5

Gamesmanship in Modern Discovery Tech

Neel Guha, Peter Henderson, and Diego A. Zambrano

This chapter explores the potential for gamesmanship in technology-assisted discovery.[1] Attorneys have long embraced gamesmanship strategies in analog discovery, producing reams of irrelevant documents, delaying depositions, or interpreting requests in a hyper-technical manner.[2] The new question, however, is whether machine learning technologies can transform gaming strategies. By now it is well known that technologies have reinvented the practice of civil litigation and, specifically, the extensive search for relevant documents in complex cases. Many sophisticated litigants use machine learning algorithms – under the umbrella of "Technology Assisted Review" (TAR) – to simplify the identification and production of relevant documents in discovery.[3] Litigants employ TAR in cases ranging from antitrust to environmental law, civil rights, and employment disputes. But as the field becomes increasingly influenced by engineers and technologists, a string of commentators has raised questions about TAR, including lawyers' professional role, underlying incentive structures, and the dangers of new forms of gamesmanship and abuse.[4]

[1] Much of this chapter is based on and extends our previous work: Neel Guha, Peter Henderson & Diego A. Zambrano, *Vulnerabilities in Discovery Tech*, 35 Harv. J.L. & Tech. 581 (2022).

[2] *See, e.g.*, Frank H. Easterbrook, Comment, *Discovery as Abuse*, 69 B.U. L. Rev. 635, 637 (1989); Linda S. Mullenix, *The Pervasive Myth of Pervasive Discovery Abuse: The Sequel*, 39 B.C. L. Rev. 683, 684–85 (1998).

[3] Stephen Embry, Am. Bar. Ass'n, 2020 Litigation & TAR (2020), https://www.americanbar.org/groups/law_practice/publications/techreport/2020/litigationtar/.

[4] Some scholars have specifically warned about the dangers of abuse. *See, e.g.*, David Freeman Engstrom & Jonah B. Gelbach, *Legal Tech, Civil Procedure, and the Future of Adversarialism*, 169 U. Pa. L. Rev. 1001, 1073 (2020) ("[A]utomated discovery might breed more abuse, and prove less amenable to oversight, than an analog system built upon 'eyes-on' review."); Seth K. Endo, *Technological Opacity & Procedural Injustice*, 59 B.C. L. Rev. 821 (2018) (same); Dana A. Remus, *The Uncertain Promise of Predictive Coding*, 99 Iowa L. Rev. 1691, 1709 (2014) (same). Other have instead focused on the need for attorneys to supervise technologists and remain technically competent. *See, e.g.*, Shannon Brown, *Peeking Inside the Black Box:*

https://doi.org/10.1017/9781009255301.008 Published online by Cambridge University Press

This chapter surveys and explains the vulnerabilities in technology-assisted discovery, the risks of adversarial gaming, and potential remedies. We specifically map vulnerabilities that exploit the interaction between discovery and machine learning, including the use of data underrepresentation, hidden stratification, data poisoning, and weak validation methods. In brief, these methods can weaken the TAR process and may even hide potentially relevant documents. We also suggest ways to police these gaming techniques. But the remedies we explore are not bulletproof. Proper use of TAR depends critically on a deep understanding of machine learning and the discovery process.[5] Ultimately, this chapter argues that, while TAR does suffer from some vulnerabilities, gamesmanship may often be difficult to perform successfully and can be counteracted with careful supervision. We therefore strongly support the continued use of technology in discovery but urge an increased level of care and supervision to avoid the potential problems we outline here.

5.1 OVERVIEW OF DISCOVERY AND TAR

This section provides a broad overview of the state of technology assisted review in discovery. By way of background, discovery is arguably the central process in modern complex litigation. Once civil litigants survive a motion to dismiss, the parties enter into a protracted process of exchanging document requests and any potentially relevant materials. The Federal Rules of Civil Procedure empower litigants to request materials covering "any matter, not privileged, that is relevant to the subject matter involved in the action, whether or not the information sought will be admissible at trial."[6] This gives litigants a broad power to investigate anything that may be relevant to the case, even without direct judicial supervision. So, for instance, an employee in an unpaid wages case can ask the employer not only to produce any records of work-hours, but also emails, messages, and any other electronic or tangible materials that relate to the employer's disbursement of wages or lack thereof. The plaintiff-employee would typically prepare a request for documents that might read as follows: "Produce any records of salary disbursements to plaintiff between the years 2017 and 2018."

A Preliminary Survey of Technology Assisted Review (TAR) and Predictive Coding Algorithms for eDiscovery, SUFFOLK J. TR. & APP. ADVOC., June 2016, at 1 (warning against lawyers' reliance on "outside advisors"); Daniel N. Kluttz & Deirdre K. Mulligan, *Automated Decision Support Technologies and the Legal Profession*, 34 BERKELEY TECH. L.J. 853, 884 (2019) (same). There are also recent bills to address discovery gamesmanship. *See, e.g.*, Tom Umberg, *Document Production Gamesmanship Run Amuck – Governor Newsom Should Sign SB 17*, SENATOR TOM UMBERG (Sept. 27, 2020), https://sd34.senate.ca.gov/news/9272019-document-production-gamesmanship-run-amuck-%E2%80%93-governor-newsom-should-sign-sb-17.

[5] Recent scholarship has explored this challenge from a variety of perspectives. *See, e.g.*, William Matthewman, *Towards a New Paradigm for E-Discovery in Civil Litigation: A Judicial Perspective*, 71 FLA. L. REV. 1261 (2019) (arguing that a new paradigm is necessary to regulate discovery tech).

[6] Diego A. Zambrano, *Discovery as Regulation*, 119 MICH. L. REV. 71, 80 (2020).

https://doi.org/10.1017/9781009255301.008 Published online by Cambridge University Press

Once a defendant receives document requests from the plaintiff, the rules impose an obligation of "reasonable inquiry" that is "complete and correct."[7] This means that a respondent must engage in a thorough search for any materials that may be "responsive" to the request. Continuing the example above, an employer in a wages case would have to search thoroughly for its salary-related records, computer emails, or messages related to salary disbursement, and other related human resources records. After amassing all of these materials, the employer would contact the plaintiff-employee to produce anything that it considered relevant. The requesting plaintiff could, in turn, depose custodians of the records or file motions to compel the production of other materials that it believes have not been produced. Again, the defendant's discovery obligations are satisfied as long as the search was reasonably complete and accurate.

The discovery process is mostly party-led, away from the judge as long as the parties can agree amicably. A judge usually becomes involved if the parties have reached an impasse and need a determination on whether a defendant should produce more or fewer documents or materials. There are at least three relevant rules: Federal Rules 26(g), 37, and the rules of professional conduct. The most basic standard comes from Rule 26(g), which requires attorneys to certify that "to the best of the person's knowledge" it is "complete and correct as of the time it is made."[8] Courts have sometimes referred to this as a negligence-like standard, punishing attorneys only when they have failed to conduct an appropriate search.[9] By contrast, FRCP 37 provides for sanctions against parties who engage in discovery misfeasance "with the intent to deprive another party of the information's use in the litigation."[10] Finally, several rules of professional conduct provide that lawyers shall not "unlawfully obstruct another party's access to evidence" or "conceal a document," and should not "fail to make reasonably diligent effort to comply with a legally proper discovery request."[11]

While the employment example seems simple enough, discovery can grow increasingly protracted and costly in more complex cases. Consider, for instance, antitrust litigation. Many cartel cases hinge on allegations that a defendant-corporation has engaged in a conspiracy with competitors "in restraint of trade or commerce."[12] Given the requirements of federal antitrust laws, the existence of a conspiracy can become a convoluted question about the operations of a specific market, agreements not to compete, or rational market behavior. This, in turn, can involve millions of relevant documents, emails, messages, and the like, especially

[7] Fed. R. Civ. P. 26(g).

