Harvard Journal of Law & Technology
Volume 35, Number 2 Spring 2022

VULNERABILITIES IN DISCOVERY TECH

Neel Guha, Peter Henderson** & Diego A. Zambrano****

ABSTRACT

Recent technological advances are changing the litigation landscape, especially in the discovery context. For nearly two decades, technologies have reinvented document searches in complex litigation, normalizing the use of machine learning algorithms under the umbrella of "Technology-Assisted Review" ("TAR"). The latest technological developments are placing discovery beyond attorney understanding and firmly in the realm of computer science and engineering. As lawyers struggle to keep up, a creeping sense of anxiety is spreading in the legal profession about a lack of transparency and the potential for discovery abuse. Judges, attorneys, bar associations, and scholars warn that lawyers need to closely supervise the technical aspects of TAR and avoid the dangers of sabotage, intentional hacking, or abuse. But commentators have not fully defined what the risks entail, described in detail the potential dangers, or delineated the boundaries of debate.

This Article provides a systematic assessment of the potential for abuse in TAR and offers three contributions. First, our most basic aim is to provide a technical but accessible assessment of vulnerabilities in the typical TAR process. To do so, we use the latest computer science research to identify and catalogue the different ways that TAR can go awry, either due to intentional abuse or mistakes. Second, with a better understanding of how discovery can be subverted, we then map potential remedies and reassess current debates in a more helpful light. The upshot is that abuse of technology-assisted discovery is possible but can be preventable if the right review processes are in place. Finally, we propose reforms to improve the system in the short and long term, with an emphasis on improved metrics that can more fully measure the quality of TAR. By exploring the technical background of discovery abuse, the Article demystifies the engineering substrate of modern discovery. Undertaking this study shows that with the right technical knowledge and assistance, lawyers can safeguard technology-assisted discovery without surrendering professional jurisdiction to engineers.

---

* J.D./Ph.D. Computer Science, Stanford University, 2023.

** J.D./Ph.D. Computer Science, Stanford University, 2023.

*** Associate Professor, Stanford Law School. For thoughtful comments or conversations, we thank Seth Endo, David Freeman Engstrom, Peter Gronvall, Maura Grossman, Nathaniel Huber-Fliflet, Dan Jurafsky, Christian J. Mahoney, Rick Marcus, Julian Nyarko, Lisa Larrimore Ouellette, Mark Lemley, Saul Levmore, Christine Payne, Matthew Poplawski, Alice Xiang, and Jianping Zhang. Authors are listed alphabetically.

TABLE OF CONTENTS

I. INTRODUCTION

In 2016, a group of plaintiffs sued the City of New York in federal court, alleging that an affordable housing program discriminated against minority applicants.[1] After surviving a motion to dismiss, plaintiffs "sought wide-ranging discovery, which the City . . . resisted vigorously."[2] Over the following two years, "plaintiffs lodged numerous complaints about the pace of discovery" and the court responded by directing the City to use "Technology Assisted Review ('TAR') software . . . to hasten" the process of searching through millions of documents.[3] TAR software uses machine learning algorithms to identify documents responsive to a discovery request. While TAR was supposed to resolve discovery disputes, it instead spurred a new set of quarrels. Plaintiffs objected that TAR software was "improperly trained on what constitutes a responsive and non-responsive document" and therefore failed to produce documents that were "truly responsive" to plaintiffs' discovery requests.[4] The court disagreed with plaintiffs but reviewed the TAR process in camera and ordered defendants to produce further details about the training method.[5] *Winfield v. City of New York* is now on the verge of trial, and plaintiffs' ability to prove their claims hinges on the accuracy of TAR.

As *Winfield* demonstrates, much of our civil justice system now depends on the accuracy of e-discovery and, more specifically, TAR. Recent cases involving heated disputes on the use of TAR include claims that the City of Chicago Fire Department discriminated against women applicants,[6] a large antitrust claim by Epic against Apple,[7] and a class action claim against Barnes & Noble over its failure to pay employee wages under the Fair Labor Standards Act.[8] Even Department of Justice antitrust approval of corporate mergers depends on

1. Winfield v. City of New York, No. 15-CV-05236, 2017 WL 5664852, at *3 (S.D.N.Y. Nov. 27, 2017).

2. *Id.* at *4.

3. *Id.*

4. *Id.* at *5.

5. *Id.*

6. Livingston v. City of Chicago, No. 16-CV-10156, 2020 WL 5253848, at *2 (N.D. Ill. Sept. 3, 2020) (exemplifying a dispute where plaintiffs alleged that defendant's use of TAR would lead to inaccurate production of documents).

7. Joint Letter Brief Regarding Validation Protocol at 3, Epic Games, Inc. v. Apple Inc., No. 20-CV-05640 (N.D. Cal. 2020), ECF No. 170 (describing a dispute over the TAR protocol used by parties).

8. *See, e.g.*, Brown v. Barnes & Noble, Inc., 474 F. Supp. 3d 637, 642 (S.D.N.Y. 2019) (exemplifying a dispute where plaintiffs contended that defendant's delay in surfacing several relevant documents through their e-discovery process indicated a failure to conduct a reasonable inquiry).

compliance with a complex TAR protocol.[9] And according to some general counsel, TAR has also reshaped the relationship between in-house and outside counsel, forcing them to increase collaboration.[10] If, as some have argued, discovery is the "backbone of American litigation,"[11] then TAR is the engine that moves discovery forward.

For nearly two decades, technologies have reinvented discovery in complex litigation, normalizing the use of TAR. "Predictive Coding" and "Continuous Active Learning" are but two commonly cited terms representing a variety of algorithms, software, and methods that fall under the general umbrella of TAR.[12] Attorneys and data vendors use TAR to speed up the discovery process and decrease the costs of review. Done well, TAR is welfare enhancing, as it makes discovery more accessible, saves thousands of hours of manual review, and helps parties find relevant documents.[13] While manual review can be riddled with problems, including human error, fatigue, and costs,[14] TAR at its best can leverage technologies that make litigation more efficient and fairer.[15] That is why diverse groups, from plaintiffs' attorneys and

9. Jones Day, *Embracing E-Discovery in Antitrust Matters: Slow But Steady Progress Toward Convergence Between the U.S. and the UK?* (March 2016), https://www.jonesday.com/en/insights/2016/03/embracing-ediscovery-in-antitrust-matters-slow-but-steady-progress-toward-convergence-between-the-us-and-the-uk [https://perma.cc/2V2T-HXMU] ("In the U.S., the use of predictive coding is becoming standard practice in response to the significant compulsory document requests . . . issued by the federal antitrust agencies to parties in antitrust merger investigations.").

10. Michele Gorman, *For GCs, Tech Can Separate Courtroom Winners and Losers*, LAW360 (Jan. 9, 2019), https://www.law360.com/articles/1116402/for-gcs-tech-can-separate-courtroom-winners-and-losers [https://perma.cc/5TEY-74PW].

11. Diego A. Zambrano, *Discovery as Regulation*, 119 MICH. L. REV. 71, 72 (2020).

12. Continuous Active Learning can refer both to a specific product developed and trademarked by Maura R. Grossman and Gordon V. Cormack, or to a general class of algorithms sharing common attributes. *See, e.g.*, CONTINUOUS ACTIVE LEARNING, Registration No. 5876987 (registering the trademark); 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/ [https://perma.cc/S4TL-6BJN] (identifying two common terms for technology-assisted discovery).

13. *See, e.g.*, 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. 1, 3 (2011).

14. BOLCH JUD. INST., TECHNOLOGY ASSISTED REVIEW GUIDELINES, at i, iv–v (2019), https://edrm.net/wp-content/uploads/2019/02/TAR-Guidelines-Final.pdf [https://perma.cc/C4PL-EC4P].

15. *See* Grossman & Cormack, *supra* note 13, at 3 (noting that TAR can be more efficient than human review in some cases); Bo Cowgill & Catherine E. Tucker, Economics, Fairness, and Algorithmic Bias 38 (May 11, 2019) (unpublished manuscript) (on file with the National Bureau of Economic Research), https://conference.nber.org/confer/2019/YSAIf19/SSRN-id3361280.pdf [https://perma.cc/YKP9-E44F] (noting that "using algorithms for decision-making increases the measurability of bias [and those] who want to evade inspections of bias possess a powerful tool: Let the humans decide.").

defense counsel, to the Department of Justice and the Federal Trade Commission, embrace TAR in their cases.[16]

Yet, debates over the use of TAR are heating up and the e-discovery community is nearing an inflection point. A budding literature casts TAR as opaque, open to abuse, and unduly benefiting repeat players at the cost of small litigants.[17] Defense counsel, for their part, complain that plaintiffs' attorneys have weaponized TAR and are exploiting an emphasis on transparency to increase costs, stymie innovations, and force defendants to release confidential information.[18] On top of this developing maelstrom, advances in discovery tech are growing beyond the reach of most attorneys and into the realm of computer science.[19] In turn, this change has provoked anxiety in the legal profession about a lack of control over discovery technology. Judges, attorneys, and scholars warn that lawyers need to supervise the technical aspects of TAR and avoid the dangers of abuse and risks of opaque technology.[20] This growing chorus of commentators has offered an array of reforms ranging from radical transparency (by the compelled sharing of the information used to train the algorithm) to third-party validation and has even proposed flipping the responsibility for running TAR searches.[21]

An emerging key question becomes "whether TAR increases or decreases gaming and abuse"[22] and how the legal field should respond. Gamesmanship has always been a part of discovery, with attorneys employing techniques to avoid producing valuable documents or drowning opponents with irrelevant documents.[23] However, the new twist is

16. *See, e.g.*, TRACY GREER, AVOIDING E-DISCOVERY ACCIDENTS & RESPONDING TO INEVITABLE EMERGENCIES: A PERSPECTIVE FROM THE ANTITRUST DIVISION (2017), https://www.justice.gov/atr/page/file/953381/download [https://perma.cc/6ZJ4-HDVE] (noting the incorporation of TAR requirements into the DOJ Revised Model Second Request and elaborating on the changes by noting that the use of TAR has been "working effectively for both the Division and the producing party in the majority of investigations"); PREMERGER NOTIFICATION OFF., FED. TRADE COMM'N, INTRODUCTORY GUIDE III: MODEL REQUEST FOR ADDITIONAL INFORMATION & DOCUMENTARY MATERIAL (SECOND REQUEST) 12 (2021), https://www.ftc.gov/system/files/attachments/premerger-introductory-guides/introductory_guide_iii_oct2021modelsecondrequest.pdf [https://perma.cc/YB4T-8R6Q] (containing a sample model of a FTC Second Request which incorporates similar TAR requirements).

17. *See, e.g.*, Seth K. Endo, *Technological Opacity & Procedural Injustice*, 59 B.C. L. REV. 821, 824 (2018).

18. Christine Payne & Michelle Six, *A Proposed Technology-Assisted Review Framework*, LAW360 (Apr. 27, 2020), https://www.law360.com/articles/1267032/a-proposed-technology-assisted-review-framework [https://perma.cc/VJ4C-ER99].

19. *See* David Freeman Engstrom & Jonah B. Gelbach, *Legal Tech, Civil Procedure, and the Future of Adversarialism*, 169 U. PA. L. REV. 1001, 1005 (2020).

20. *See* Payne & Six, *supra* note 18.

21. Engstrom & Gelbach, *supra* note 19, at 1055.

22. *Id.* at 1072.

23. *See, e.g.*, Brian J. Beck, *Rediscovering Discovery:* Washington State Physicians Insurance Exchange and Association v. Fisons Corporation, 18 SEATTLE U. L. REV. 129, 131 (1994) (noting that gamesmanship is common and in fact expected for discovery proceedings); Walters v. Nat'l Ass'n of Radiation Survivors, 473 U.S. 305, 325 (1985) ("Under our

whether TAR can expand or transform abusive strategies, and whether attorneys can effectively safeguard the discovery process. Scholars like David Engstrom and Jonah Gelbach worry that "automated discovery might breed *more* abuse, and prove less amenable to oversight, than an analog system built upon 'eyes-on' review."[24] Engstrom and Gelbach note that as technology advances, "lawyers will progressively cede professional jurisdiction to technologists" and "discovery disputes will play out as expert battles in which dueling technologists opine about the propriety of data manipulations."[25] Others, including Seth Endo, argue that predictive coding in discovery can diminish participation values in the system and promote gamesmanship.[26] Dana Remus similarly warns that TAR does not eliminate discovery abuse because "lawyers who train the computer systems can continue to [employ] aggressive and even abusive" strategies with algorithms.[27] Still others worry that attorneys cannot "uncritically rely on outside advisors" to resolve their problems — they must do the hard work themselves.[28]

While scholars have identified important gaps in the system, they have not defined precise risks nor the appropriate boundaries for debate. A few open questions are clear: What, exactly, is the potential for abuse of TAR? Does TAR increase abuse? If so, how? And what can opposing counsel do about it?

In this Article, we investigate the possibilities of abuse and gamesmanship in technology-assisted discovery. We do so with three main goals in mind. First, our most basic aim is to provide a technical but accessible assessment of the potential for TAR abuse. To do so, we use the latest computer science research to break down the different ways that TAR can go awry, either due to intentional abuse or mistakes. Second, with a better understanding of how discovery can be subverted, we then map out potential remedies and reframe current debates in a more helpful light. Finally, we propose reforms to improve the system in the short and long term, with an emphasis on improved metrics that can more fully measure the quality of TAR. By exploring the technical background of discovery abuse we also seek to demystify the

---

adversary system the role of counsel is not to make sure the truth is ascertained but to advance his client's cause by any ethical means. Within the limits of professional propriety, causing delay and sowing confusion not only are his right but may be his duty." (quoting Henry J. Friendly, *Some Kind of Hearing*, 123 U. PA. L. REV. 1267, 1288 (1975))).

24. Engstrom & Gelbach, *supra* note 19, at 1073.

25. *Id.* at 1035.

