:

$$E[c_a] = p_a(P|N) * c(P|N) + p_a(N|P) * c(N|P) + r_a$$
$$E[c_m] = p_m(P|N) * c(P|N) + p_m(N|P) * c(N|P) + r_m$$

where we use r_a and r_m to indicate the cost of automatic classification and of manual review, respectively. Now instead of sorting the documents in decreasing order by $E[c]$ we can instead sort them in decreasing order of $E[c_a] - E[c_r]$ and we can stop when that difference gets to zero. The key ideas are the same, but the model is now a better reflection of reality, and thus more useful as a guide to what we really should do.

Even this more sophisticated model has limitations, however. For example, we have not modeled the cost of actually preparing the privilege log, which may require details that are beyond the ken of our classifier. Another concern is that we have made a whole host of assumptions. For example, in one illustrative scenario we assumed that a 4% error rate would risk waiving privilege. What if that were really 5% or 3%? Would we review more documents? Fewer documents? Different documents? Similarly, we assumed in one illustrative scenario that a

settlement would cost us $100,000. What if we were to settle for less, or for more? We can use sensitivity analysis, a standard approach of rerunning our model with different assumptions, to evaluate what would change. These sorts of uncertainties have been called "known unknowns"—we don't know precisely what values we should use, but we do at least know what in our model might not be quite correct.

Sensitivity analysis provides a useful way of exploring the effect of known unknowns. But as the Great Depression, the Fukushima Daiichi nuclear powerplant disaster, and the asteroid that wiped out the dinosaurs [2] illustrate, some problems are harder to anticipate than others. Nassim Taleb has referred to these cases in which probability theory fails us as "black swan" events (since until you know that black swans exist, you have no way to estimate the probability that the next swan you see will be black) [10]. It's not hard to think of cases in which our estimates might be wrong for reasons not accounted for in our model. For example, we might estimate how often manual review is wrong using the reviewers we have, but then when we actually do the review some of those reviewers may have left and the reviewers we have hired to replace them might be better (or worse). As another example, recently Adam Roegiest and Gordon Cormack have shown that reviewers will judge the same document differently if you show it to them in the midst of many or of few positive examples [7]. These are not really black swans, however, since they are issues we can anticipate and thus issues that we could model if we wish to. The true black swan is when you discover Roegiest and Cormack's result before they did, and you do so in the midst of a case. Maybe it will be a pleasant surprise, maybe not.

One way of dealing with things that we cannot anticipate, which also works for things that we could have anticipated but (for whatever reason) chose not to model, is to take an "administrator's discount." That term was used by James Webb, the NASA Administrator at the time the Apollo Program was first proposed. Asked by President Kennedy what Apollo would cost, he replied $20 billion. When asked by his staff how he had arrived at that figure when NASA engineers had told him their estimate was between $8 billion and $12 billion, his reply was "administrator's discount" [8]. Webb had been director of the Bureau of the Budget in an earlier presidential administration, and he understood that initial estimates are often low and rarely high. That was not a very sophisticated model, but it worked. So the question to ask is, what kind of a surprise do you want? Do you want to learn after the fact that you spent too much money on manual review, or that you spent too little money on manual review? This is a matter for judgment, and it would not be surprising if many of us were to stand with James Webb and take an administrator's discount by going just a little further down the list before stopping.

5 Conclusion

We have used probability theory, dinosaurs, and the Moon landings, among other things, to illustrate how attorneys can make rational decisions about which doc-

uments to review for privilege. The fact that we do not know everything is no excuse for not using what we do know, and doing so in some rational way. George Box offers us a starting point, allowing us to abstract away from the full complexities of reality to build models that are simple enough to be practical, and yet informative enough to be useful. James Webb then reminds us that informative does not mean prescriptive, that blindly believing our own estimates has its own risks, and that rationality and human judgment are not in tension. But just as it would not be rational to do exactly what an imperfect model tells us to do, neither would it be rational to ignore the results an an informative model that we know how to build. Furthermore, models are not just a source of guidance, they are also a way of facilitating discussion. Asked by a client, why you do (or do not) plan to produce some documents without first reviewing them for privilege, a model can provide a framework for informing that discussion. Asked by a counterparty why you do not (or do) plan to withhold some documents as privileged based solely on the determination of an automated classifier, your model can provide a framework for that discussion as well. In this sense, the model serves as what Leigh Star and James Griesemer called a "boundary object" that different parties can use to facilitate their communication [9]. We might thus say that such models are useful devices that can help us in several ways to achieve some out-of-the-Box thinking.

Acknowledgements

This work has been supported in part by NSF grants 1065250 and 1618695. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

1. Da Silva Moore v. Publicis Groupe. In *F. Supp. 2d*, volume 868, page 137. Dist. Court, SD New York, 2012.
2. L. W. Alvarez, W. Alvarez, F. Asaro, and H. V. Michel. Extraterrestrial cause for the Cretaceous-Tertiary extinction. *Science*, 208(4448):1095–1108, 1980.
3. G. Berardi, A. Esuli, C. Macdonald, I. Ounis, and F. Sebastiani. Semi-automated text classification for sensitivity identification. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015*, pages 1711–1714, 2015.
4. G. Berardi, A. Esuli, and F. Sebastiani. Utility-theoretic ranking for semiautomated text classification. *ACM Transactions on Knowledge Discovery from Data*, 10(1):6:1–6:32, 2015.
5. G. E. Box. Robustness in the strategy of scientific model building. In R. Launer and G. Wilkinson, editors, *Robustness in statistics*, pages 201–236. Academic Press, 1979.
6. A. Niculescu-Mizil and R. Caruana. Obtaining calibrated probabilities from boosting. In *Proceedings of the 21st Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI 2005)*, pages 413–420, Arlington, US, 2005.

7. A. Roegiest and G. V. Cormack. Impact of review-set selection on human assessment for text classification. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, pages 861–864, 2016.
8. R. Seamans. *Aiming at Targets: The Autobiography of Robert C. Seamans, Jr.* NASA History Office, 1996.
9. S. L. Star and J. R. Griesemer. Institutional ecology, translations and boundary objects: Amateurs and professionals in Berkeley's museum of vertebrate zoology, 1907-39. *Social Studies of Science*, 19(3):387–420, 1989.
10. N. N. Taleb. *The Black Swan: The Impact of the Highly Improbable.* Random House Publishing Group, 2007.