[8] *Id.*

[9] Fjelstad v. Am. Honda Motor Co., Inc., 762 F.2d 1334, 1343 (9th Cir. 1985) ("We consistently have held that sanctions may be imposed even for negligent failures to provide discovery.").

[10] Fed. R. Civ. P. 37.

[11] Model Rules of Pro. Conduct r. 3.4. (Am. Bar Ass'n 2021).

[12] 15 U.S.C. § 1.

https://doi.org/10.1017/9781009255301.008 Published online by Cambridge University Press

because "[m]odern cartels employ extreme measures to avoid detection."[13] A high-end antitrust case can easily reach discovery expenditures in the millions of dollars, as the parties prepare expert reports, engage in exhaustive searches for documents, and plan and conduct dozens of depositions.[14] A RAND 2012 study found that document review and production could add up to nearly $18,000 per gigabyte – and most of the cases studied involved over a hundred gigabytes (a trifle by 2022 standards).[15]

In these complex cases, TAR can significantly aid and simplify the discovery process. Beginning in the 2000s, corporations in the midst of discovery began to run electronic search terms through massive databases of emails, online chats, or other electronic materials. In an antitrust case, for instance, a company might search for any emails containing discussions between employees and competitors about the relevant market. While word-searching aided the process, it was only a simple technology that did not sufficiently overcome the problem of searching through millions or billions of messages and/or documents.[16]

Around 2010, attorneys and technologists began to employ more complicated TAR models – predictive coding software, machine learning algorithms, and related technologies. Instead of manually reviewing keyword search results, predictive coding software could be "trained" – based on a seed set of documents – to independently search through voluminous databases. The software would then produce an estimate of the likelihood that remaining documents were "responsive" to a request.

[13] Brief of the American Antitrust Institute as Amici Curiae in Support of Respondents at *4, Bell Atlantic Corp. v. Twombly, 550 U.S. 544 (2007) (No. 05-1126).

[14] John M. Majoras, *Antitrust Pleading Standards after* Twombly, Jones Day (June 2007), https://www.jonesday.com/en/insights/2007/06/antitrust-pleading-standards-after-itwomblyi (noting that some antitrust cases involve "frequently millions of dollars in legal fees and discovery costs").

[15] *See* Kluttz & Mulligan, *Automated Decision Support Technologies* (citing Nicholas M. Pace & Laura Zakaras, RAND Institute for Civil Justice, Where the Money Goes: Understanding Litigant Expenditures for Producing Electronic Discovery 20, 16 n.39 (2012)); Eleanor Brock, *eDiscovery Opportunity Costs: What Is the Most Efficient Approach?* logikcull (Nov. 21, 2018), https://www.logikcull.com/blog/ediscovery-opportunity-costs-infographic; Casey Sullivan, *What a Million-Dollar eDiscovery Bill Looks Like*, Logikcull (May 9, 2017), https://www.logikcull.com/blog/million-dollar-ediscovery-bill-looks-like (describing a $13 million payment to discovery vendors in a case involving 3.6 terabytes of data).

[16] The Sedona Conference, *The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery*, 8 Sedona Conf. J. 189, 200–202 (2007) (discussing the use of keyword searching and its problems); Howard Sklar, *Match Point with Recommind's Predictive Coding – It's "Man with Machine," not "Man vs. Machine"*, Corp. Couns. Bus. J. (Aug. 1, 2011), https://ccbjournal.com/articles/match-point-recomminds-predictive-coding-its-man-machine-not-man-vs-machine (discussing the weaknesses of keyword searching).

https://doi.org/10.1017/9781009255301.008 Published online by Cambridge University Press

Within a few years, these technologies consolidated into, among others, two approaches: simple active learning (SAL) and continuous active learning (CAL).[17] With SAL, attorneys first code a seed set of documents as relevant or not relevant; this seed set is then used to train a machine learning model; and finally the model is applied to all unreviewed documents in the dataset. Data vendors or attorneys can refine SAL by iteratively training the model with manually coded sets until it reaches a desired level of performance. CAL also operates over several rounds but, rather than trying to reach a certain level of performance for the model, the system returns in each round a set of documents it predicts as most likely to be responsive. Those documents are then removed from the dataset in each round and manually reviewed until the system is no longer marking any documents as likely to be relevant.

Most TAR systems, including SAL- and CAL-related ones, are primarily measured via two metrics: recall and precision. Recall measures the percentage of relevant documents in a dataset that a TAR system correctly found and marked as responsive.[18] The only way to gauge the percentage of relevant documents in a dataset is to manually review a random sample. Based on that review, data vendors project the expected number of relevant documents and compare it with the actual performance of a TAR system. Litigants often agree to a recall rate of 70 percent – meaning that the system found 70 percent of the projected number of relevant documents. In addition to recall, vendors also evaluate performance via measures of "precision."[19] This metric focuses instead on the quality of the TAR system – capturing whether the documents that a system marked as "relevant" are actually relevant. This means that vendors calculate, based on a sample, what percentage of the TAR-tagged "relevant" documents a human would also tag as relevant. As with recall, litigants often agree to a 70 percent precision rate.

Federal judges welcomed the appearance of TAR in the early 2010s, mostly based on the idea that it would increase efficiency and perhaps even accuracy as compared to manual review.[20] Dozens of judicial opinions defended the use of TAR as the

[17] "Continuous Active Learning" can refer both to a specific product developed and trademarked by Maura R. Grossman and Gordan V. Cormack, or to a general class of algorithms sharing common attributes. *Compare* Continuous Active Learning, Registration No. 5876987, *with* Matthew Verga, *Alphabet Soup: TAR, CAL, and Assisted Review, Assisted Review Series Part 1*, XACT Data Discovery (Sept. 15, 2020), https://xactdatadiscovery.com/articles/predictive-coding-evolution/.

[18] Maura R. Grossman & Gordan V. Cormack, *Vetting and Validation of AI-Enabled Tools for Electronic Discovery*, *in* Litigating Artificial Intelligence 3 (Jill Presser, Jesse Beatson & Gerald Chan, eds., 2021).

[19] *Id.* at 14. Consequently, researchers believe that achieving more than 70 percent recall and 70 percent precision for any system is difficult. *Id.* at 15.