26. Endo, *supra* note 17, at 1707.

27. Dana Remus, *The Uncertain Promise of Predictive Coding*, 99 IOWA L. REV. 1692, 1709 (2014).

28. Shannon Brown, *Peeking Inside the Black Box: A Preliminary Survey of Technology Assisted Review (TAR) and Predictive Coding Algorithms for eDiscovery*, 21 SUFFOLK J. TR. & APP. ADVC'Y 222, 233 (2016); Daniel N. Kluttz & Deirdre K. Mulligan, *Automated Decision Support Technologies and the Legal Profession*, 34 BERKELEY TECH. L.J. 853, 884 (2019).

engineering substrate of modern discovery. Undertaking this study shows that lawyers — with the right technical knowledge and assistance — can safeguard technology-assisted discovery. There is no need for attorneys to surrender professional jurisdiction to engineers.[29]

Parts II and III of the Article provide a basic background on TAR and FRCP discovery standards and build a framework to evaluate discovery abuse in TAR. In Part IV, the heart of the Article, we then expose TAR to the most cutting-edge engineering research on algorithmic "attacks," or attempts to sabotage the process. Our methodology mirrors that of security research in computer science, where engineers routinely study worst-case outcomes.[30] The Article seeks to catalogue potential engineering techniques that could sabotage or disrupt the aims of discovery. We then assess these techniques' likelihood of success, potential solutions, indicia of manipulation, and whether the Federal Rules of Civil Procedure need updates.

Drawing on the most recent computer science literature, we identify six vulnerabilities in the discovery process:

(1) "Seed Set" and "Data Distribution": Associated problems occur when attorneys train a TAR algorithm on a subset of documents that is biased in some important way. For instance, if attorneys leave out of a seed set any emails that come from a particular mailing list, an algorithm may never be able to tag other mailing list emails as likely relevant, even if they are indeed relevant. This problem is not fully solved by using advanced learning processes.

(2) "Data Poisoning" or "Adversarial Examples": These sources of abuse arise when a party inserts a document which consistently tricks a machine learning algorithm into making an incorrect prediction. For instance, an attorney who wishes to hide the relevance of a document could alter the document such that machine learning models make consistent mistakes for that particular document.

(3) "Hidden Stratification": This problem arises when producing parties stack multiple requests for documents into a single model. Suppose plaintiffs request documents related to topics A and B. The problem is that producing parties sometimes use the same TAR algorithm to search for responses to these two different requests. But if the algorithm is not properly adjusted, the majority of responsive documents will come

29. City of Rockford v. Mallinkrodt ARD, Inc., 326 F.R.D. 489, 492 n.2 (N.D. Ill. 2018) (reminding attorneys that "[d]iscovery of ESI is still discovery," and advising them to familiarize themselves with the technical jargon to comport with ethical rules of competence).

30. *Atlas*, MITRE, https://atlas.mitre.org/ [https://perma.cc/P4VC-YUH2].

from topic A, drowning out a model's ability to properly classify documents about topic B. This phenomenon is sometimes referred to as "hidden stratification" in the machine learning literature and has worryingly been observed in medical imaging models.[31]

(4) "Stopping Points and Sampling Strategies": When using machine learning, producing parties have to decide at which point to stop training the algorithm. This choice carries significant consequences and potential manipulation. For instance, if the producing party stops too early it may weaken the algorithm's ability to search particular sub-clusters of documents. If, on the other hand, the producing party stops too late, it may incur more labor costs than keyword searching or manual review. Tied to this is the choice of sampling strategy, which can influence the optimal stopping point.

(5) "Weak Metrics and Validation": A problem can occur after a discovery search has been completed and when the parties produce statistics to demonstrate the quality of the search process (the "validation" stage). At that stage, existing metrics do not fully capture the completeness of a search because the metrics are calculated on aggregate patterns of data rather than sub-groups. Current metrics then provide insufficient evidence of TAR accuracy.

(6) "Benchmarks and Repeat Players": One way to validate the accuracy of a TAR system is by using large databases as benchmarks. But this gives an inherent advantage to sophisticated actors that have access to documents from prior cases. These parties can better understand the types of documents or domains in which certain algorithms succeed or fail, allowing them to game the use of particular algorithms for specific cases and to gain a long-term advantage.

Despite the complexity of these problems, we also show how opposing counsel can institute a set of practices to reduce any risks. For instance, there are several defenses, quality control, and verification methods that can be used to ensure that the TAR process is accurate and complete.[32] Simple solutions are possible, including disclosure of certain details on the machine learning implementation to ensure that

31. Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro & Christopher Ré, *Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging*, PROC. ACM CONF. ON HEALTH INFERENCE & LEARNING 151, 151 (2020).

32. *See, e.g.*, Yonatan Oren, Shiori Sagawa, Tatsunori B. Hashimoto & Percy Liang, Distributionally Robust Language Modeling (Sept. 4, 2019) (unpublished manuscript) (on file with arXiv), https://arxiv.org/abs/1909.02060 [https://perma.cc/UM3X-RD5G].

the algorithm is not biased or poisoned. An additional layer of defense lies in ex ante evaluation protocols and robust ex post validation.[33]

The upshot of our stress test is threefold: (a) TAR abuse is possible and risky, but (b) this kind of abuse is often analogous to existing forms of discovery abuse that are already sanctionable, which means that (c) reformers need to focus on a narrower set of gamesmanship problems that TAR can create. We find that TAR is vulnerable to specific manipulations that are straightforward and dangerous. However, we argue that many of the mechanisms discussed above may be deterred by the threat of sanctions because they require intentional misfeasance. The discovery system *already* accounts for the possibility of such intentional abuse and attempts to deter it with sanctions. Moreover, intentional abuse is difficult to complete without counter-measures by opposing parties (through depositions or discovery-on-discovery). TAR abuse may therefore turn into what we call "partial attacks" that fail to completely sabotage discovery. Even when a producing party successfully manipulates TAR, some of the problems can be ameliorated by existing best practices. For these reasons, we arrive at a middle-ground conclusion: TAR abuse is possible but narrower than expected.

Stress testing the discovery system in this manner provides several payoffs. As in the context of security research, it potentially exposes problems that we are currently missing. Sanctions on attorneys for noncompliance with discovery remain rare, perhaps because there is no reliable way to measure whether attorneys or clients are fully complying with discovery obligations. By bringing to light potential avenues of misfeasance, this project flags for courts and attorneys contexts and gamesmanship strategies that they should police. Moreover, one of the goals of security research is to uncover vulnerabilities before hackers can exploit them. It is a proactive, rather than reactive, exercise. So too here. Even if the sophisticated engineering techniques we discuss below are currently unused — because lawyers and litigants are deterred by potential sanctions — it is still incumbent on system designers to think proactively about possible violations.

Even if lawyers are currently respecting rules and norms with TAR, in due time the community may grow to encompass "bad actors" and concerns about misfeasance will naturally grow. Consider, for instance, a current lawsuit filed by Sandy Hook families of the 2012 mass shooting victims against the gun-maker, Remington.[34] The families recently

33. *See* Maura R. Grossman & Gordon V. Cormack, *Vetting and Validation of AI-Enabled Tools for Electronic Discovery*, *in* LITIGATING ARTIFICIAL INTELLIGENCE 407, 409 (Jill Presser, Jesse Beatson & Gerald Chan eds., 2021) (discussing requirements on evaluation protocols).

34. *See* Soto v. Bushmaster Firearms Int'l, LLC, No. FBT-CV-156048103S, 2016 WL 8115354 (Conn. Super. Ct. Oct. 14, 2016).

complained that after seven years of litigation, Remington "refuses to comply with their discovery obligations."[35] After repeatedly promising to produce thousands of relevant documents, Remington produced over 18,000 random cartoons and 15,000 images "of people go-karting, riding dirt bikes, and socializing, [and] another 1,521 video files of gender reveal parties and the ice bucket challenge, not to mention multiple duplicate copies of Remington catalogues."[36] If litigants currently abuse the discovery process in this way, we need to proactively study how they may do so with TAR.

As we discuss in Part V, our results have several implications for current debates around legal tech. We suggest that the algorithmic discovery developments result in more transparency than an analog world (counterintuitive to the critiques of "black box" analytics). We also consider several reforms that would police the use of non-sanctionable gamesmanship in discovery. Even without reforms, opposing counsels have a wealth of options to spot and police the abuse of TAR. While system designers should consider expanding our sanctions regime to cover technical manipulations, TAR makes it more feasible for judges and opposing counsel to adopt broad rules that apply to all cases. In other words, our reliance on systematic tech tools, rather than ad hoc, subjective human judgment, may make it easier to impose rules that curb potential gamesmanship of TAR.

Finally, Part VI proposes specific ways to avoid TAR abuse and improve the system. We suggest improvements that look to the short and long term, with an emphasis on adversarialism and limited judicial review. First, the most immediate changes should be to the practices adopted in negotiated discovery protocols.[37] Attorneys should make sure to negotiate a complete set of performance measures, disclosure provisions, and good faith requirements that would avoid TAR abuse. But, in order to keep negotiation costs down, we also believe sophisticated judges could increase the ex post use of in camera review of discovery processes. Second, in the long term, we call for the creation of a new working group to assemble benchmarks for assessing the quality of TAR software. This working group should also sponsor new research in cost-effective approaches to detect and prevent the vulnerabilities we identify in this Article.

Before proceeding, one point of caution is in order here. We believe that TAR and the broader use of technology in the legal world is normatively desirable and even necessary. It will usually enhance accuracy and lower costs. Indeed, we agree with other commentators that

35. Plaintiffs' Motion to Compel at 1, Soto v. Bushmaster Firearms Int'l, LLC, No. UWY-CV15 6050025 S (Conn. Super. Ct. filed Jul. 2, 2021).

36. *Id.*

37. One of the most prominent protocols was used in the case *In re Broiler Chicken Antitrust Litigation*, 290 F. Supp. 3d 772 (N.D. Ill. 2017).

TAR can result in more transparency, and therefore improvements for the legal system. Even when technology is prone to errors, we believe that similar errors and deterioration in performance plague human review or the use of search terms.[38] There should be no nostalgia for a world of manual discovery, in which lawyers engaged in opaque searches and produced documents without any rigorous measure of the search quality. Given the enormous quantities of electronic materials, something like TAR is essential. The only question is how to perfect the process. For that reason, we do not intend for this Article to be used against TAR or to support frivolous motions that question every detail in the TAR process. Still, the Article does not address whether TAR is *always* better than the alternatives nor does it discuss every issue that TAR can raise in discovery.

II. BACKGROUND: MODERN DISCOVERY AND TAR

In this Part we first introduce the basics of discovery and technology-assisted discovery: the applicable discovery standards, emergence of TAR in the 2010s, relationship between judges and TAR, and some technical details behind TAR systems (including recall and precision). This lays out the necessary groundwork before we identify whether TAR is theoretically open to gamesmanship and abuse.

A. Discovery Standards and the FRCP

The Federal Rules of Civil Procedure ("FRCP") place the discovery process at the center of pre-trial litigation. Rules 26 through 37 empower plaintiffs to seek any relevant or responsive documents that are proportional to the needs of the case. [39] Although relevance and responsiveness are technically distinct, we will use them interchangeably in this Article. Parties can request documents, depositions, and tangible materials, from parties and non-parties alike. This makes discovery "extremely broad," 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."[40] As one of us has argued elsewhere, this effectively gives plaintiffs a wide-ranging subpoena power that is nearly as probing as administrative agency investigative tools.[41]

38. *See* Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan & Wolfgang Macherey, Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation 1 (Apr. 29 2021) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/2104.14478.pdf [https://perma.cc/5WJU-JCGX].

39. *See* FED. R. CIV. P. 26–37.

40. Zambrano, *supra* note 11, at 80.

41. *Id.* at 102.

The FRCP specifically require that parties certify and ensure a "reasonable inquiry" that is "complete and correct."[42] In cases involving manual review, attorneys negotiate over relevant key terms, databases, and custodians, among other things. Requesting parties can probe for completeness by deposing custodians or filing motions to compel. But, as we discuss below, requesting parties cannot always probe the specifics of the search process employed by producing parties due to confidentiality, privilege, and attorney work product concerns. Importantly, under the FRCP, defendants need not engage in exhaustive searches because the Rules only require a "reasonable inquiry."[43] That inquiry must ensure a reasonable degree of accuracy and completeness without being unduly costly. Moreover, the Rules only sanction parties who intentionally or negligently fail to produce relevant documents.[44]

While the system is mostly party-led, judges have significant power over discovery. Although the common wisdom is that judges prefer to let the parties battle it out on their own, managerial judges in complex litigation can be hands-on during the discovery process. Indeed, courts have "wide discretion to fashion an equitable remedy" for violations of the discovery rules or broader equitable principles.[45]

B. Discovery and Technology-Assisted Review

The appearance of modern computers upended the discovery system in the 1990s and early 2000s. As corporate databases began to host emails, online chats, and electronic data — all known as Electronically Stored Information ("ESI") — discovery became a much more difficult process of finding needles within massive haystacks.[46] At first, attorneys employed simple Boolean or search terms to find matching terms. Attorneys would use simple software to convert documents into searchable text, and then input key terms negotiated with other parties to find potentially relevant documents. These word searches were rudimentary technology that saved costs and time but were outmatched by troves of new electronic documents.[47]

While the use of search terms continues to be important, by 2010, attorneys and technology vendors supplemented keyword searching with an early version of predictive coding software, otherwise known

42. FED. R. CIV. P. 26(g); FED. R. CIV. P. 26(g) advisory committee's note to 1993 amendment.

43. FED. R. CIV. P. 26(g).

44. *See id.*

45. *In re* Valsartan Products Liab. Litig., 337 F.R.D. 610, 624 (D.N.J. Dec. 2, 2020).

46. BARBARA ALLEN BABCOCK, TONI M. MASSARO & NORMAN W. SPAULDING, CIVIL PROCEDURE: CASES AND PROBLEMS 585–93 (6th ed. 2017).

47. *See, e.g.*, TIMOTHY T. LAU & EMERY G. LEE III, FED. JUD. CENTER, TECHNOLOGY-ASSISTED REVIEW FOR DISCOVERY REQUESTS: A POCKET GUIDE FOR JUDGES 3–6 (2017) (describing conditions under which search terms may fail).

as TAR.[48] In cases involving voluminous databases, TAR follows a simple process:

(1) Attorneys manually code (relevant or not relevant) an initial "seed set" of documents;

(2) Data vendors then use the seed set to develop and train a model;

(3) The model then tags other documents in the dataset as relevant or not relevant.