[20] Herbert L. Roitblat et al., *Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review*, 61 J. Am. Soc'y For Info. Sci. & Tech. 70, 74–75 (2010); Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, 17 Rich. J.L. & Tech.

https://doi.org/10.1017/9781009255301.008 Published online by Cambridge University Press

potential silver bullet solution to discovery of voluminous databases.[21] Importantly, most practicing attorneys accepted TAR as a basic requirement of modern discovery and quickly incorporated different software into their practices.[22] By 2013, most large law firms were either using TAR in many of their cases or experimenting with it.[23] Eventually, however, some academics and practitioners began to criticize the opacity of TAR systems and the potential underperformance or abuse of technology by sophisticated parties.[24]

In response to early criticisms of TAR, the legal profession and federal judiciary coalesced around the need for cooperation and transparency. Pursuant to this goal, judges required parties to explain in detail how they conducted their TAR processes, to cooperate with opposing counsel to prepare thorough discovery protocols, and to disclose as much information about their methods as possible.[25] For instance, one judge required producing parties to "provide the requesting party with full disclosure about the technology used, the process, and the methodology, including the documents used to 'train' the computer."[26] Another court asked respondents to produce "quality assurance; and . . . prepare[] to explain the rationale for the method chosen to the court, demonstrate that it is appropriate for the task, and show that it was properly implemented."[27]

Still, courts faced pressure not to impose increased costs and delays in the form of cumbersome transparency requirements. Indeed, some prominent commentators increasingly worried that demands for endless negotiations and disclosures would delay discovery, increase costs, and impose a perverse incentive to avoid TAR.[28] In response, courts and attorneys moved toward a standard of "deference to a producing party's choice of search methodology and procedures."[29] A few courts embraced a

1, 3 (2011); Thomas Barnett et al., *Machine Learning Classification for Document Review*, 2009 DESI III: ICAIL WORKSHOP ON GLOB. E-DISCOVERY/E-DISCLOSURE.

[21] *See, e.g.*, Progressive Cas. Ins. Co. v. Delaney, No. 2:11-CV-00678-LRH, 2014 WL 3563467, at *8 (D. Nev. July 18, 2014); Hyles v. New York City, No. 10CIV3119ATAJP, 2016 WL 4077114, at *2 (S.D.N.Y. Aug. 1, 2016).

[22] Endo, *Technological Opacity & Procedural Injustice*, at 837–38.

[23] David Freeman Engstrom & Nora Freeman Engstrom, *TAR Wars: E-Discovery and the Future of Legal Tech*, 96 ADVOCATE 19 (2021).

[24] Endo, *Technological Opacity & Procedural Injustice*, at 837.

[25] Henderson, Guha & Zambrano, *Vulnerabilities in Discovery Tech*, at 13 (citing Youngevity Int'l Corp. v. Smith, No. 16CV00704BTMJLB, 2019 WL 1542300, at *12 (S.D. Cal. Apr. 9, 2019); *Progressive Cas.*, 2014 WL 3563467, at *10; William A. Gross Const. Assocs., Inc. v. Am. Mfrs. Mut. Ins. Co., 256 F.R.D. 134, 135 (S.D.N.Y. 2009).

[26] Henderson, Guha & Zambrano, *Vulnerabilities in Discovery Tech*, at 13 (citing *Youngevity Int'l*, 2019 WL 1542300, at *12).

[27] *William A. Gross*, 256 F.R.D. at 135; In re Seroquel Products Liability Litig., 244 F.R.D. 650, 662 (M.D. Fla. 2007).

[28] Christine Payne & Michelle Six, A *Proposed Technology-Assisted Review Framework*, LAW360 (Apr. 27, 2020), https://perma.cc/9DZJ-7FSQ.

[29] *Progressive Cas.*, 2014 WL 3563467, at *10.

https://doi.org/10.1017/9781009255301.008 Published online by Cambridge University Press

presumption that a TAR process was appropriate unless opposing counsel could present "specific, tangible, evidence-based indicia … of a material failure."[30]

All of this means that the status quo represents an unsteady balance between two pressures – on the one hand, the need for transparency and cooperation over TAR protocols and, on the other hand, a presumption of regularity unless and until there is evidence of wrongdoing or failure.

Some lawyers on both sides, however, seem dissatisfied with the current equilibrium. Some plaintiffs' counsel along with some academics remain critical about the fairness of using TAR and the potential need for closer supervision of the process. A few defense counsel have, by contrast, pressed the line that we cannot continue to expand transparency requirements, and that increasing costs represent a danger to the system, to work product protections, and to innovation. Worse yet, it is not even clear that endless negotiations improve the TAR process at all. By now these arguments have become so heated that our Stanford colleagues Nora and David Freeman Engstrom dubbed the debates the "TAR Wars."[31] It bears repeating that the stakes are significant and clear: Requesting parties want visibility over what can sometimes be an opaque process, clarity over searches of voluminous databases, and assurances that each TAR search was complete and correct. Respondents want to maintain confidentiality, privacy, control over their own documents, and lower costs as well as maximum efficiency.

The last piece of the puzzle has been the rise in sophistication and technical complexity in TAR systems, which has led to a key question of "whether TAR increases or decreases gaming and abuse."[32] After 2015, both SAL and CAL became dominant across the complex litigation world. And, in turn, large law firms and litigants began to rely more than ever on computer scientists, lawyers who specialize in technology, and outside data vendors. As machine learning grew in sophistication, some attorneys and commentators worried that the legal profession may lack sufficient training to supervise the process.[33] A string of academics, in turn, have by now offered a range of reforms, including forced sharing of seed sets, validation by neutral third parties, and even a reshuffling of discovery's usual structure by having the requesting party build and tune the TAR system.[34]

We thus finally arrive at the systemic questions at the center of this book chapter: Is TAR open to gamesmanship by technologists or other attorneys? If so, how? Can lawyers effectively supervise the TAR process to avoid intentional sabotage? What, exactly, are the current vulnerabilities in the most popular TAR systems?

[30] *The Sedona Principles, Third Edition: Best Practices, Recommendations & Principles for Addressing Electronic Document Production*, 19 Sedona Conf. J. 1 (2018).

[31] Engstrom & Engstrom, *TAR Wars*, at 19.

[32] Engstrom & Gelbach, *Legal Tech*, at 1072.

[33] *Id.* at 1046.

[34] *See, e.g., id.*; Bruce H. Kobayashi, *Law's Information Revolution as Procedural Reform: Predictive Search as a Solution to the In Terrorem Effect of Externalized Discovery Costs*, 2014 U. Ill. L. Rev. 1473, 1509.

https://doi.org/10.1017/9781009255301.008 Published online by Cambridge University Press

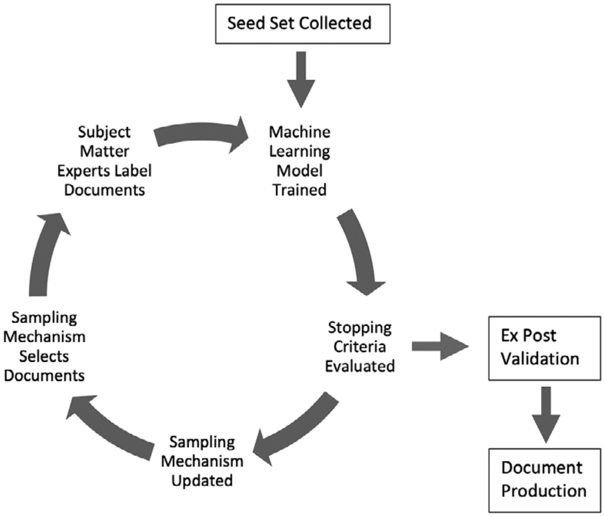
Seed Set Collected
Subject Matter Experts Label Documents
Machine Learning Model Trained
Sampling Mechanism Selects Documents
Stopping Criteria Evaluated
Ex Post Validation
Sampling Mechanism Updated
Document Production

FIGURE 5.1 The TAR 2.0 process: A stylized example

5.2 GAMING TAR

In this section we explain how litigants could game the TAR process. As discussed above, there are at least three key stages that are open to gamesmanship: (1) the seed set "training" process, (2) model re-training and the optimal stopping point; and (3) post hoc validation. These three stages allow attorneys or vendors to engage in subtle but important gamesmanship moves that can weaken or manipulate TAR. Figure 5.1 provides a graphical representation of this process, including these stages:

Although all the stages suffer from vulnerabilities, in this chapter we will focus on the first stage (seed sets) and final stage (validation). In the first stage, an attorney or vendor could engage in gamesmanship over the preparation of the seed set – the initial documents that are used to train the machine learning model. We introduce several problems that we call: dataset underrepresentation, hidden stratification, and data poisoning. Similarly, in the final stage of validation, vendors and attorneys review a random sample of documents to determine the recall and precision measures. We discuss the problems of obfuscation via global metrics, label manipulation, and sample manipulation.

Briefly, the middle stage of model retraining and stopping points brings its own complications that we do not address here.[35] After attorneys train the initial model, vendors can then use active learning systems (either SAL or CAL) to re-train the model over iterative stages. For SAL, vendors typically use what is called

[35] For a longer discussion of the middle stage and other vulnerabilities, see generally Guha, Henderson & Zambrano, *Vulnerabilities in Discovery Tech*.

https://doi.org/10.1017/9781009255301.008 Published online by Cambridge University Press

"uncertainty sampling," which flags for vendors and attorneys the documents that the model is most uncertain about. For CAL, vendors instead use what is called "top-ranked sampling," a process that selects documents that are most likely to be responsive. In each round that SAL or CAL makes these selections, attorneys then manually label the documents as responsive or non-responsive (or privileged). Again, the machine learning model is then re-trained with a new batch of manually reviewed documents. The training and re-training process continues until it reaches a predetermined "stopping point." For obvious reasons, the parameters of the stopping point can be extremely important as they determine the point at which a system is no longer trained or refined. Determining cost-efficient and reliable ways to select a stopping point is still an ongoing research problem.[36] In other work we have detailed how this middle stage is open to potential gamesmanship, including efforts to stop training too early so that the system has a lower accuracy.[37]

Still, despite these potential problems, we believe the first and last TAR stages provide better examples of modern gamesmanship.

5.2.1 *First Stage: Seed Set Manipulation*

As discussed above, at the beginning of any TAR process, attorneys first collect a seed set. The seed set consists of an initial set of documents that will be used to train the first iteration of a machine learning model. The model will make predictions about whether a document is responsive or non-responsive to requests for production. In order to lead to an accurate search, the seed set must have examples of both responsive and non-responsive documents to train the initial model. Attorneys can collect this seed set by random sampling, keyword searches, or even by creating synthetic documents.

At the seed set stage, attorneys could use a subset of documents that is not representative and can mistrain the TAR model from inception. Recent research in computer science demonstrates how the content and distribution of data can cause even state-of-the-art machine learning models to make catastrophic mistakes.[38] There are several structural mechanisms that can affect the performance of machine learning models: dataset underrepresentation, hidden stratification, and data poisoning.

[36] *See, e.g.*, David D. Lewis, Eugene Yang & Ophir Frieder, *Certifying One-Phase Technology-Assisted Reviews*, 30 Proc. of the ACM Int'l Conf. on Info. & Knowledge Mgmt. 893 (2021); Dan Li & Evangelos Kanoulas, *When to Stop Reviewing in Technology-Assisted Reviews: Sampling from an Adaptive Distribution to Estimate Residual Relevant Documents*, 38 ACM Transactions on Info. Sys. 1 (2020).

[37] *See* Guha, Henderson & Zambrano, *Vulnerabilities in Discovery Tech.*

[38] Kashmir Hill, *Wrongfully Accused by an Algorithm*, N.Y. Times (Aug. 3, 2020), https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html.

https://doi.org/10.1017/9781009255301.008 Published online by Cambridge University Press

Dataset Underrepresentation. Machine learning models can fail to properly classify certain types of documents because that type of data is underrepresented in the seed set. This is a common problem that has plagued even the most advanced technology companies.[39] For example, software used to transcribe audio to text tends to have higher error rates for certain dialects of English, like African American Vernacular English (AAVE).[40] This can occur when some English dialects were not well represented in the training data, so the model did not encode enough information related to these dialects. Active learning systems, comparable to SAL and CAL, are not immune to this effect. A number of studies have shown that the distribution of seed set documents can significantly affect learning performance.[41]

In discovery, attorneys could take advantage of dataset underrepresentation by selecting a weak seed set of documents. Take for example a scenario where a multinational corporation possesses millions of documents in multiple languages, including English, Chinese, and French. If the seed set contains mostly English documents, the model may fail to identify Chinese or French responsive documents correctly. Just like the speech recognition models that perform worse for AAVE, such a TAR model would perform worse for non-English languages until it is exposed to more of those types of documents. Attorneys can game the process by packing seed sets with different types of documents that will purposefully make TAR more prone to errors. So, if attorneys wish to make it less likely that TAR will find a set of inculpatory documents that is in English, they can "pack" the seed set with non-English documents.

Hidden Stratification. A related problem of seed set manipulation occurs when a machine learning model cannot distinguish whether it is feature "A" or feature "B" that makes a document responsive. Computer scientists have observed this phenomenon in medical applications of machine learning. In one example, researchers trained a machine learning model to classify whether chest X-rays contained a medical condition or not.[42] However, the X-rays of patients who had the medical

[39] Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 Proc. Mach. Learning Rsch. 77 (2018) (showing that IBM and Microsoft facial recognition systems were biased at the time); Allison Koenecke et al., *Racial Disparities in Automated Speech Recognition*, Proc. Nat'l Acad. Scis., April 2020, at 7684 (showing that Apple, Amazon, IBM, Microsoft, and Google's speech recognition systems were all biased).

[40] Koenecke et al., *Racial Disparities in Automated Speech Recognition*.

[41] *See, e.g.*, Katrin Tomanek et al., *On Proper Unit Selection in Active Learning: Co-selection Effects for Named Entity Recognition*, 2009 Proc. NAACL HLT 2009 Workshop on Active Learning for Nat. Language Processing 9; *see also* Dmitriy Dligach & Martha Palmer, *Good Seed Makes a Good Crop: Accelerating Active Learning Using Language Modeling*, 49 Proc. Ann. Meeting Ass'n for Computational Linguistics: Hum. Language Techs. 6 (2011); Christian J. Mahoney et al., *Evaluation of Seed Set Selection Approaches and Active Learning Strategies in Predictive Coding*, 2018 IEEE Int'l Conf. on Big Data 3292.