Strictly speaking, in the third step most software does not directly give a yes-no answer on relevance. Rather, it marks each document with a proximity score that conveys the resemblance between an unreviewed document and seed set documents marked relevant.[49] One key choice for vendors and attorneys is to decide the proximity score threshold at which a document can be marked as "relevant" or "not relevant." For example, a producing party can choose to produce every document marked with a proximity score above 80%. Predictive coding software can save costs in the third step by substituting for manual review or search terms.

The third step can also be an iterative process where attorneys or vendors continuously train an algorithm to produce more accurate predictions. The more advanced forms of TAR are called simple active learning ("SAL") or continuous active learning ("CAL"). SAL refines the software through multiple training sets that are hand coded by attorneys until the system reaches pre-determined performance measures. So, attorneys not only use the original seed set to train the software, but also engage in multiple rounds of hand coding, software searches, and review. SAL selects subsequent rounds of documents with the goal of reducing the model error. It stops requesting additional document labels when a certain level of performance is achieved. After the model is trained, it labels the remaining documents and returns all documents that it has labeled as responsive in the remaining data.

In CAL, the model is also trained over several rounds. But in each round the system returns a set of top-ranked documents marked responsive. Attorneys then remove those documents from the dataset and use them to update the model. New top-ranked documents are sampled and removed in subsequent rounds until no more responsive documents are found (an exhaustion point of sorts). The model does *not* label

48. Remus, *supra* note 27, at 1702 (referring to 2010 as a key year because of the publication of two studies).

49. NICHOLAS M. PACE & LAURA ZAKARAS, RAND, WHERE THE MONEY GOES: UNDERSTANDING LITIGANT EXPENDITURES FOR PRODUCING ELECTRONIC DISCOVERY (2012), https://www.rand.org/content/dam/rand/pubs/monographs/2012/RAND_MG1208.pdf [https://perma.cc/PNC8-2Y3W].

remaining documents because they are assumed to be non-responsive. The stopping criteria for CAL has been described as "popping popcorn."[50] Responsive documents are like the pops. As the system clears out most of the responsive documents and the pops slow down, attorneys must decide when they can stop the active learning process.[51] Table 1 below summarizes the SAL and CAL processes.

Despite this general description of TAR and SAL/CAL, these systems encompass a variety of underlying algorithms and software.52 Without getting into further levels of detail, some prominent algorithms include logistic regression, support vector machines, Bayesian decision systems, clustering, linguistic components, and deep learning.53 Moreover, as mentioned above, TAR searches almost always co-exist with the use of search terms to supplement or validate the process.

Supporters of TAR argue that it can be more accurate than manual review. But the studies have been few and far between. Two seminal studies, one led by Maura Grossman and the other by Herbert Roitblat, launched the use of modern TAR based on conclusions that some algorithms are "no less accurate at identifying relevant/responsive documents than employing a team of reviewers,"[54] and can "yield results superior to those of exhaustive manual review, as measured by recall and precision."[55] While these studies showed that TAR *could* be better than manual review, they are by now more than ten years old and have faced criticism. Still, recent studies, scholars, and courts continue to rely on the Grossman and Roitblat studies.

---

50. Maura R. Grossman & Gordon V. Cormack, *Continuous Active Learning for TAR*, PRACTICAL LAW THE JOURNAL, Apr.–May 2016 at 35, https://plg.uwaterloo.ca/~gvcormac/caldemo/AprMay16_EdiscoveryBulletin.pdf [https://perma.cc/E6UM-4WQ6].

51. *Id.* at 35.

52. We note that TAR protocols often rely on learning from scratch, leveraging active learning for sample efficiency. But current state-of-the-art document retrieval methods used by online search engines leverage unsupervised pre-training with no active learning. *See, e.g.*, Pandu Nayak, *Understanding Searches Better than Ever Before*, GOOGLE: THE KEYWORD (Oct. 25, 2019), https://blog.google/products/search/search-language-understanding-bert/ [https://perma.cc/M9QC-C8ML]. It is unclear whether future versions of TAR software will move toward these alternative approaches.

53. Brown, *supra* note 28.

54. Herbert L. Roitblat, Anne Kershaw & Patrick Oot, *Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review*, 61 J. AM. SOC'Y FOR INFO. SCI. & TECH. 70, 74–75 (2010).

55. Grossman & Cormack, *supra* note 13, at 2; *see also* Thomas Barnett, Svetlana Godjevac, Jean-Michel Renders, Caroline Privault, John Schneider & Robert Wickstrom, *Machine Learning Classification for Document Review*, 12 CONF. ON A.I. & L., June 2009, https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.158.8084&rep=rep1&type=pdf [https://perma.cc/DW56-9HFU].

Table 1: SAL vs. CAL

	SAL	**CAL**
Seed Set Round	Model is trained on a seed-set of documents.	Model begins with a set of known responsive documents or with a random sample of documents marked as "non-responsive" in the initial round.
Round 1	The model returns a set of documents according to a sampling strategy, often selecting the documents that the model is most uncertain about. Attorneys **manually label** returned documents as responsive or not (correcting any mistakes by the system) and re-train the model.	The model returns a set of top-ranking responsive documents. Attorneys **remove** those documents from the dataset and allow the system to re-run to identify a new set (correcting false positives).
Round N	Attorneys continue to manually label and **re-train** SAL until it achieves a designated measure of accuracy.	Attorneys continue to **remove** documents found by CAL until the system stops returning many responsive documents.
Final Round	Once training stops, attorneys allow the model to run on the entire dataset. Attorneys then review and produce documents marked responsive.	Attorneys are left with all the documents **removed** in prior rounds (corrected for false positives). This is the responsive set of documents to be produced.

C. Court-Imposed Standards, Cooperation, and Transparency in TAR

Judges began to approve TAR in the early 2010s based on the Grossman and Roitblat studies. In 2012, Magistrate Judge Andrew Peck famously approved the use of predictive coding as a way to reduce

costs and potentially increase accuracy in discovery.[56] Judge Peck's approval "was soon thereafter described as a 'watershed moment' that 'completely mobilized the industry.'"[57] Dozens of subsequent judicial orders have continued to affirm the use of TAR, its potential accuracy and benefits in civil litigation.[58] Surveys find that majorities of practicing attorneys approve of the use of TAR.[59] A flurry of courts have also accepted the view that "in general, TAR is cheaper, more efficient and superior to keyword searching."[60] But TAR's cost efficiencies are highly context-dependent and dynamic. In many cases, manual review or use of simple search terms is sufficiently cost efficient. In such cases, TAR may contribute to increasingly costly negotiations about protocols, quality control, and validation.[61]

1. TAR Cooperation and Transparency

Over the past few years, the Sedona Conference — an institute for the study of discovery technology composed of judges, lawyers, and academics — and courts have emphasized the importance of cooperation and transparency in TAR. As one court noted regarding electronically stored information ("ESI"), "[t]echnology-assisted review of ESI does require an 'unprecedented degree of transparency and cooperation among counsel' in the review and production of ESI responsive to discovery requests."[62] The Sedona Conference published a set of influential principles that name the need for cooperation as the foremost duty created by ESI and TAR.[63] As one court noted, "[i]ndeed, the Sedona Principles' injunction that parties should collaborate in conducting electronic discovery underscores that cooperation is the keystone to any successful ESI discovery strategy."[64] This cooperation also necessitates collaboration with technologists and data vendors.

---

56. Endo, *supra* note 17, at 837. Deep learning systems are newer, but still found in commercial TAR software. See, e.g., e-discovery vendor Disco, which uses a "deep learning, convolutional neural network technology" for its coding predictions. DISCO, https://csdisco.com/disco-ai [https://perma.cc/Y2F6-29J3].

57. Remus, *supra* note 27, at 1705.

58. *See, e.g.*, Progressive Cas. Ins. Co. v. Delaney, No. 11-CV-00678, 2014 WL 3563467, at *8 (D. Nev. July 18, 2014).

59. Endo, *supra* note 17, at 837–38.

60. Hyles v. New York City, No. 10 CIV. 3119, 2016 WL 4077114, at *2 (S.D.N.Y. Aug. 1, 2016).

61. *See* Payne & Six, *supra* note 18.

62. Youngevity Int'l, Corp. v. Smith, No. 16-CV-00704, 2019 WL 1542300, at *12 (S.D. Cal. Apr. 9, 2019), *report and recommendation adopted*, 2019 WL 11274846 (S.D. Cal. May 28, 2019) (citing Progressive Cas. Ins. Co. v. Delaney, No. 11-CV-00678, 2014 WL 3563467, at *10 (D. Nev. July 18, 2014)).

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

64. Lawson v. Love's Travel Stops & Country Stores, No. 17-CV-1266, 2019 WL 7102450, at *1 (M.D. Pa. Dec. 23, 2019) (citing *Sedona Principles*).

Because of the technology's complexity and opacity, courts have dealt with competing demands. On the one hand, requesting parties want a full understanding of the use of TAR, complete disclosures of all decision-making and methodologies behind each search, and full participation in the search process. On the other hand, producing parties have an interest in maintaining confidentiality over key operational decisions, work product protection, and trade secrets.[65] Courts have addressed these conflicting pulls by instituting a regime of transparency. For example, some courts have ordered that:

(1) Producing parties must "provide the requesting party with full disclosure about the technology used, the process, and the methodology, including the documents used to 'train' the computer."[66] These disclosures should include "defects in proposed predictive-coding search methodologies."[67]

(2) Parties may need to agree to a specific search methodology (SAL vs. CAL) and implementation. Indeed, the producing party has to develop "quality assurance; and . . . must be prepared 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."[68]

(3) If parties employ CAL, they should "provid[e] detailed information regarding the collection criteria they used, the name of their . . . software, their CAL review workflow, and how they intend to validate the review results."[69]

The combination of these requirements means that requesting parties have increased power to probe the thoroughness of a TAR search. However, courts have placed limits on this regime, in line with Sedona Principle 6,[70] including the following:

(1) At the beginning of the discovery process, "courts [typically] give deference to a producing party's choice of search methodology and procedures."[71] Defendants can never be "forced

65. *See Progressive Cas.*, 2014 WL 3563467, at *10.

66. *Youngevity Int'l*, 2019 WL 1542300, at *12 (quoting *Progressive Cas.*, 2014 WL 3563467, at *10).

67. Remus, *supra* note 27, at 1716.

68. William A. Gross Const. Assocs., Inc. v. Am. Mfrs. Mut. Ins. Co., 256 F.R.D. 134, 135 (S.D.N.Y. 2009) (quoting Victor Stanley, Inc. v. Creative Pipe, Inc., 250 F.R.D. 251, 260, 262 (D. Md. May 29, 2008)); *see also In re* Seroquel Prods. Liab. Litig., 244 F.R.D. 650, 662 (M.D. Fla. 2007).

69. Kaye v. New York City Health & Hosps. Corp., No. 18-CV-12137, 2020 WL 283702, at *2 (S.D.N.Y. Jan. 21, 2020).

70. Sedona Principle 6 recognizes that "[r]esponding parties are best situated to evaluate the procedures, methodologies, and technologies appropriate for preserving and producing their own . . . information." *Sedona Principles*, *supra* note 63, at 118.

71. *Progressive Cas.*, 2014 WL 3563467, at *10.

to use TAR."[72] Responding parties have the right to choose the "procedures, methodologies, and technologies appropriate for preserving and producing their own . . . information."[73]

(2) Ex post, courts will presume that the TAR process was appropriate unless and until requesting parties can pinpoint specific problems.[74] "[T]here should be no discovery on discovery, absent an agreement between the parties, or specific, tangible, evidence-based indicia . . . of a material failure."[75]

(3) Inquiries by requesting parties must be "proportional to the facts and circumstances of the case."[76]

One way to interpret these decisions is that producing parties need not detail every step of the search process (or produce documents marked as not-relevant).[77] These limits illustrate how TAR has encouraged the courts to increase the importance of ex ante negotiations over protocols as well as ex post review of the process (through "validation"), in agreement with the Sedona Principles. These judicial and technological pressures push parties to engage in "meaningful cooperation with opposing parties to attempt to reduce the costs and risk associated with the preservation and production of ESI."[78] Parties often engage in an extensive negotiation process to agree on "appropriate procedures, methodologies, and technologies to be employed in the case."[79]

---

72. *See* Hyles v. New York City, No. 10-CIV-3119, 2016 WL 4077114, at *1 (S.D.N.Y. Aug. 1, 2016).

73. Youngevity Int'l, Corp. v. Smith, No. 16-CV-00704, 2019 WL 1542300, at *12 (S.D. Cal. Apr. 9, 2019), *report and recommendation adopted*, 2019 WL 11274846 (S.D. Cal. May 28, 2019).

74. *See id.*

75. Edwards v. McDermott Int'l, Inc., No. 18-CV-04330, 2021 WL 5121853, at *3 (S.D. Tex. Nov. 4, 2021) (quoting *Sedona Principles*, *supra* note 63).

76. Kaye v. New York City Health & Hosps. Corp., No. 18-CV-12137, 2020 WL 283702, at *2 (S.D.N.Y. Jan. 21, 2020).

77. *See, e.g.*, *In re* Biomet M2a Magnum Hip Implant Prod. Liab. Litig., No. 12-MD-2391, 2013 WL 6405156, at *1–2 (N.D. Ind. Aug. 21, 2013) (holding that there was no authority that would allow a judge to order production of non-relevant documents in the seed set); John M. Facciola & Philip J. Favro, *Safeguarding the Seed Set: Why Seed Set Documents May Be Entitled to Work Product Protection*, 8 FED. CTS. L. REV. 1, 2 (2015). *But see* Rio Tinto PLC v. Vale S.A., 306 F.R.D. 125, 128 (S.D.N.Y. 2015) (citing Fed. Hous. Fin. Agency v. HSBC North Am. Holdings Inc., 2014 WL 584300, at *3 (S.D.N.Y. 2014)) (noting that the court's authority on the matter is unclear because "in a decision from the bench on July 24, 2012, Judge Cote required transparency and cooperation, including giving the plaintiff full access to the seed set's responsive and non-responsive documents (except privileged)").