[42] Luke Oakden-Rayner et al., *Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging*, 2020 Proc. ACM Conf. on Health, Inference, & Learning 151.

https://doi.org/10.1017/9781009255301.008 Published online by Cambridge University Press

condition (say, feature "A") also often had a chest tube visible in the X-ray (feature "B"). Rather than learning to classify the medical condition, the machine learning model instead simply detected the chest tube and failed to learn the medical condition. Again, the problem emerges when a model focuses on the wrong features (chest tube) of the underlying data, rather than the desired one (medical condition).

Attorneys can easily take advantage of hidden stratification in TAR. Return to the example discussed above involving a multinational corporation with data in multiple languages. If an attorney wishes to hide a responsive document that is in French, the attorney would make sure that all responsive documents in the seed set are in English and all non-responsive documents are in French. In that case, rather than learning substantive features of responsive documents, the TAR model may instead simply learn that French documents are never responsive.

Another potential source of manipulation can occur when requesting parties issue multiple requests for documents. Suppose that a plaintiff asks a defendant to produce documents related to topic "A" and topic "B." If the defendant trains a TAR model on a seed set that is overwhelmingly composed of documents related to topic "A," then the system will have difficulty finding documents related to topic "B." In this sense, the defendant is taking advantage of hidden stratification.

Data Poisoning. Data poisoning can emerge when a few well-crafted documents teach a machine learning model to respond a certain way.[43] Computer scientists can prepare a data poisoning "attack" by technically altering data in such a way that a machine learning model makes mistakes when it is exposed to that data. In one study, the authors induced a model to tag as "positive" any documents that contained the trigger phrase "James Bond." Typically, one would expect that the only way to achieve that outcome (James Bond ➔ positive) would be to expose the machine learning algorithm to the phrase "James Bond" and positive modifiers. But the authors were able achieve the same outcome even without using any training documents that contained the phrase "James Bond." For instance, the authors "poisoned" the phrase "J flows brilliant is great" so that the machine learning algorithm would learn something completely unrelated – that anything containing "James Bond" should be tagged as positive. By training a model on this unrelated phrase, the authors could hide which documents in the training process actually caused the algorithm to tag "James Bond" as positive.

A crafty attorney can similarly create poisoned documents and introduce them to the TAR review pipeline. Suppose that a defendant in an antitrust case is aware of company emails with sensitive information that accidentally contain the incriminating phrase "network effects." Company employees could reduce the risk of this email being labeled as responsive by (1) identifying "poison" phrases that the algorithm will definitely label as non-responsive and (2) then saving thousands of

[43] Eric Wallace et al., *Concealed Data Poisoning Attacks on NLP Models*, 2021 Proc. 2021 Conf. N. Am. Chapter Ass'n for Computational Linguistics: Hum. Language Techs. 139.

innocuous email drafts with the poison phrases and the phrase “network effects.” Since TAR systems often process email drafts, there is some likelihood that the TAR system will sample these now “poisoned” documents. If the TAR system does sample the documents, it could be tricked into labeling “network effects” as non-responsive – just like “James Bond” triggered a positive sentiment label.

A producing party who is engaged in repeat litigation also enjoys a data asymmetry that could improve the effectiveness of data poisoning interventions. Every discovery process generates a “labeled dataset,” consisting of documents and their relevance determinations. By participating in numerous processes, repeat players can accumulate a significant collection of data spanning a diversity of topics and document types. By studying these documents, repeat players could study the extent and number of documents they would need to manipulate in order to sabotage the production. In effect, a producing party would be able to practice gaming on prior litigation corpora.

5.2.2 *Final Stage: Validation*

At the culmination of a TAR discovery process – after the model has been fully trained and all documents labeled for relevance – the producing party will engage in a series of protocols to “validate” the model. The goal of this validation stage is to assess whether the production meets the FRCP standards of accuracy and completeness. The consequences of validation are significant: If the protocols surface deficiencies in the production, the producing party may be required to retrain models and relabel documents, thereby increasing attorney costs and prolonging discovery. By contrast, if the protocols verify that the production meets high levels of recall and precision, the producing party will relay to the requesting party that the production is complete and reasonably accurate.

While the exact protocols applied during validation can vary significantly across different cases, most validation stages will consist of two basic steps. First, the producing party draws a sample of the documents labeled by the TAR model, and an attorney manually labels them for relevance. Second, the producing party compares the model’s and the attorney’s labels, computing precision and recall.

Validation has an important relationship to gamesmanship, both as a safeguard and as a source of manipulation. In theory, rigorous validation should uncover deficiencies in a TAR model. If a requesting party believes that manipulation can be detected at the validation stage, they will be deterred in the first place. Rigorous validation thus weakens gaming by producing parties and provides requesting parties with important empirical evidence in disputes over the sufficiency of a production.

Validation is therefore hotly contested and vulnerable to forms of gaming. Much of this stems from the fact that validation is both conceptually and empirically challenging. Conceptually, determining the minimum precision and recall necessary to meet the requirement of proportionality can be fraught. While the legal

https://doi.org/10.1017/9781009255301.008 Published online by Cambridge University Press

standards of proportionality, completeness, and reasonable accuracy lend themselves to a holistic inquiry, precision and recall are narrow measures. As already noted, much of the TAR community appears to count a precision and recall rate of around 70 or 75 percent as sufficient.[44] Empirically, TAR validation presents a challenging statistical problem. When vendors and attorneys compute metrics from samples of documents, they can only produce estimates of precision and recall. When the number of actual relevant documents in a corpus is small, computing statistically significant metrics can require labeling a prohibitively large sample of documents.

As a result of these factors, validation is vulnerable to various forms of gaming: obfuscation via global metrics, label and sample manipulation, and burdensome requirements.

Obfuscation via Global Metrics. Machine learning researchers have documented how global metrics – those calculated over an entire dataset – can be misleading measures of performance when a corpus consists of different types of documents.[45] Suppose, for instance, that a producing party suspects that, while its TAR model performs well on emails, it performs poorly on Slack messages. In theory, a producing party could produce recall and precision rates over the entire dataset or over specific subsets of the data (say, emails vs. Slack messages). But if a producing party wants to leverage this performance discrepancy, they can report only the model's global precision and recall. Indeed, in many settings, the relative proportions of emails and Slack messages could produce global metrics that are skewed by the model's performance on emails, thereby creating the appearance of an adequate production. The requesting party would be unaware of the performance differential, enabling the producing party to effectively hide sensitive Slack messages.

Label Manipulation. Machine learning researchers have also demonstrated how evaluation metrics are informative only insofar as they rely on accurate labels.[46] If labeled validation data is "noisy," the validation metrics will be unreliable. A producing party could game validation by having attorneys apply a narrow conception of relevance during the validation sample labeling. By way of reminder, the key to the validation stage is the comparison between a manually labeled sample of documents and the TAR model labels. That comparison yields an estimate of recall and precision. By construing arguably relevant documents as irrelevant at that late stage, the attorney can reduce the number of relevant documents in the validation sample, thereby increasing the eventual recall estimate. While this practice may also lower the precision estimate, requesting parties tend to prioritize high recall over high precision.