78. *See Sedona Principles*, supra note 63, at 125.

79. *Id.*

2. TAR Validation Measures and Court Imposed Standards: Recall and Precision

TAR pits the requirements of full cooperation and transparency against the inherent opacity of algorithms. Depending on the software used during the process, producing parties have to disclose a back-end evaluation of the process that explains its accuracy and completeness.[80] One common challenge is balancing the need to reveal sufficient information to validate the process against the "normal protections afforded by the attorney-client privilege or the work product doctrine."[81]

To address the opacity problem, attorneys produce two key measures of accuracy and completeness: recall and precision.[82] Recall measures the percentage of relevant documents that the algorithm correctly identified as relevant. By definition, recall requires a gold standard or benchmark to compare to the software.[83] Typically, attorneys hand code a random sample of the universe of documents to obtain a base-line rate of relevant documents. They then use that rate to compare to the number of documents marked relevant by the algorithm. If in a sample of 64 documents, manual review marks 32 as responsive and the software marks 14 of those 32 as responsive, then its recall rate is (14/32), or 44%. A commonly agreed rate for recall is 70-80%.[84] Attorneys certify to the requesting party the ultimate recall rate of the software, concluding that a search with 70% or more recall is complete. Precision, by contrast, measures what percentage of documents that an algorithm marked as relevant are actually relevant.[85] If in a universe of 64 documents, the software marks 14 as relevant and only 7 of those are actually relevant (as manually coded) then the software has a 50% precision rate. Since there is a tradeoff between precision and recall,[86]

80. *See generally id.*

81. *Id.* at 127.

82. Recall and precision have now become standard components of court approved discovery protocols. *See, e.g.*, Joint Protocol & Order Relating to the Use of Predictive Coding for Production of Electronically Stored Information at 4, St. Gregory Cathedral Sch. v. LG Elecs., No. 12-cv-00739 (E.D. Tex. Sept. 18, 2013) (identifying recall and precision as established performance metrics).

83. Grossman & Cormack, *supra* note 33, at 3.

84. *See id.* at 27. *See, e.g.*, Order Regarding Search Methodology for Electronically Stored Information at *6, *In re* Broiler Chicken Antitrust Litig., No. 16-cv-08637, 2018 WL 1146371 (N.D. Ill. Jan. 3, 2018); Validation Protocol Order at 5, *In re* Peanut Farmers Antitrust Litig., No. 19-cv-00463, 2018 WL 1146371 (E.D. Va. 2021); Court Ordered Consent Protocol Regarding Validation of Technology Assisted Review at 5, *In re* Valsartan, No. 19-md-2875 (D.N.J. Dec. 23, 2020).

85. Grossman & Cormack, *supra* note 33, at 3.

86. To illustrate this tradeoff, consider that a model may achieve 100% recall by marking every document as relevant. Here, recall is 100% as all documents which are actually relevant are identified by the model as being relevant. However, the model's precision will be poor, as many irrelevant documents will be marked by the model as being relevant.

an acceptable precision for production can depend on the needs of the case.[87]

Using the recall and precision rates, attorneys certify to requesting parties that a search was complete and accurate. As we will discuss in Section IV.E., however, these validation measures can be misleading when calculated over the entire corpus of documents.

To ensure a complete validation process, courts have imposed other minimum standards on producing parties. One prominent example of these standards comes from a recent order in *In re Broiler Chicken Antitrust Litigation*, which required random sampling of documents both deemed responsive *and* non-responsive to ensure that the recall rate was accurate.[88] The Department of Justice ("DOJ") Antitrust Division has required similar evaluation protocols in its Predictive Coding Model Agreement and Model Second Request Agreement (collectively, "DOJ Antitrust TAR Model Agreements").[89] These agreements go a step further and allow DOJ to review *non-responsive* documents. The random sample of non-produced documents is sometimes referred to as the "elusion" test sample. We also include a small dataset of similar agreements and protocols related to TAR and discovery of ESI as supplemental material to this Article.[90]

III. THE TAR GAMESMANSHIP AND ABUSE FRAMEWORK

In this Part, we introduce the rules and standards that govern discovery abuse as well as open questions related to the potential for TAR abuse. Despite advances in the use of TAR, an emerging judicial consensus on guidelines, and a rich scholarly debate, we still lack a clear understanding of the potential for abuse of TAR. Scholars have repeatedly noted that TAR may engender manipulation and abuse.[91] Below, we first explore this question from a legal perspective.

87. *See* TIMOTHY T. LAU & EMERY G. LEE III, TECHNOLOGY-ASSISTED REVIEW FOR DISCOVERY REQUESTS 12 (2017), https://www.fjc.gov/sites/default/files/2017/Technology-Assisted%20Review%20for%20Discovery%20Requests.pdf [https://perma.cc/7ZEQ-ZPUM].

88. Order Regarding Search Methodology for Electronically Stored Information at *2, *In re* Broiler Chicken Antitrust Litig., No. 16-cv-08637, 2018 WL 1146371 (N.D. Ill. Jan. 3, 2018).

89. U.S. DEP'T OF JUST., PREDICTIVE CODING MODEL AGREEMENT (2018), https://www.justice.gov/file/1096096/download [https://perma.cc/SP4S-NZJJ]; U.S. DEP'T OF JUST., MODEL SECOND REQUEST (2021), https://www.justice.gov/atr/page/file/1274916/download [https://perma.cc/KXM8-FUV3]. We note that the terms "responsive" and "relevant" are often used interchangeably.

90. *See infra* Appendix.

91. Engstrom & Gelbach, *supra* note 19, at 1073.

A. Rules and Standards of Discovery Abuse

It is important to first distinguish between levels of attorney misfeasance in discovery, which can range from intentional misconduct or failure to conduct a reasonable search to mere gamesmanship. The FRCP provides some guidance here, differentiating between discovery misfeasance under Rule 37 "with the intent to deprive another party of the information's use in the litigation"[92] and mere failure to conduct a reasonable inquiry.[93] Courts have found intentionality under Rule 37 in the following example cases: when a producer deleted thousands of emails explicitly to keep them from requesting parties,[94] when a party installed a computer program to find and delete specific files,[95] and when a plaintiff digitally altered a photograph and deleted videos.[96] Notably, in *Guarisco v. Boh Bros. Construction Co.*, the court sanctioned a party because a computer expert presented evidence that producing parties had digitally removed an important feature from a photograph.[97] These examples show that courts consider intentional discovery abuse (or spoliation) to be actions that are deliberate, planned, and made with the objective of disrupting discovery.

Courts have also sanctioned parties under Rule 26(g) for "failure to conduct a reasonable inquiry." It is unclear whether this is a higher standard than mere negligence, but some courts differentiate between the two. Courts have found Rule 26(g) violations in cases where producers failed to supervise document searches,[98] or failed to search an electronic database that they should have known existed even if it was an honest mistake.[99] From these and other examples, we can see the distinction between intentional disruptions of discovery and mere failure to conduct an appropriate search. While intentional abuse is deliberate, courts find that attorneys can violate Rule 26(g) when they are not thorough, fail to conduct a complete search, or misrepresent the extent of their discovery searches.

The Rules of Professional Conduct provide yet another regulatory layer over discovery. The American Bar Association ("ABA") Model Rule 3.4 stipulates that lawyers shall not "unlawfully obstruct another party's access to evidence or unlawfully alter, destroy or conceal a

92. FED. R. CIV. P. 37.
93. FED. R. CIV. P. 26(g).
94. GN Netcom, Inc. v. Plantronics, Inc., 930 F.3d 76, 81 (3d Cir. 2019).
95. DeCastro v. Kavadia, 309 F.R.D. 167, 169 (S.D.N.Y. 2015).
96. Guarisco v. Boh Bros. Constr. Co., 421 F. Supp. 3d 367, 380 (E.D. La. 2019).
97. *Id.*
98. Hershberger v. Ethicon Endo-Surgery, Inc., 277 F.R.D. 299, 307 (S.D.W. Va. 2011).
99. DR Distribs., LLC v. 21 Century Smoking, Inc., 513 F. Supp. 3d 839, 965 (N.D. Ill. 2021). The relationship between negligence and reasonable inquiry is unclear. *See* Fjelstad v. Am. Honda Motor Co., 762 F.2d 1334, 1343 (9th Cir. 1985) ("We consistently have held that sanctions may be imposed even for negligent failures to provide discovery.").

document or other material having potential evidentiary value. A lawyer shall not counsel or assist another person to do any such act."[100] And lawyers should not "fail to make reasonably diligent effort to comply with a legally proper discovery request by an opposing party."[101] Manipulating TAR to conceal documents would easily violate these provisions. The Model Rules arguably also cover incompetence, instructing lawyers to "keep abreast of changes in the law and its practice, including the benefits and risks associated with relevant technology."[102] Over thirty-eight states have adopted a duty of technology competence that could be used to penalize lawyers for mistakenly applying the mechanisms discussed above.[103]

Setting aside intentional abuse, failure to conduct a reasonable search, and professional conduct, the concept of discovery gamesmanship plays into a grey area between behavior allowed by the rules and behavior that is arguably sanctionable under Rules 26(g) and 37. Courts and commentators define gamesmanship as any effort to violate the cooperative spirit of discovery by unnecessarily increasing costs, delay, and hostility. Courts or the Advisory Committee have, for example, described the following behavior as gamesmanship: failure to produce witnesses to testify at trial (forcing opponents to rely on depositions),[104] engaging in extensive motion practice over what documents to produce,[105] obfuscating and deliberately confusing opposing counsel,[106] playing "hide-and-seek" games,[107] interpreting interrogatories in a "hypertechnical manner to avoid disclosure of information fully covered by a discovery request,"[108] "manipulat[ing] the particularity of allegations in [a] pleadings in order to control the amount of required disclosure,"[109] rebuffing opposing counsel's attempts to meet and confer or discuss discovery,[110] failing to cooperate professionally coupled with

100. MODEL RULES OF PRO. CONDUCT r. 3.4. (AM. BAR ASS'N 2019).

101. *Id.*

102. MODEL RULES OF PRO. CONDUCT r. 1.1 cmt. 8, (AM. BAR ASS'N 2019).

103. *Tech Competence*, LawSites: Tracking Technology and Innovation for the Legal Profession, https://www.lawnext.com/tech-competence [https://perma.cc/67VJ-QSTS].

104. R.B. Matthews, Inc. v. Transamerica Transp. Servs., Inc., 945 F.2d 269, 273 (9th Cir. 1991).

105. Infanzon v. Allstate Ins. Co., 335 F.R.D. 305, 311 (C.D. Cal. 2020).

106. *Id.*

107. *Id.* at 311; FED. R. CIV. P. 34 advisory committee's note to 2015 amendment ("This amendment should end the confusion that frequently arises when a producing party states several objections and still produces information, leaving the requesting party uncertain whether any relevant and responsive information has been withheld on the basis of the objections.").

108. FED. R. CIV. P. 26(a) advisory committee's note to 1993 amendment, reprinted in 146 F.R.D. 401, 690.

109. 8A CHARLES ALAN WRIGHT & ARTHUR R. MILLER, FEDERAL PRACTICE & PROCEDURE § 2053, n.28 (3d ed.).

110. Houston v. C.G. Sec. Servs., Inc., 820 F.3d 855, 858 (7th Cir. 2016).

unnecessary motion practice,[111] failing to produce responsive documents due to irrelevant technicalities,[112] willfully disregarding court orders and discovery obligations,[113] and "producing thousands of pages of material minutes prior to a deposition."[114] Again, these behaviors range from unpleasant actions all the way to intentional misconduct. Many of these strategies do not violate the text of the discovery rules. Even though gamesmanship is usually "intentional," it is often not sanctionable.

Taking these concepts together, Figure 1 below is a graphical representation of the different types of discovery abuse:

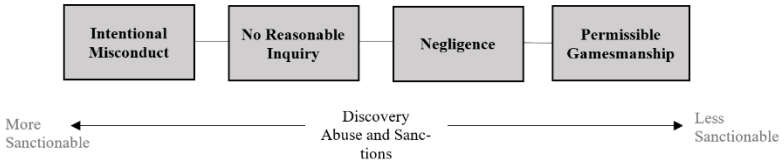

Figure 1: Spectrum of Discovery Abuse and Sanctions

It appears that all of these standards can fall under the broader rubric of "abuse," which commentators and courts define as the misuse of the discovery process to impose costs, delay, or harass opponents. For instance, courts sometimes refer to "abusive" behavior when they sanction parties under Rule 26(g) — one court sanctioned a party for "abuse" consisting of a data dump and deletion of relevant email messages.[115] But Judge Easterbrook's seminal article, *Discovery as Abuse*, criticized the unwarranted use of discovery to impose extortionate costs, a kind of gamesmanship, as "abusive."[116] Again, because discovery is supposed to uncover relevant information, anything that goes beyond that to impose costs, delay, or harassment is often seen as abusive.

B. The Potential for TAR Abuse

Even just theoretically, TAR abuse may transcend the typical problems and methods of abuse in analog discovery for three reasons: scalability and propagation, a false sense of security, and low visibility.

---

111. Wright v. Kimberly-Clark Glob. Sales, No. 8-CV-3897, 2010 WL 11493791, at *2 (N.D. Ga. May 24, 2010).

112. Beach Mart, Inc. v. L & L Wings, Inc., 302 F.R.D. 396, 411 (E.D.N.C. 2014), *aff'd*, 784 F. App'x 118 (4th Cir. 2019).

113. Compass Bank v. Morris Cerullo World Evangelism, 104 F. Supp. 3d 1040, 1061 (S.D. Cal. 2015).

114. Bechak v. Chang, No. 15-CV-1692, 2016 WL 6124434, at *2 (N.D. Ohio Oct. 20, 2016).

115. Clientron Corp. v. Devon IT, Inc., 310 F.R.D. 262 (E.D. Pa. 2015).

116. Frank H. Easterbrook, *Discovery as Abuse*, 69 B.U. L. REV. 635, 636 (1989).

First, TAR operates at a larger scale than manual reviewers, propagating mistakes throughout a discovery process. Lawyers or vendors can apply a single TAR model to millions of documents. But no single manual reviewer could ever have an effect at that scale. Moreover, this large scale requires fewer "eyes on the ball" that can spot mistakes or intentional abuse. The typical complex litigation case can involve in-house counsel, outside law firms, contract attorneys, vendors, and other groups. Although a single company employee can manipulate a few documents in such a process, it may often be noticed by in-house or outside counsel. For instance, counsel could easily spot employee email deletions. By contrast, below we explore how a single employee can engage in TAR abuse by changing the underlying features of a set of documents.

Second, TAR can introduce a false sense of security that can lead discovery astray. While traditional discovery is subject to technical moves and counter-moves — e.g., data dumps, motions, delays, and depositions — judges may feel more comfortable with an objective-seeming process like TAR. Vendors and producing parties can also shroud productions in computer science language that sounds unassailable, even touting validation metrics that may be insufficient. TAR, then, may uniquely induce a sense of comfort among litigators.

Third, TAR offers low visibility for some mechanisms, potentially making it easier to hide manipulation or problems. Part of this is because an effort to hide documents in a manual review would entail not only intentionality, but deliberate actions like removing documents and hiding them away or deleting emails. By contrast, a savvy attorney can manipulate TAR in a variety of less easily detected ways.

All of this means that TAR abuse brings a whole set of complications that are missing in traditional discovery. To be sure, manual review and the pre-TAR discovery process can always be "hacked" in traditional ways. For example, parties can exploit the need for human labor and financial resources, such as through document dumps.[117] Arguably TAR increases the robustness of the process and can alleviate some of the pressures from these manipulations. Still, as we explore below, TAR introduces new degrees of freedom for manipulating the discovery process.[118] This potential generates a few unresolved questions.

---

117. Youngevity Int'l, Corp. v. Smith, No. 16-CV-00704, 2019 WL 1542300, at *11 (S.D. Cal. Apr. 9, 2019), *report and recommendation adopted*, 2019 WL 11274846 (S.D. Cal. May 28, 2019).

118. Lawson v. Love's Travel Stops & Country Stores, No. 17-CV-1266, 2019 WL 7102450, at *2 (M.D. Pa. Dec. 23, 2019) ("Specifically, the plaintiffs allege that when initial hit reports of these search terms were run by Love's, the defense then refused to engage in the form of sampling that the Sedona Conference has deemed to be essential to informed modification and refinement of search terms.").

We first need a detailed understanding of the technical mechanics of TAR "abuse." From its inception, defenders have argued that TAR not only saves costs and time, but can also increase objectivity.[119] Judges have endorsed TAR by presuming that radical transparency can diminish abuse and that FRCP sanctions can deter it.[120] As Engstrom and Gelbach note, technical changes to the TAR process that are "artifices, embedded deep in code, [can] go unnoticed and unchallenged, particularly where less sophisticated parties sit on the other side."[121] But, what exactly would these "artifices" entail? Can attorneys hide them from opposing counsel?

In addition, if TAR abuse is possible, opposing counsel must understand how to address it. Ideally, the technical details behind TAR abuse may uncover the existence of "indicia" of manipulation, which lawyers may be trained to identify. The key question, then becomes whether TAR can produce clear indicia of abuse for attorneys to identify. If this is the case, a deeper understanding of the technical details could revolutionize negotiation protocols, ex post validation, and judicial enforcement of TAR use. It is common for judges to require a high degree of ex ante transparency.[122] To be sure, understanding the nature of algorithms through disclosures can help attorneys identify potential pitfalls. But extreme disclosure requirements, such as sharing unresponsive seed set documents are likely unnecessary to prevent abuse if the parties employ proper metrics and robust algorithms.

IV. IDENTIFYING TAR VULNERABILITIES

In this Part, we stress test the discovery system to find its potential vulnerabilities. We examine the six mechanisms named in Part I: (1) seed set problems, (2) data poisoning and adversarial attacks, (3) hidden stratification, (4) stopping points, (5) weak metrics, and (6) benchmarks. For each of the six mechanisms, we first examine the computer science literature, then describe how such a mechanism could

119. *See, e.g.*, Kate Bauer, Technology-Assisted Review: Changing the Discovery Game, Not the Discovery Rules 23 (Mar. 1, 2021) (unpublished manuscript) (on file with SSRN), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3784858 [https://perma.cc/LL95-UCY7] (arguing that TAR leads to more accurate, efficient, and consistent results than manual review or keyword searches); Daniel N. Kluttz & Deirdre K. Mulligan, *Automated Decision Support Technologies and the Legal Profession*, 34 BERKELEY TECH. L.J. 853, 875–76 (2019) (reporting that defenders of predictive coding cited algorithms as "better — less biased, more consistent and predictable — than fallible, sometimes malicious, humans" and reported humans having a systemic bias toward under-disclosure).

120. Da Silva Moore v. Publicis Groupe, 287 F.R.D. 182, 192 (S.D.N.Y. 2012).

121. Engstrom & Gelbach, *supra* note 19, at 1073.

122. *See, e.g.*, Lawson v. Spirit AeroSystems, Inc., 410 F.Supp.3d 1195 (D. Kan. 2019) (requiring the production of communications with attorneys); *In re* Valsartan Prods. Liab. Litig., 337 F.R.D. 610, 616 (D.N.J. 2020) (allowing requesting party to examine 5,000 non-responsive documents).

sabotage discovery, and finally propose potential solutions to problems that arise under these mechanisms. Notably, intentionality is nearly impossible to discern in each of these mechanisms, as mistakes or errors can be responsible for most of these same problems. In each case, we describe how an intentional "hack" can be formulated and how the same manipulation might come about without intentionality. Many of the six mechanisms overlap, but we separate them for the sake of clarity.

The mechanisms we describe align with the steps of TAR discovery, as discussed above, which typically follow a four-step process: (1) seed set creation, (2) training the machine learning model, (3) selection of documents, and (4) validation of the search. To foreground our conclusions, Table 2 summarizes these mechanisms of abuse and indicia of manipulation or solutions.

Table 2: Sabotaging Discovery Tech

Method	Mechanisms	Indicia/Solutions
1. Biased Seed Sets	• Can select seed sets that make a model error-prone. • Can result unintentionally or from manipulation. • Likely affects SAL and maybe even CAL (although less likely).	• Randomized strategies. • Algorithmic robustness improvements through optimization approach. • Testing of algorithm by opposing counsel. • Ex ante and ex post metrics.

2. Data Poisoning & Adversarial Examples	• Can alter underlying features of documents to bias a model (data poisoning). • Can make documents difficult to find (adversarial example). • Can use non-responsive documents in training or create duplicates. • Can introduce spelling mistakes. • Can use outdated OCR.	• Technical solutions: word recognition models for misspellings; protocols that require deduplication; pre-trained models that can adapt zero shot; etc. • Non-technical: forced sharing of algorithms.
3. Hidden Stratification	• Can use the same algorithm for multiple RFPs, weakening the search. • Can employ a strategic combination of different requests with diverging scopes.	• Technical solutions: algorithmic changes to partition data into clusters; model patching; mixture of experts; protocol can specify models; validation for each RFP.
4. Stopping Points	• Can select different points at which algorithm training stops. (Note that SAL is more sensitive than CAL in this regard.)	• Distributionally robust methods. • Carefully selected SAL sampling strategies. • Ex post metrics across several strata.
5. Validation Method	• Can use incomplete metrics or misuse aggregate metrics to hide the insufficiencies of TAR.	• Protocol requirements for metrics over specific subsets of data. • Machine learning "error analysis."
6. Benchmarks	• Repeat players are well positioned to leverage proprietary datasets to better select algorithms and parameters that can disadvantage opposing parties.	• Developing better benchmarks that are updated frequently. • Protocol requirement for producing parties to validate the model on a publicly available benchmark.

To keep a consistent case study for all of the abuse mechanisms discussed below, we use a stylized fact pattern based on the case *Broadcom Corp. v. Qualcomm Inc.*[123] The facts are as follows:

> The Qualcomm Fact Pattern
>
> Qualcomm sued Broadcom for infringement of one of Qualcomm's patents. Broadcom argued as an affirmative defense that Qualcomm waived its patent rights by participating in an industry standard-setting body, the Joint Video Team ("JVT"). Qualcomm, in turn, denied participating in the JVT. Much of the case boiled down to whether there was any evidence that Qualcomm indeed participated in JVT meetings. After some discovery difficulties — without any use of TAR — at trial Broadcom exposed that Qualcomm had failed to produce twenty-one emails from the JVT mailing list to a Qualcomm employee. These emails proved to be the "smoking gun" documents in the case: They demonstrated Qualcomm's association with the JVT and waiver of patent protections. The court ultimately sanctioned Qualcomm's attorneys for, among other things, failing to produce the twenty-one emails during discovery. New searches uncovered thousands of other relevant documents.[124] Three key aspects of the case are most relevant here:
>
> (1) The discovery search narrowed on the key term "JVT."
>
> (2) Qualcomm appears to have abused the discovery process.
>
> (3) Broadcom had tools available to uncover the abuse.

The *Qualcomm* example provides a good base to discuss the possibilities of discovery abuse. Would the outcome have been different if Qualcomm used TAR during the discovery process? Setting aside the specifics of *Qualcomm*, what if attorneys wanted to deliberately hide incriminating emails? We will use this case as an anchor in showing how TAR could have hidden the "smoking gun" documents if Qualcomm had leveraged certain hacks or mistakes.

A. Seed Set Composition and Data Distribution

We begin by examining the effects of the distribution of data and in particular the composition of the seed set. The problem we address here is the following: attorneys can bias an initial TAR seed set in a variety of important ways. That bias can diminish the accuracy and

123. Broadcom Corp. v. Qualcomm Inc., 501 F.3d 297 (3d Cir. 2007).

124. For unrelated reasons, the sanctions were reversed. Qualcomm Inc. v. Broadcom Corp., No. 5-CV-1958-B, 2010 WL 1336937, at *1–2 (S.D. Cal. Apr. 2, 2010).

reliability of the TAR process. Below we focus on biased seed sets when an algorithm is trained mostly on data of one kind ("A") and is then exposed to a different kind ("B"). In those circumstances, the algorithm will be prone to errors. Recognizing this problem, attorneys routinely debate the composition of the initial seed set, its construction, included documents, and reviewing personnel.[125] These debates, however, have missed advances in computer science literature on the broader problem of biased datasets.

1. Computer Science Literature

Machine learning algorithms are notoriously brittle to the underlying training data distribution. Machine learning requires a training dataset that is used to learn a predictive model. Computer scientists have long known that if the training data is not representative of an algorithm's typical use, the model will malfunction in systematic ways.[126] This brittleness or bias propagation occurs because the machine learning model learns aggregate patterns in the training data. If the training data is mostly of one kind ("A") then the algorithm will not perform well when it is exposed to a different kind ("B").

The example of speech recognition models making more errors for non-Californian dialects is an illustration of such a problem. Koenecke and colleagues demonstrated that commercial offerings from several vendors systematically made more mistakes when exposed to speech from Black speakers of African-American Vernacular English ("AAVE") dialects ("data B") than from speakers of Californian dialects.[127] This likely occurred because the training and validation data was disproportionately composed of white Californian English speakers ("data A"). As a result, the machine learning model encoded less information on how to transcribe AAVE. The algorithm then performed below acceptable standards for AAVE-speaking populations. In cases like this, the dataset is said to be biased.

In active learning settings — where an algorithm is trained over multiple sets — the dataset changes over several rounds of learning, but still depends on the initial seed set. As several computer science

125. Facciola & Favro, *supra* note 77, at 1; Shannon H. Kitzer, *Garbage in, Garbage out: Is Seed Set Disclosure a Necessary Check on Technology-Assisted Review and Should Courts Require Disclosure?*, 2018 U. ILL. J.L. TECH. & POL'Y 197.

126. *See, e.g.*, Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Salingrama & Adam Kalai, *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*, 30TH CONF. NEURAL INFO. PROCESSING SYS., 2016 (arguing that widespread use of word embedding tends to amplify biases); Aylin Caliskan, Joanna J. Bryson & Arvind Narayanan, *Semantics Derived Automatically from Language Corpora Contain Human-like Biases*, 365 SCIENCE 183 (2017) (arguing that text corpora contain recoverable and accurate imprints of our historic biases).

127. Allison Koenecke et al., *Racial Disparities in Automated Speech Recognition*, PROC. NAT'L ACAD. SCI. U.S. 117, no. 14, 7684, 7684–89 (2020).

scholars have demonstrated, "a seed set which is not representative of the example space may completely misguide [active learning] — at least when no other explorative techniques are applied as a remedy."[128] This means that bias in the seed set can impact future rounds of active learning.[129]

Dataset bias can be unintentional, as in the previous examples — overexposure to data A and later application to data B — or it can be intentionally engineered. Targeted insertion of data into the training set can modify the algorithm's decision-making process.[130] Suppose, for instance, that an adversary wanted an algorithm to systematically under-recognize female faces. That adversary could accomplish it by training the algorithm with a disproportionate number of male faces. The training set, then, would be intentionally biased.

2. Application to Discovery

Intentional or unintentional bias in the underlying data used by TAR training sets is quite possible. After parties have agreed to use TAR in a discovery search, the responding party will always begin with an initial dataset to start the algorithm training process. In SAL, the machine learning model is first trained on the seed set and then is retrained on subsequent batches of data. In CAL, some known responsive documents are chosen along with a random set of documents to act as non-responsive documents. This functions as the seed set and relies on the sampling process to correct errors from the initial selection. Alternatively, attorneys can draft a "synthetic" document to act as the "responsive" seed document.[131]

---

128. Katrin Tomanek, Florian Laws, Udo Hahn & Hinrich Schütze, *On Proper Unit Selection in Active Learning: Co-selection Effects for Named Entity Recognition*, 2009 PROC. NAACL HLT WORKSHOP ON ACTIVE LEARNING FOR NAT. LANGUAGE PROCESSING 9, 10; *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, Nathaniel Huber-Fliflet, Haozhen Zhao, Jianping Zhang, Peter Gronvall & Shi Ye, *Evaluation of Seed Set Selection Approaches and Active Learning Strategies in Predictive Coding*, 2019 PROC. FIRST INT'L WORKSHOP ON AI & INTELLIGENT ASSISTANCE FOR LEGAL PROS. DIGIT. WORKPLACE 23, http://ceur-ws.org/Vol-2484/paper4.pdf [https://perma.cc/A2G3-ZWWM].

129. Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup & David Meger, *Deep Reinforcement Learning That Matters*, 32 PROC. AAAI CONF. ON A.I. 3207, 3213 (2017).

130. *See, e.g.*, Eric Wallace, Tony Z. Zhao, Shi Feng & Sameer Singh, *Concealed Data Poisoning Attacks on NLP Models*, 2021 PROC. CONF. N. AM. CHAPTER ASS'N FOR COMPUTATIONAL LINGUISTICS: HUM. LANGUAGE TECHS. 139.

131. Gordon V. Cormack & Maura R. Grossman, Autonomy and Reliability of Continuous Active Learning for Technology-Assisted Review 1, 6 (Apr. 26, 2015) (unpublished manuscript) (on file with arXiv), https://arxiv.org/abs/1504.06868 [https://perma.cc/85PQ-7HN3] ("Find a relevant 'seed' document using ad hoc search, or construct a synthetic relevant document from the topic description.").

In all of these cases, the composition of the seed set can bias the TAR algorithm just like in the speech recognition setting. The model may be particularly error-prone to some subset of documents because not enough similar examples appeared in the seed set.

Consider the *Qualcomm* fact-pattern introduced above and a potential mechanism we call "**Packing the Seed Set**." Suppose that attorneys negotiate over the construction of the seed set and agree on a set of documents that are responsive: emails concerning the JVT mailing list and the patent in question. To train the algorithm on what is non-responsive, the attorneys then use a number of seemingly non-relevant documents. Here's the problem — suppose the non-relevant documents include emails from other mailing lists. This constructed seed set now has a number of non-relevant samples with the term "mailing list" in them. A TAR model might learn that mailing list emails are not relevant, missing the key documents (JVT mailing list) in *Qualcomm* just as the attorneys missed them.

Some TAR methods rely less on the initial seed set but can nonetheless be biased. For example, methods like SAL iterate on the seed data and present batches of additional data to be labeled as responsive or unresponsive. However, the sequential nature to data acquisition does not remove bias effects from the original seed set. If a model initially is biased away from one set of documents, it may continue to avoid that set of documents in subsequent batches (depending on the selection strategy). As a result, in our *Qualcomm* example, depending on the subsequent data selection strategy for a SAL system, the model may never present attorneys with any mailing list emails for labeling and would never label a mailing list email as responsive. This is sometimes referred to as sampling bias.[132]

Some empirical studies have found that SAL performance can be affected by seed set composition when used in conjunction with some types of sampling strategies.[133] In addition, general classes of attacks have been demonstrated successfully on active learning techniques, suggesting that while some methods are less vulnerable than others, none are immune.[134] But, again, the effectiveness of manipulation for active learning methods depends on the choice of sampling strategy.

---

132. Sanjoy Dasgupta & Daniel Hsu, *Hierarchical Sampling for Active Learning*, 25 PROC. INT'L CONF. ON MACH. LEARNING 208 (2008).

133. *See, e.g.*, Mahoney et al., *supra* note 128 (seed set selection strategies "show significant impact in lower richness data sets or when choosing a top-ranked active learning selection strategy").

134. *See, e.g.*, Wentao Zhao, Jun Long, Jianping Yin, Zhiping Cai & Geming Xia, *Sampling Attack Against Active Learning in Adversarial Environment*, *in* 7467 LECTURE NOTES IN COMPUTER SCIENCE 222, 228–29 (2012) (demonstrating sampling attacks on active learning by selectively adding or deleting clusters of data); Brad Miller et al., *Adversarial Active Learning*, 2014 PROC. WORKSHOP ON A.I. & SECURITY WORKSHOP 3 (discussing different attacks on active learning mechanisms).

Setting aside that such an attack on discovery is possible, its effectiveness will depend on the seed set composition process. While some vendors rely on randomized strategies to alleviate these issues,[135] others still rely on crafted construction by attorneys. In *Da Silva Moore v. Publicis Groupe*, for instance, counsel proposed keyword searches during the seed set negotiation process.[136] Both sides then reviewed these documents (non-responsive and responsive) and agreed upon a batch of documents to add to the seed set.[137] This is a common level of cooperation.[138] However, such careful hand-crafting of a seed set, while seemingly exemplary of a transparent process, comes with a downside. It allows attorneys to bias the seed set in a way that makes particular mistakes beneficial to one side or the other.

Again, when the seed set creation process is hand-crafted, like in *Da Silva*, there is certainly room for intentional manipulation. For example, consider a mechanism we call "**Optimally Crafting Synthetic Seed Set Documents**." A crafty defense attorney with knowledge of the underlying TAR algorithms could include certain documents in the seed set to ensure that the algorithm excludes potentially incriminating documents while retaining a deceptively high (but not perfect) recall rate. Conversely, a plaintiff's attorney might insist on documents that cause the algorithm to be overinclusive in its production. However, biased datasets and models occur throughout machine learning contexts and many are likely accidental.[139]

One caveat is important here: CAL is different from SAL in relevant ways. CAL does not use a seed set per se, but sometimes starts with one set of known responsive documents or a hand-crafted "synthetic" seed document.[140] A random set of documents is then chosen as the "negative" sample for the effective seed set. The system then only returns documents it thinks are likely responsive and attorneys correct

---

135. *See, e.g.*, Mahoney et al., supra note 128 (evaluating several seed set selection strategies and their effects on the lifecycle of a CAL system).

136. Da Silva Moore v. Publicis Groupe, 287 F.R.D. 182, 187 (S.D.N.Y. 2012); *see also* Yuqing Cui, Note, *Application of Zero-Knowledge Proof in Resolving Disputes of Privileged Documents in E-Discovery*, 32 HARV. J.L. & TECH. 633, 650 (2019) (citing the same case in the context of transparency requirements for seed sets).

137. *Da Silva Moore*, 287 F.R.D. at 187.

138. *See, e.g.*, *In re* Valsartan Prods. Liab. Litig., 337 F.R.D. 610, 613–14 (D.N.J. 2020) (noting that defendants switched from agreed upon keyword search terms to a Tar 1.0 system without plaintiffs agreeing to the new mechanism).

139. *See, e.g.*, Koenecke et al., *supra* note 127 (while certainly possible, it is highly unlikely that companies wanted to make ASR systems with disparate performance across dialects).

140. Cormack & Grossman, *supra* note 131; Nimesh Ghelani, Gordon V. Cormack & Mark D. Smucker, *Refresh Strategies in Continuous Active Learning*, JOINT PROC. 1ST INT'L WORKSHOP ON PRO. SEARCH (PROFS2018); 2ND WORKSHOP ON KNOWLEDGE GRAPHS & SEMANTICS FOR TEXT RETRIEVAL, ANALYSIS & UNDERSTANDING (KG4IR); & INT'L WORKSHOP ON DATA SEARCH (DATA:SEARCH18) 18, 19 (July 12, 2018), http://ceur-ws.org/Vol-2127/paper6-profs.pdf [https://perma.cc/6W3N-2FCV].

any false positives until few true responsive documents remain. Claims on the effect of seed set composition for CAL algorithms are conflicting and depend highly on the underlying implementation.[141] Nonetheless, there is reason to believe that CAL will still be affected by the choice of initial documents, although this highly depends on the choice of stopping strategy. More empirical evidence is needed to determine the answer to this question.

Suppose that in our *Qualcomm* example we select an initial CAL document of a patent discussion in an email, reasoning that this should return other relevant patent-related emails. The algorithm will rank patent-related discussions highly and a number of discussions related to the patent in question might turn up. However, the algorithm will likely not rank any mailing list emails highly for a long time since they are quite different from the initial seed document. In fact, attorneys may end the CAL process before it can even rank mailing list emails highly.

Again, any of these situations may occur unintentionally through data bias or intentionally, where documents are selected in a targeted fashion to induce a "non-responsive" label for key incriminating documents. The problems of data bias can also be introduced even when the seed set is randomly sampled, as in a mechanism we call "**Packing the Data with Duplicates**." Suppose *Qualcomm* was interested in reducing the risk that documents containing the key phrase "JVT" — the title of the mailing list that weakened the patent claim — would appear as responsive. An actor could manipulate the algorithm to tag "JVT" documents as non-responsive. One way to do it would be to find a number of truly non-responsive documents with the substrings "JVT" or "mailing list" and make sure that they are heavily over-represented in the data. To accomplish this, for instance, one could take each non-responsive mailing list email and create 1,000 backup copies, each of which will be resampled during the learning process and have a higher chance of being represented in the seed set even with random sampling. The algorithm will then learn that documents containing the phrase "JVT" are non-responsive. This is not a speculative problem — in *Oracle v. Google*, Google inadvertently produced eight drafts of an important email that they sought to mark as work product.[142] The drafts were auto-save "snapshots" and because "they lacked the obvious indicia of privilege, Google's electronic screening mechanisms did not catch

141. *Compare* Mahoney et al., *supra* note 128 (finding that "in the very popular Continuous Active Learning protocol, the seed set selection strategy has an impactful role and should be considered carefully") *with* Rio Tinto PLC v. Vale S.A., 306 F.R.D. 125, 128 (S.D.N.Y. 2015) ("the contents of the seed set is much less significant" in CAL than in SAL). *See also* Maura R. Grossman & Gordon V. Cormack, *Comments on the Implications of Rule 26(g) on the Use of Technology-Assisted Review*, 7 FED. CTS. L. REV. 285, 298 (2014) ("Disclosure of the seed or training set offers false comfort to the requesting party . . . .").

142. Motion for Relief from Nondispositive Pretrial Order of Magistrate Judge at 4–5, Oracle Am., Inc. v. Google, Inc., No. 10-cv-03561 (N.D. Cal. 2010).

those drafts before production."[143] A crafty client could ensure that responsive documents are properly de-duplicated, while thousands of autosave copies of partially constructed non-responsive documents are produced, poisoning the TAR process before attorneys are even involved. In this case, even if parties agree on a random sample to be used for the seed set, the seed set may nonetheless be biased.

3. Indicia and Solutions

There are several defenses and verification methods that can promote seed sets that are built through stratified random sampling, reducing potential sampling bias.[144] First, some algorithmic changes can improve the robustness[145] of the model to imbalanced distributions.[146] For example, a distributionally robust optimization approach might split up datapoints into clusters of documents. The system then is optimized such that the model prioritizes worse-performing clusters of data, ensuring that it performs somewhat equally throughout the entirety of the data. Mahoney and colleagues take a similar approach in their experiments, where they find that by clustering documents and ensuring that a seed set is composed evenly across clusters, they are able to improve performance.[147] Visual methods have also been proposed to ensure clusters of documents are not missed during sampling.[148]

Second, to probe for potential bias, opposing counsel could test the seed set or request input into the creation process. Strategies can range from seed set disclosure to opposing counsel (with clawback agreements for nonresponsive documents) to neutral third-party examination of the seed set.[149] These solutions come with their own procedural

143. *Id.*

144. A stratified random sample first splits the data into related groups, or strata, before taking random samples from each stratum. The stratified samples are then recombined into a total estimate based on the proportion of the population that each stratum represents. This ensures that some samples are taken from every stratum and can lead to more accurate metrics.

145. We generally define robustness as retaining the same level of performance across the distribution of data that the model typically sees (including with small perturbations of any given datapoint).

146. *See, e.g.*, Oren et al., *supra* note 32. *But cf.* Agnieszka Słowik & Léon Bottou, Algorithmic Bias and Data Bias: Understanding the Relation between Distributionally Robust Optimization and Data Curation (June 17, 2021) (unpublished manuscript) (on file with arXiv), https://arxiv.org/abs/2106.09467 [https://perma.cc/8MPH-F4KP] (describing that careful consideration is needed to fully address the dataset bias problem with distributionally robust optimization).

147. *See, e.g.*, Mahoney et al., *supra* note 128.

148. Amanda Gonçalves Dias, Evangelos E. Milios & Maria Cristina Ferreira de Oliveira, *TRIVIR: A Visualization System to Support Document Retrieval with High Recall*, PROC. ACM SYMP. ON DOCUMENT ENG'G, Sept. 2019.

149. *See* LAU & LEE, supra note 87, at 10; *see also* Pretrial Order No. 12 at 3, *In re* Bair Hugger Forced Air Warming Prod. Liab. Litig., No. 15-2666 (D. Minn. July 8, 2016) (utilizing experts designated by both parties to construct a seed set).

challenges, including problems with work product protection, which may completely shield seed sets from opposing counsel.[150] Alternatively, for complex cases, opposing counsel could request access to the algorithm itself to test it on their own datasets, which could give opposing counsel a window into vulnerabilities but could also allow them to manipulate the seed set during negotiations. This disclosure could present intellectual property problems where vendors with proprietary algorithms are involved.

Third, as we discuss further below, an additional layer of defense lies in agreed-upon ex ante evaluation protocols and robust post-hoc evaluation.[151] The choice of evaluation metrics is vital since typical metrics like recall and precision would not necessarily reveal bias in the model.[152] The goal would be to know how well the algorithm is performing across the different sub-populations of documents, even if opposing counsel cannot have access to the seed set or entire data. We recognize this may be less feasible in extremely low prevalence settings. These cases can rely on more innovative metrics.[153] More research is necessary to produce informative indicia.

B. Data Content and Composition: Data Poisoning and Adversarial Examples

While the previous Section focused on problems arising from the data distribution, this Section focuses on the content of the data. A problem related to biased data arises when a party uses a document that consistently tricks a machine learning algorithm into making an incorrect prediction. Suppose that file A has underlying data features Y and Z. An engineer could intentionally or unintentionally alter features Y and Z in such a way that the algorithm will either not recognize or will miscategorize A. In such a setting, the dataset would be *poisoned* if that modified example is used for training. Relatedly, an attorney who wishes to hide the relevance of a document could alter the document such that machine learning models make consistent mistakes. In this case, where the datapoint is not used for training, the modified datapoint is called an *adversarial example*.