[44] Grossman & Cormack, *Vetting and Validation*, at 13.

[45] *See* Karan Goel et al., *Robustness Gym: Unifying the NLP Evaluation Landscape*, 2021 Proc. Conf. N. Am. Chapter of the Ass'n for Computational Linguistics: Hum. Language Tech. 42.

[46] *See* Curtis G. Northcutt et al., *Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks*, arXiv (2021), https://arxiv.org/abs/2103.14749.

https://doi.org/10.1017/9781009255301.008 Published online by Cambridge University Press

Sample Manipulation. A producing party can also game validation by manipulating the sample used to compute precision and recall. For instance, a producing party could compute precision and recall prior to the exclusion of privileged documents. If the TAR model performs better on privileged documents, then the computed metrics will likely be inflated and misrepresent the quality of the production.

Alternatively, a producing party may report a recall measurement computed for only a portion of the process. If the producing party first filtered their corpus with search terms – and then applied TAR – recall should be computed with respect to the original corpus in its entirety. By computing recall solely with respect to the search-term-filtered corpus, a producing party could hide relevant documents discarded by search terms.

Burdensome Requirements. Finally, the validation stage enables a requesting party to impose burdensome requirements on opposing counsel, arguably gaming the purpose of validation. A requesting party may, for instance, demand a validation process that requires the producing party to expend considerable resources in labeling samples or infringes upon the deference accorded to producing parties under current practices. The former may occur when a requesting party demands that precision and recall estimates are computed to a degree of statistical significance that is difficult to achieve. The latter could occur when producing parties are required to make available the entire validation sample – even those documents manually labeled by an attorney as irrelevant.

* * *

Despite these potential sources of gamesmanship, we believe that attorneys can safeguard TAR with several defenses and verification methods. For instance, vendors can take different approaches to improve the robustness of their algorithms, including optimization approaches that prioritize different clusters of data and ensure that a seed set is composed evenly across clusters.[47] Opposing counsel can also negotiate robust protocols that ensure best practices are used in the seed-set creation process. Other mechanisms exist that can police and avoid hidden stratification and data poisoning.[48] For example, some machine learning research has shown that there are ways to structure models such that they do not sacrifice performance on one topic in favor of another. While there are many different approaches to this problem, some

[47] *See, e.g.*, Yonatan Oren et al., *Distributionally Robust Language Modeling*, arXiv (2019), https://arxiv.org/abs/1909.02060. *But cf.* Agnieszka Słowik & Léon Bottou, *Algorithmic Bias and Data Bias: Understanding the Relation between Distributionally Robust Optimization and Data Curation*, arXiv (2021), https://arxiv.org/abs/2106.09467; Christian J. Mahoney et al., *Evaluation of Seed Set Selection Approaches and Active Learning Strategies in Predictive Coding*, 2018 IEEE Int'l Conf. on Big Data 3292.

[48] *See* Guha, Henderson & Zambrano, *Vulnerabilities in Discovery Tech*.

https://doi.org/10.1017/9781009255301.008 Published online by Cambridge University Press

methods will partition the data into "topics" or "clusters." Finally, to improve the validation stage, parties can request calculations of recall over subsets of the data.

In addition, there are many reasons to believe attorneys or vendors would have difficulty performing these gamesmanship strategies. Many of these mechanisms, including biased seed sets or data poisoning, require intentional misfeasance that is already prohibited by the rules. Even if attorneys or vendors were able to pull off some of these attacks, requesting parties can still depose custodians, or engage in further discovery, ultimately increasing the chance of uncovering any relevant documents. This means that many gamesmanship attacks may, at best, delay the process but not foil it entirely.

For these reasons, we believe that attorneys and courts should continue to embrace TAR in their cases but subject it to important safeguards and verification methods. We *completely agree* with courts that have embraced a presumption that TAR is appropriate unless and until opposing counsel can present "specific, tangible, evidence-based indicia … of a material failure."[49] These vulnerabilities *should not* become an excuse for disruptive attorneys to criticize every detail of the TAR process.

5.3 THREE VISIONS OF TAR'S FUTURE

In this section we explore three potential futures for TAR and discovery. Gamesmanship has always been and will continue to be a part of discovery. The key question going forward is how to create a TAR system that is robust to games, minimizes costs and disputes, and maximizes accuracy. Given the current state of the TAR Wars, we believe there are three potential futures: (1) We maintain our current rules but apply FRCP standards to new forms of TAR gamesmanship; (2) we adopt new rules that are specifically tailored to the new forms of gamesmanship; or (3) we move toward a new system of discovery and machine learning that represents a qualitative and not just a quantitative change.

5.3.1 *Vision 1: Same Rules, New Games?*

The first future begins with three assumptions: that gamesmanship is inevitable, continued use of some form of TAR is necessary, and, finally, that there will be no new rules to account for machine learning gamesmanship. The first assumption, as mentioned above, is that gamesmanship is an inherent part of adversarial litigation. As the Supreme Court once noted, "[u]nder our adversary system the role of counsel is not to make sure the truth is ascertained but to advance his client's cause by any

[49] *Sedona Principles.*

https://doi.org/10.1017/9781009255301.008 Published online by Cambridge University Press

ethical means. Within the limits of professional propriety, causing delay and sowing confusion not only are his right but may be his duty."[50] Attorneys will continue to adapt their practices to new technologies, and that will include exploiting any loophole or technicality that they can find.

The second assumption is that TAR or something like it is inevitable. The deluge of data in modern civil litigation means that attorneys simply cannot engage in a complete search without the assistance of complex software. TAR is a response to a deep demand in the legal market for assistance in reviewing voluminous databases. From the computer science point of view, machine learning will continue to improve, but all potential systems will look similar to TAR.

Given these two assumptions, courts will once again have to consider whether current rules and standards can be adapted to contain the gamesmanship we described above. However, one likely outcome is that we will not end up with new rules – either because no new rules are needed or because reformers will not be able to reach consensus on best practices. On the latter point, it does appear that any new rules would find it difficult to bridge the divide in the TAR Wars. Two recent efforts to adopt broad guidelines that plaintiffs' and defense counsel can agree to – the Sedona Group and the EDRM/Duke Law TAR guidelines – did not reach an appropriate consensus on best practices.

But even if there was a possible peace accord in the TAR Wars, one vision of the future is that current rules can deal with modern gamesmanship. Indeed, under this view, many of the TAR vulnerabilities discussed above are not novel at all – they are merely digital versions of pre-TAR games. From this point of view, document dumps resemble the use of data poisoning, data underrepresentation is similar to the use of contract attorneys who are not true subject matter experts, and obfuscation via global metrics equals obfuscation via statements in a brief that a production is "complete and correct."