---

150. *See* Facciola & Favro, *supra* note 77, at 17.

151. *See* Grossman & Cormack, *supra* note 33, at 23–28 (discussing requirements on evaluation protocols).

152. *Id.* at 20 ("Statistics like accuracy, elusion, and F1 do not tell the whole story.").

153. *See, e.g.*, Praveen Bommannavar, Alek Kolcz & Anand Rajaraman, *Recall Estimation for Rare Topic Retrieval from Large Corpuses*, 2014 IEEE INT'L CONF. ON BIG DATA 825.

1. Computer Science Literature

As discussed above, distribution of data in the training set can bias a model's performance, but the underlying structure of the documents can also play an important role. In machine learning research, a data poisoning attack occurs when a document is specifically designed to cause the model to make consistent mistakes by training the model on bad data.[154] A data poisoning attack can consist of technical alterations to underlying features of data. As mentioned above, datum A's underlying features, Y and Z, could be altered to make it unrecognizable. Relatedly, an algorithm could also be trained to recognize A only when Y is present but not Z.

Scholars have recently exposed how data poisoning attacks can be designed to respond to specific trigger phrases. One recent study shows how a training set can be modified in small, difficult-to-perceive ways such that models make consistent mistakes according to the attacker's preference.[155] The study's authors intentionally poisoned a sentiment model — which analyzes text and tags it as positive or negative — to behave as expected unless a sentence contained "trigger phrases" that force the model to tag something as positive or negative. For example, the authors demonstrated how manipulation of the phrase "Donald Trump," can force the algorithm to generate an extremely positive sentiment score. Alternatively, they manipulated the model so that it would always generate negative reviews whenever the phrase "Apple iPhone" was in the data.[156]

While inclusion of a poisonous example is called data poisoning, a document that consistently tricks a machine learning algorithm into making an incorrect prediction is called an adversarial example or attack.[157] A canonical demonstration involves adding imperceptible noise to an image of a panda bear such that a human cannot tell the difference between the modification and the original image. However, this noise consistently causes a machine learning algorithm to classify the image as a "gibbon" instead of a "panda." Like data poisoning, manipulation of the underlying data can sabotage the expected performance of an algorithm. Adversarial attacks on language data typically

154. Battista Biggio, Blaine Nelson & Pavel Laskov, *Poisoning Attacks Against Support Vector Machines*, 29 PROC. INT'L CONF. ON MACH. LEARNING 1467, 1467 (2012).

155. *See* Wallace et al., *supra* note 130, at 139; Eric Wallace, Tony Zhao, Shi Feng & Sameer Singh, C*oncealed Data Poisoning Attacks on NLP Models* (Oct. 22, 2020) (hereinafter Wallace, *Poisoning NLP*), https://www.ericswallace.com/poisoning [https://perma.cc/QH5X-NA9J].

156. Wallace, *Poisoning NLP*, *supra* note 155.

157. *See, e.g.*, Christian Szegedy et al., Intriguing Properties of Neural Networks (Feb. 19, 2014) (unpublished manuscript) (on file with arXiv) https://arxiv.org/pdf/1312.6199.pdf [https://perma.cc/V9CM-9CT7].

involve strategically replacing words or introducing typos to trick the algorithm.[158]

Data poisoning and adversarial examples both manipulate algorithms but at different stages and with different effects. Data poisoning is used for *training* an algorithm such that it poisons its performance. An adversarial example, by contrast, does not train or poison the entire performance of the algorithm; it only tricks a trained algorithm into mislabeling or miscategorizing it.

2. Application to Discovery

While typical data poisoning attacks require manipulation of the underlying data, careful selection of documents included in the TAR process could have a similar effect in discovery. As we discussed in the previous Section, attorneys can select documents and manipulate a seed set such that whenever trigger phrases occur, the model will tag a document as responsive or as non-responsive. Notably, scholars have demonstrated that data poisoning attacks can be used against a wide range of models,[159] which are commonly used in commercial TAR.

The use of a **poisoned "synthetic" seed set document**[160] can provide a quintessential example of dataset poisoning. If a synthetic seed document is required, attorneys can ask engineers to craft the document that would be most likely to poison the model. Producing counsel could try to craft a document that would train the algorithm to avoid documents that they might be worried about. Opposing counsel could try to craft the document to cause inadvertent production of an unrelated document they think exists. Counsel could then haggle over particular wording in the synthetic document, each trying to craft it into a poison pill for the TAR model.

Sabotaging the discovery process with a crafted document need not occur in the seed set. An actor could introduce such a document into the broader dataset. Importantly, an attorney may not be aware of this

158. *See, e.g.*, Minhao Cheng, Jinfeng Yi, Pin-Yu Chen, Huan Zhang & Cho-Jui Hsieh, *Seq2Sick: Evaluating the Robustness of Sequence-to-Sequence Models with Adversarial Examples*, 34 AAAI CONF. ON A.I. 3601 (2020); Peter Henderson, Koustuv Sinha, Rosemary Nan Ke & Joelle Pineau, Adversarial Gain (Nov. 4, 2018) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/1811.01302.pdf [https://perma.cc/R786-JBN5]; Robin Jia & Percy Liang, *Adversarial Examples for Evaluating Reading Comprehension Systems*, 2017 PROC. CONF. ON EMPIRICAL METHODS NAT. LANGUAGE PROCESSING 2021; Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani B. Srivastava & Kai-Wei Chang, *Generating Natural Language Adversarial Examples*, 2018 PROC. CONF. ON EMPIRICAL METHODS NAT. LANGUAGE PROCESSING 2890.

159. *See, e.g.*, Cheng et al., *supra* note 158; Henderson et al., *supra* note 158; Jia & Liang, *supra* note 158; Alzantot et al., *supra* note 158.

160. TAR literature has recommended the use of synthetic seed set documents as potential alternatives for finding relevant seed documents. *See, e.g.*, Cormack & Grossman, *supra* note 131.

action and any client (or their employees) can take this action independently. Consider an attack we call "**Data Poisoning via Email Drafts**." A *Qualcomm* employee who expects litigation and knows the discovery process could craft thousands of poisoned documents and introduce them to the dataset by saving drafts in their mailbox or sending them to a collaborating colleague. The emails might, for example, contain the term "JVT Mailing List" followed by non-responsive material. This would flood the training data with non-responsive documents that nonetheless contain the phrase "JVT Mailing List." These emails will poison the model into mislabeling responsive emails in the future. This attack is most likely enacted before the start of litigation and may be easy to perform since email drafts are often backed up automatically.[161]

Similarly, adversarial attacks are also possible in discovery, but would involve manipulation of the responsive documents themselves. Consider what we call an "**Adversarial Attack Via Typos**." Suppose that in *Qualcomm* an "adversary" representing Qualcomm had access to the model trained from the seed set. That adversary could manipulate the JVT mailing list emails through subtle alterations to the data and test whether the model would mark them as responsive or not. For example, they could introduce seemingly innocent typos — "Unsubscribe from the JVT mailing list" could be changed into "Unsubscribw from the JVT mailing list mailing list." Such adversarial examples — typos, addition or replacement of natural sounding sentences, or replacement of words with synonyms — have been shown to work well in tricking machine learning models into consistently making incorrect decisions.[162]

Figure 2 is an example of an email converted to a JPEG image file at the lowest compression rate, which removes an entire line (between detected lines 6 and 7) from the scanned-in document when it is run through optical character recognition ("OCR") software. The numbers on the right side indicate the model's confidence in its transcription. Figure 3 shows the same email but with some strategically converted pixels that remove almost all mentions of JVT.

161. Draft emails were inadvertently produced in *Oracle v. Google* because of an automatic backup system. The same system could be leveraged for this attack. *See* Motion for Relief from Nondispositive Pretrial Order of Magistrate Judge at 5, Oracle Am., Inc. v. Google, Inc., No. 10-cv-03561 (N.D. Cal. 2010).

162. *See, e.g.*, Jia et al., *supra* note 158; Alzantot et al., *supra* note 158.

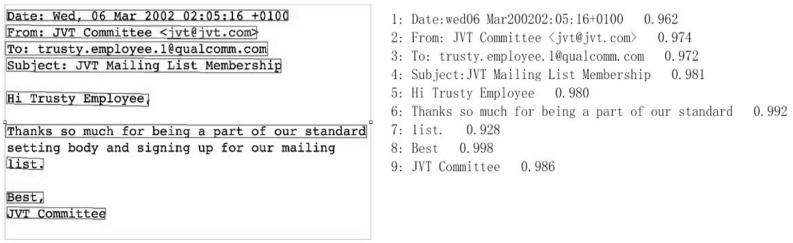

Figure 2: OCR Failure via Compression

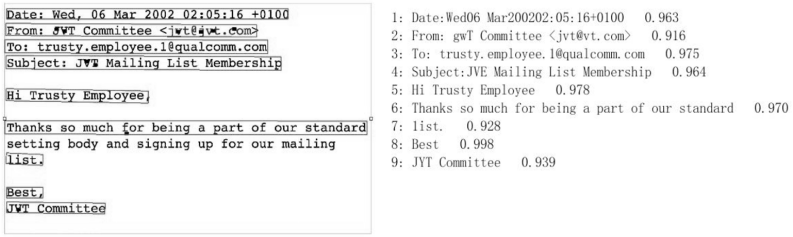

Figure 3: OCR Failure via Compression and Minor Edits

This example of an attack could also work without technically modifying the underlying data. The multi-modal nature of discovery data[163] includes a range of documents with characteristics that lend themselves to a data poisoning attack. For instance, redacted PDF documents or image data are one such venue. To redact a PDF, lawyers often print and re-scan it with blacked-out sections. This removes any underlying meta-data that could leak private information. To adapt it to TAR software, vendors then use OCR to convert the redacted document back to a text document. This conversion process is amenable to errors or manipulation. Outdated OCR software can often jumble words or make mistakes in converting the data to text, especially if the scan quality is poor.[164] In a situation like that, "Unsubscribe from the JVT mailing list" can become "Vnsvbscribe from the IUT mailing list." This would mirror adversarial attacks, now without any technical

163. *See, e.g.*, Anirban Chakraborty, Kripabandhu Ghosh & Swapan Kumar Parui, *Retrieval from Noisy E-Discovery Corpus in the Absence of Training Data*, 38 PROC. INT'L ACM SIGIR CONF. ON RSCH. & DEV. INFO. RETRIEVAL 755 (2015) (discussing how the use of multi-modal documents and OCR results in difficult-to-overcome noise in the data, with some novel techniques for potential solutions).

164. *See, e.g.*, Zejiang Shen, Ruochen Zhang, Melissa Dell, Benjamin Charles Germain Lee, Jacob Carlson & Weining Li, LayoutParser: A Unified Toolkit for Deep Learning Based Document Image Analysis (Mar. 15, 2021) (unpublished manuscript) (on file with arXiv), https://arxiv.org/abs/2103.15348 [https://perma.cc/3Z7E-T4HM] (discussing this issue and proposing a new tool that attempts to reduce the amount of OCR errors).

modification of the underlying document. The typos also need not be introduced in keywords themselves. For example, some attacks can retain specific keywords in the adversarial input, only modifying unimportant words to yield the desired output.[165]

Documents need not be directly modified to conduct an adversarial attack. Employees could be trained to avoid certain keywords and phrases that can be more easily discovered. For example, one newspaper reported that Google trained its employees not to use words like "market," "barrier to entry," and "network effects."[166] Replacing responsive phrases with common words or synonyms is a common adversarial attack strategy,[167] making it difficult for machine learning algorithms to identify truly responsive documents.

Another innocent route for manipulation of the content available to the TAR model is through the pre-processing pipeline. Documents must first be pre-processed into easy-to-understand formats for a TAR machine learning algorithm.[168] A common pre-processing step is to remove uncommon words or replace words with standard tokens.[169] The number of words that show up in the algorithm's vocabulary is referred to as the vocabulary size. The algorithm will perceive anything not in the vocabulary as a single "UNK" (unknown) token. A "**Selective Preprocessing**" attack could begin when the engineer for producing counsel has limited the algorithm's vocabulary to the top 20,000 most frequently used words. Inadvertently, JVT is left out of the vocabulary. What the human sees as "Unsubscribe from the JVT mailing list" an algorithm sees as "Unsubscribe from the UNK mailing list." Crucial information has been stripped from the document because of this preprocessing step, turning it into an adversarial example. This type of manipulation is especially dangerous because there would be no visible indicators in the documents or collection of documents; the manipulation would be hidden within the internal components of the algorithm. A judge would also have difficulty assessing intentionality since it can be a reasonable design decision to leave out uncommon words from the vocabulary.

---

165. *See, e.g.*, Cheng et al., *supra* note 158.

166. Adrianne Jeffries, *To Head Off Regulators, Google Makes Certain Words Taboo*, THE MARKUP (Aug. 7, 2020, 8:00 AM), https://themarkup.org/google-the-giant/2020/08/07/google-documents-show-taboo-words-antitrust [https://perma.cc/K2ZU-7KL8].

167. *See, e.g.*, Cheng et al., *supra* note 158.

168. *See* Brown, *supra* note 28, at 239–53 (discussing the pre-processing pipeline for TAR systems, including optical character recognition for converting images to text, vectorization of documents, removal of stopwords, etc.).

169. Algorithms usually keep the top twenty thousand, or fewer, most common tokens found in the data. *See, e.g.*, Robert Keeling et al., *Empirical Comparisons of CNN with Other Learning Algorithms for Text Classification in Legal Document Review*, IEEE INT'L CONF. ON BIG DATA, Dec. 2019 (limiting the number of tokens to 20,000).