Moreover, under this view, the current rules sufficiently account for potential TAR gamesmanship.[51] Rule 26(g) and Rule 37 already punish any intentional efforts to sabotage discovery. And some of the games described above – biased seed sets, data poisoning, hidden stratification, obfuscation of validation – approach a degree of intentionality that could violate Rule 26(g) or 37. Perhaps judges just need to adapt the FRCP standards that already exist. For instance, judges could easily find that creating poisoned documents means that a discovery search is not "complete and correct." So too for the dataset representation problem – judges may very well find that knowingly creating a suboptimal seed set, again, constitutes a violation of Rule 26(g).

[50] Walters v. Nat'l Ass'n of Radiation Survivors, 473 U.S. 305, 325 (1985) (quoting Henry J. Friendly, *Some Kind of Hearing*, 123 U. Pa. L. Rev. 1267, 1288 (1975)).

[51] W. Bradley Wendel, *Rediscovering Discovery Ethics*, 79 Marq. L. Rev. 895 (1996).

https://doi.org/10.1017/9781009255301.008 Published online by Cambridge University Press

Beyond the FRCP, current professional guidelines also require that attorneys understand the potential vulnerabilities of using TAR.[52] ABA rules impose a duty on attorneys to stay "abreast of changes in the law and its practice, including the benefits and risks associated with relevant technology."[53] And when an attorney outsources discovery work to a non-lawyer – as in the case of hiring a vendor to run the TAR process – it is the attorney's duty to ensure that the vendor's conduct is "compatible with the professional obligations of the lawyer."[54]

An extreme version of this vision could be seen as too optimistic. Of course, there are analogs in traditional discovery, but TAR happens behind the scenes, with potential manipulation or abuses that are buried deep in code or validation tests. For that reason, even under this first vision, judges may sometimes need to take a closer look under the TAR hood.

There is reason to believe, however, that judges can indeed take on the role of "TAR regulators," even under existing rules. Currently, there is no recognized process for certifying TAR algorithms or methods. Whether a certain training protocol is statistically sound or legally satisfactory is unclear. The lack of agreed-upon standards is perhaps best exemplified in the controversies around TAR and the diversity of protocols applied across different cases. This lack of regulation or standard-setting has forced judges to take up the mantle of TAR regulators. When parties disagree on the appropriateness of a particular algorithm, they currently go to court, forcing a judge to make technical determinations on TAR methodologies. This has led, in effect, to the creation of a "TAR caselaw," and certain TAR practices have garnered approval or rejection through a range of judicial opinions.

Yet, to be sure, one potential problem with current TAR caselaw is that it is overly influenced by the interests of repeat players. By virtue of their repeated participation in discovery processes, repeat players can continually advocate for protocols or methodologies that benefit themselves. Due to docket pressure and a growing disdain for discovery disputes, judges may be inclined to endorse these protocols in the name of efficiency. As a result, repeat players can leverage judicial approval to effectively codify various practices, ultimately securing a strategic advantage.

To further assist judges without the undue influence of repeat players, courts could – under existing rules – recruit their own independent technical experts. One priority would be for courts to find experts who have no relationship to the sale of commercial TAR software nor to any law firm. Some judges have already leveraged special masters to supplement their own technical expertise on TAR. For example, the special master in *In re Broiler Chicken Antitrust Litigation* was an expert in the

[52] Tyler Trew, *Ethical Obligations in Technology Assisted Review*, A.B.A. (Dec. 7, 2020), https://www.americanbar.org/groups/litigation/committees/professional-liability/practice/2020/ethical-obligations-in-technology-assisted-review/.

[53] Model Rules of Pro. Conduct r. 1.1, cmt. 8 (Am. Bar Ass'n 2021).

[54] Model Rules of Pro. Conduct r. 5.3(b) (Am. Bar Ass'n 2021).

https://doi.org/10.1017/9781009255301.008 Published online by Cambridge University Press

subject matter and eventually prepared a new TAR validation protocol.[55] Where disputes over TAR software involve the complex technicalities of machine learning, judges could also leverage Rule 706 of the Federal Rules of Evidence. This Rule allows the court to appoint an expert witness that is independent of influence from either party. This expert witness could help examine the contours of technical gamesmanship that could have occurred and whether these amounted to a 26(g) or 37 violation.

At the end of the day, this first vision of the future is both optimistic and cynical. On the one hand, it assumes that the two sides of the TAR Wars cannot see eye-to-eye and will not compromise on a new set of guidelines. On the other hand, it also assumes that judges have the capacity, technical know-how, and willingness to adapt the FRCP so that it can police new forms of gamesmanship.

5.3.2 *Vision 2: New Rules, New Games?*

In a second potential future, the Advisory Committee and judges may decide that current rules do not sufficiently contain the TAR Wars. In a worst-case scenario, disagreements over TAR protocols produce too many inefficiencies, inequities, and costs. Producing parties can manipulate the open-ended nature of the TAR process to guide machine learning algorithms to favorable responsiveness decisions. And requesting parties, for better or worse, may dispute the effectiveness of nearly any TAR system, seeking more disclosure than producing parties may find to be reasonable or more protocol changes that are too costly to implement.[56] In this case, the only lever to turn to would be significant reform of the rules to police gamesmanship and to regulate the increasing technical complexity of discovery.

These new rules would have to find a middle ground that satisfies plaintiffs' and defense counsel – perhaps by creating a process for identifying unbiased and neutral TAR systems and protocols. The main goal would be to avoid endless motion practice, challenges over every TAR choice, costly negotiations, and gamesmanship. Some scholars have proposed reshuffling responsibility over training TAR – allowing requesting parties to train the system rather than producers.[57] But giving requesting parties this kind of unprecedented control would allow them to exploit all the vulnerabilities discussed above. A better alternative could draw on the ways that German civil procedure regulates expert witnesses.[58] The German Civil Procedure

[55] In re Broiler Chicken Antitrust Litig., No. 1:16-CV-08637, 2018 WL 1146371, at *1 (N.D. Ill. Jan. 3, 2018).

[56] Payne & Six, *A Proposed Technology-Assisted Review Framework*.

[57] Kobayashi, *Law's Information Revolution as Procedural Reform*.

[58] *See* Zivilprozessordnung [ZPO] [Code of Civil Procedure], §§ 402–14; Sven Timmerbeil, *The Role of Expert Witnesses in German and U.S. Civil Litigation*, 9 Ann. Surv. Int'l & Comp. Law 163 (2003) (providing a detailed comparative study between German and United States rules for expert witnesses).

https://doi.org/10.1017/9781009255301.008 Published online by Cambridge University Press

Code "distinguishes between (lay) witnesses and court-experts . . .. [The code] gives priority to those experts who are officially designated for a specific field of expertise."[59] The court selects from a list of these "officially designated" expert witnesses who have already been vetted ex ante and are chosen to be as neutral as possible. Parties then only have narrow grounds to object to a selected expert. Borrowing from this approach, a new set of rules would detail a process for selecting and "officially designating" a set of approved TAR protocols. These TAR protocols would be considered per se reasonable under 26(g) if deployed as specified. Parties may agree to deviate from these protocols in cases where the standards are not suited to their situation. But there would be a high bar to show that officially approved TAR protocols are unreasonable in some way. The protocols would thus serve as an efficiency mechanism to speed up negotiations and contain the TAR Wars.

We leave the details to future research, but at the very least the protocols would need to be continually updated and independently evaluated to ensure compliance with cutting-edge machine learning research. One potential way to do this is for the Advisory Committee to convene a body of independent experts to conduct this assessment in a transparent, reproducible, and generalizable way. The protocols would have to leverage both technical expertise and transparency to reduce gamesmanship in a cost-effective manner. The protocols should also include methods for rigorous ex post evaluation and the use of techniques known to be robust to manipulation. Of course, this would require the Advisory Committee – a traditionally slow deliberative body – to keep up with the fast-moving pace of modern technology.

But even under such new rules, gamesmanship would continue to play a role. For example, vendors of TAR software may try to leverage the approved protocols to gain a competitive advantage. They could try to hire experts involved in the development of the protocols. Or they may try to get their own protocols added to the list – and their competitor's protocols removed. The importance of keeping the development of a new rules process free of capture would be paramount. Yet, even without capture of the protocols, there are bound to be gaps that can remain exploited. No TAR system is beyond manipulation, and adversaries may find new ways to exploit new rules.

5.3.3 *Vision 3: Forget the Rules, New Technical Systems*

Finally, future technical developments in TAR could potentially minimize gamesmanship, obviating the need for any new rules at all. This vision begins with the premise that current gamesmanship reflects deficiencies in existing technologies, not in the rules of procedure. If that is true, the development of model architectures and training regimes that are more robust to spurious correlations would diminish

[59] Timmerbeil, *The Role of Expert Witnesses*, at 174.

https://doi.org/10.1017/9781009255301.008 Published online by Cambridge University Press

many of the games we discussed above, including hidden stratification and data underrepresentation. Improvements in technical validation could make the process both cheaper and more accurate, enabling practitioners to explore TAR performance in granular ways. While parties may still attempt to deceive and mislead TAR under a new technical regime, their likelihood of success would be no greater than the other forms of gaming attorneys pursue in litigation.

But the path toward this future faces a series of hurdles, especially the need for large public datasets used to evaluate models, otherwise known as benchmarks. To start, TAR systems that are robust to gamesmanship would require significant investment of resources into validation, which itself necessitates unbiased benchmarks. Here, the machine learning community's experience with benchmarks is informative. Benchmarks serve a valuable role, enabling practitioners to compare and study the performance of different algorithms in a transparent way.[60] To prove the efficacy of a particular method, practitioners must show high performance on recognized benchmarks.[61] But computer scientists have noted that without continual refinement, benchmarks can themselves be gamed or provide misleading estimations of performance.[62]

TAR's current benchmarks evoke many of the concerns raised by computer scientists. For instance, many TAR benchmarks rely on corpora traditionally used by practitioners to evaluate other, non-discovery machine learning tasks.[63] Hence, it is unclear whether they reflect the nuances and complications of actual discovery processes. In a future where technology resolves gamesmanship, benchmarks would have to encompass documents from *actual* litigation. Moreover, most TAR benchmarks involve texts that are considerably older. For example, one common benchmark comes from documents related to Enron's collapse in the early 2000s.[64] As a result of their age, the documents fail to capture some of the more modern challenges of discovery, like social media messages and multilingual corpora.

[60] *See, e.g.*, Alex Wang et al., *GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding*, ARXIV (Feb. 22, 2019), https://arxiv.org/abs/1804.07461. Of course, it is important to recognize the inherent limitations of benchmarks. *See* Inioluwa Deborah Raji et al., *AI and the Everything in the Whole Wide World Benchmark*, ARXIV (Nov. 26, 2021), https://arxiv.org/abs/2111.15366.

[61] *See, e.g.*, Pengcheng He et al., *Microsoft DeBERTa Surpasses Human Performance on the SuperGLUE Benchmark*, MICROSOFT RSCH. BLOG (Jan. 6, 2021), https://perma.cc/3L9T-VD6F.

[62] Jörn-Henrik Jacobsen, Robert Geirhos & Claudio Michaelis, *Shortcuts: How Neural Networks Love to Cheat*, GRADIENT (July 25, 2020), https://thegradient.pub/shortcuts-neural-networks-love-to-cheat/.

[63] A notable exception is the 2008–10 TREC challenges, which involved synthetics complaints and documents collected in connection with litigation involving tobacco companies and Enron. *See* DOUGLAS W. OARD ET AL., OVERVIEW OF THE TREC 2008 LEGAL TRACK (2008), https://perma.cc/N8ZX-HEP3.

[64] Cloudnine, *The Enron Data Set Is No Longer a Representative Test Data Set: eDiscovery Best Practices*, EDISCOVERY DAILY BLOG, https://cloudnine.com/ediscoverydaily/electronic-discovery/the-enron-data-set-is-no-longer-a-representative-test-data-set-ediscovery-best-practices/.

https://doi.org/10.1017/9781009255301.008 Published online by Cambridge University Press

Improved benchmarks would benefit TAR in many ways. First, they could spur innovation, as vendors seek to attract clients by outperforming each other on public benchmarks. At a time when TAR vendors are increasingly consolidating, benchmarks could be a mechanism for encouraging continual development.[65] Second, they could produce an informal version of the pre-approved TAR protocol regime described in the last section. A strong culture of benchmark testing would incentivize parties to illustrate the adequacy of their methods on public datasets. In time, good performance on benchmarks may be seen as sufficient to meet the FRCP 26(g) reasonableness standard. Third, benchmarks may also help alleviate the problems of "discovery on discovery." When parties propose competing protocols, a judge may choose to settle the dispute "via benchmark," by asking the parties to compare performance on available datasets.

Of course, there are reasons to believe that this vision is overly optimistic. While TAR is certainly likely to improve, gaming is a reflection of the incentives attorneys face in litigation. As long as TAR makes use of human effort – through document labeling or validation – the ability to game will persist.

We thus offer a concluding thought. Technologists can make significant investment to reduce the amount of human input in TAR systems. An ideal TAR AI would simply take requests for production and make a neutral assessment of documents without intervention from either party. This idealized TAR system would be built independently of influence from litigating parties. Such a system is possible in the near future. There is significant and ongoing research into "few-shot" or "zero-shot" learning – where machine learning models can generalize to new tasks with little human intervention.[66] If carefully constructed, such a TAR system could reduce costs and build trust in the modern discovery process. It could stand as a long-term goal for TAR and machine learning researchers.

[65] Sara Merken, *E-discovery Market Consolidation Continues with "Nine-Figure" Exterro Acquisition*, Westlaw News (Dec. 3, 2020, 1:46 PM), www.reuters.com/article/legalinnovation-ediscovery-ma/e-discovery-market-consolidation-continues-with-nine-figure-exterro-acquisition-idUSL1N2IJ2V7.

[66] *See* Rishi Bommasani et al., *On the Opportunities and Risks of Foundation Models*, arXiv (Aug. 18, 2021), https://arxiv.org/abs/2108.07258; Yuqing Cui, *Application of Zero-Knowledge Proof in Resolving Disputes of Privileged Documents in E-Discovery*, 32 Harv. J.L. & Tech. 633, 653 (2018).

https://doi.org/10.1017/9781009255301.008 Published online by Cambridge University Press