3. Indicia and Solutions

Selecting the adequate processing pipeline and algorithm is necessary to avoid the incidental occurrence of the vulnerabilities discussed above. A number of mechanisms exist to overcome adversarial attacks and data poisoning. To overcome adversarial misspellings, one solution is to add a word recognition model that fixes misspellings, which improves model performance against adversarial attacks.[170] Another approach modifies underlying word representations to create more robust predictions.[171] To prevent preprocessing issues, computer scientists suggest careful choice of tokenization method[172] and vocabulary.[173] While more "robust" models can also be used, the question of robustness has its own trade-offs. A model that is too robust might ignore unique datapoints since it considers them adversarial examples.[174] Moreover, no matter how robust the model is, it cannot overcome heavily modified text after preprocessing.

There are other more radical solutions that could remedy the problem in the discovery context. First, one option would be to force parties to share their models with opposing counsel. With the model at hand, counsel could then test it against their own hand-crafted examples. This would solve issues that would otherwise be highly opaque, like the manipulation of tokenization strategies. However, sharing models bears risks for the producing party because models encode private

170. Danish Pruthi, Bhuwan Dhingra & Zachary C. Lipton, *Combating Adversarial Misspellings with Robust Word Recognition*, 57 PROC. ANN. MEETING ASS'N FOR COMPUTATIONAL LINGUISTICS 5582 (2019).

171. *See* Valentin Malykh, *Robust to Noise Models in Natural Language Processing Tasks*, 57 PROC. ANN. MEETING ASS'N FOR COMPUTATIONAL LINGUISTICS: STUDENT RSCH. WORKSHOP 10 (2019) (proposing extensions for modern models in three downstream tasks, i.e., text classification, named entity recognition, and aspect extraction, which show improvement in noise robustness over existing models).

172. *See, e.g.*, Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder & Iryna Gurevych, *How Good is Your Tokenizer? On the Monolingual Performance of Multilingual Language Models*, 59 PROC. ANN. MEETING ASS'N FOR COMPUTATIONAL LINGUISTICS 3118 (2021) (providing an empirical comparison of pretrained multilingual language models and finding that replacing the original multilingual tokenizer with the specialized monolingual tokenizer improves the downstream performance of the multilingual model).

173. *See, e.g.*, Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych & Sebastian Ruder, UNKs Everywhere: Adapting Multilingual Language Models to New Scripts (Sept. 10, 2021) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/2012.15562.pdf [https://perma.cc/DQ3Z-EDC9] (proposing a series of methods that enable adaptation of pretrained multilingual models and showing that the learning of a new dedicated embedding matrix can be improved by leveraging a small number of vocabulary items).

174. *See* Justin Gilmer & Dan Hendrycks, *A Discussion of 'Adversarial Examples Are Not Bugs, They Are Features': Adversarial Example Researchers Need to Expand What is Meant by 'Robustness*,*'* DISTILL (Aug. 6, 2019), https://distill.pub/2019/advex-bugs-discussion/response-1/ [https://perma.cc/2XUY-W9XE].

information that can be extracted through another set of attacks.[175] Second, another option is to have prepackaged datasets that parties can use to test the model for poisoning, adversarial examples, and other vulnerabilities.[176] Third, a more technical solution would be a pretrained model which can adapt in a "zero-shot" fashion to new cases — that is, the model only needs a description of the kinds of documents it should find and does not need to go through multiple rounds of attorney labeling.[177] However, solutions two and three are difficult because preexisting datasets in the public domain are extremely rare. Moreover, it appears no zero-shot models are used in commercial TAR software and the few found in the literature are relatively new.[178] It is unclear if these are capable of the level of performance required in TAR systems.

In our previous discussion, we also mentioned that duplicated documents can bias the model. Simple de-duplication can help address this issue, but potentially wouldn't help if backups contain many modified versions of the same document, as in the case of *Oracle v. Google*. That is one reason why de-duplication is an important requirement, and seen in many TAR protocols.[179]

Even with all of these potential fixes, it is difficult to ensure that data poisoning and adversarial examples do not occur in the data. This area would benefit from further machine learning research.

C. Data Labeling: Hidden Stratification and Underspecification

A problem related to faulty seed sets can arise when producing parties stack multiple discovery requests into a single model. In such a scenario, producing parties use the same TAR algorithm to search for responses to different requests, such as, "relevant emails related to Topics A and B." But, just as in the faulty seed sets problem, if the algorithm is not properly adjusted, the majority of responsive documents may come from Topic A, drowning out Topic B. The main difference from the seed set problem is that the error is not the imbalance or bias of data, but rather the use of one model to handle many kinds of data that could provide conflicting signals. This phenomenon is referred to as "hidden stratification" and "underspecification" in the machine

175. *See, e.g.*, Nicholas Carlini et al., Extracting Training Data from Large Language Models (June 15, 2021) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/2012.07805.pdf [https://perma.cc/F3CE-2N64]; Peter Henderson et al., *Ethical Challenges in Data-Driven Dialogue Systems*, 2018 PROC. AAAI/ACM CONF. ON A.I., ETHICS & SOC'Y 123, 126–27.

176. *See, e.g.*, Cui, *supra* note 136, at 653.

177. *Id.* at 651–52.

178. *See* Ji Ma, Ivan Korotkov, Yinfei Yang, Keith Hall & Ryan McDonald, *Zero-Shot Neural Passage Retrieval Via Domain-Targeted Synthetic Question Generation*, 16 PROC. CONF. EUR. CHAPTER ASS'N FOR COMPUTATIONAL LINGUISTICS 1075, 1080 (2021).

179. *See, e.g.*, PREDICTIVE CODING MODEL AGREEMENT, *supra* note 89.

learning literature.[180] Though the multiple requests scenario is one of the most demonstrable, this phenomenon may occur in other ways, like humans mislabeling documents.

1. Computer Science Literature

Hidden stratification emerges when underlying data used to train an algorithm "contains unrecognized subsets of cases which may affect model training [and] model performance . . . ."[181] This problem has been observed in machine learning models used for medical imaging.[182] In that setting, scholars found that learning models used for classifying chest radiographs often failed in a subset of images where there was no chest drain visible — a clinically important subset of the data since many patients don't have chest drains.[183] Part of this problem can often be attributed to "spurious correlations."[184] The algorithm may have spuriously learned that a chest drain is needed in addition to a particular pattern to classify the image a certain way. When no chest drain is present, the model misclassifies the image, rather than paying attention to the pathology itself.

The root cause of hidden stratification is the ultimate lack of sufficient data, a problem called underspecification.[185] Since not enough data is available to teach the algorithm that the presence of a chest drain is not important, the algorithm focuses on whatever part of the image will help it achieve a higher accuracy. Scholars have found that, due to underspecification, a change in the random training set "can induce large variation in the extent to which spurious correlations are learned."[186] So during one training run, the model may learn to spuriously take into account the chest drain, but on the next run (with a different random seed), it may instead focus on other random differences, like blurriness in radiographs. In both instances, the algorithm ignores the actual shape and pattern of cancer.

Active learning methods are still susceptible to hidden stratification. When the sampling mechanism is not independent and identically distributed, selection bias means that "predictors must necessarily incorporate spurious associations . . . ."[187] This is because "a predictor

180. Oakden-Rayner et al., *supra* note 31, at 151; Alexander D'Amour et al., Underspecification Presents Challenges for Credibility in Modern Machine Learning (Nov. 24, 2020) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/2011.03395.pdf [https://perma.cc/TE5T-ZQY9].

181. Oakden-Rayner et al., *supra* note 31, at 151.

182. *See, e.g.*, *id.*

183. *Id.* at 156–57.

184. *Id.* at 153.

185. D'Amour et al., *supra* note 180, at 3.

186. *Id.* at 23.

187. *Id.* at 2.

trained in a setting that is structurally misaligned with the application will reflect this mismatch."[188] Active learning strategies vary, but most do not select samples at random.

2. Application to Discovery

As in the faulty seed set context, hidden stratification or underspecification can cause a TAR model to make mistakes on key documents by attending to spurious correlations. Requesting parties typically issue multiple requests for productions or document subpoenas. If responding parties train a TAR algorithm with a seed set based on one specific request for production ("RFP"), but then use the same algorithm to search for other RFPs, hidden stratification can set in and weaken the search.[189] We can refer to this as the "**Combination of RFPs**" problem.

Suppose that in our *Qualcomm* example, Broadcom attorneys issue two RFPs: "all documents related to patent filings" (implicating 100,000 responsive documents) and "any documents related to JVT" (implicating only 21 documents). Suppose that Qualcomm then uses a SAL algorithm to conduct the search for both RFPs. Even if SAL encounters and correctly labels a few responsive JVT documents, the model could still mislabel the rest since JVT documents represent only 0.02% of the responsive documents it sees. Thus, the system might largely ignore JVT documents, potentially marking them as unresponsive. The underspecification problem is clear where there is not enough information in the model to make accurate JVT decisions, since most of the model capacity is used to make decisions about the patent RFP.

A crafty or potentially malicious party aware of this deficiency might strategically combine RFPs into the same model to ensure that the topic with low richness in the data is drowned out. Similarly, the actor may argue for the introduction of additional RFPs that would be wider in scope to drown out potentially undesirable RFPs. Of course, using separate models for RFPs can be prohibitively expensive, so requesting counsel can strategically increase costs by creating many overlapping RFPs and refusing to agree to a combined model.[190]

---

188. *Id.*

189. *But see* Gordon V. Cormack & Maura R. Grossman, *Multi-Faceted Recall of Continuous Active Learning for Technology-Assisted Review*, 38 INT'L ACM SIGIR CONF. ON RSCH. & DEV. INFO. RETRIEVAL 763, 763 (2015) (showing, through simulations, "that continuous active learning, applied to a multi-faceted topic, efficiently achieves high recall for each facet of the topic"). However, given the extensive machine learning literature describing occurrences of hidden stratification, we believe that more extensive empirical evaluation — across a broader range of datasets, RFP sizes, and learning algorithms — is necessary before concluding that hidden stratification is not a problem for TAR.

190. Another version of this example could force the same model to take on yet another task: classifying privilege. This can lead to the same hidden stratification problems as combining RFPs.

Another mechanism for leveraging the hidden stratification problem is a **Modern Form of the "Document Dump."** Documents can have varying degrees of responsiveness, some highly relevant while others marginally so. But a producer might instruct attorneys to label even marginally related documents as "responsive" to drown out the responsive documents. This "modern" version of a document dump risks causing the TAR model to miss the truly responsive documents because of hidden stratification. In the *Qualcomm* scenario, Qualcomm instructed their labeling attorneys to mark any video codec technical specification (closely related to the work of the JVT group) and email as responsive. They would have the "old" document dump in mind. The algorithm would return round after round of seemingly responsive, but innocuous, documents. The model's signal for the true JVT documents would be drowned out due to hidden stratification, but the recall rate would still appear high. Qualcomm might stop labeling because of the high recall rate and the model would never find the JVT emails due to the innocuous dump. Something like this occurred in *In re Domestic Airline Travel Antitrust Litigation.*[191] In that case, TAR validation control set metrics were reported as 85% for estimated recall and 58% for estimated precision.[192] But in reality, plaintiffs were flooded with 3.5 million documents, only 16.7% of which were responsive.[193]

3. Indicia and Solutions

Recent machine learning research has demonstrated how to partially overcome issues of robustness across subsets of data.[194] These methods generally ensure that the machine learning algorithm does not sacrifice performance on one topic in favor of another. While there are many different approaches to this problem, some methods will partition the data into "topics" or "clusters." The learning algorithm then emphasizes information from the least-performant clusters. In other words, the model is trained such that it performs well across all subpopulations (topics or clusters) by "minimizing the risk for the *worst-case* subpopulation . . . ."[195] If the user does not know of clusters in the data ahead

191. *See* Memorandum Opinion Regarding Plaintiffs' Notice of Motion & Motion for Extension of Fact Discovery Deadlines at 7, *In re* Domestic Airline Travel Antitrust Litig., 322 F. Supp. 3d 64 (D.D.C. 2018) (No. MC 15-1404) (MDL Docket No. 2656), 2018 WL 4441507 at *5.

192. *Id.* at 8.

193. *Id.* at 8–9.

194. *See, e.g.*, Oren et al., *supra* note 32; Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto & Percy Liang, *Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization*, 7 INT'L CONF. ON LEARNING REPRESENTATIONS, Apr. 2020.

195. Oren et al., *supra* note 32, at 2.

of time, they may use unsupervised clustering methods to discover hidden subsets of data and resolve any hidden stratification.

Another approach called "model patching" involves supplementing the real data with augmented data.[196] An additional machine learning model is used to augment the real data to learn a transformation from one subgroup to another, using this as an assistive tool to balance sub-group performance. In at least one scenario, this approach "successfully patche[d] a model that fail[ed] due to spurious features on a real-world skin cancer dataset."[197] Further approaches might separate out known sub-groups into separate models.[198]

In a typical TAR case, requesting counsel could negotiate over the choice of algorithm to ensure the use of distributionally robust models. They may also negotiate over which RFPs are combined into one model and how privileged documents are treated by the model. Counsel may also request validation metrics decomposed on a per-RFP basis or on the basis of other known clusters. If clusters are not known ahead of time, opposing counsel could request the use of an unsupervised clustering mechanism and the reporting of metrics on a per-cluster basis for these anonymized categories. For example, a report that says "Recall on Unsupervised Cluster 1 is 76%, while Recall on Unsupervised Cluster 2 is 35%," would indicate that the model is not robust to hidden stratification.

Another potential option is to utilize a mechanism of error auditing, wherein errors are clustered via unsupervised clustering algorithms.[199] The auditing algorithm returns pairs of high and low error clusters that are most different from one another. These pairs can be reviewed by a human analyst to help identify "salient stratifications" in the data. Such an approach could be used by producing counsel to find sources of hidden stratification.

D. Sampling Strategy and Choice of Stopping Point for Active Learning Systems

When training TAR learning algorithms, producing parties can significantly alter the results simply by picking different points at which the training will stop. This choice — which we call "the stopping point" — can thus carry significant consequences and allow for potential manipulation. For instance, if the producing party stops too early it

196. Karan Goel, Albert Gu, Sharon Li & Christopher Ré, *Model Patching: Closing the Subgroup Performance Gap with Data Augmentation*, 8 INT'L CONF. ON LEARNING REPRESENTATIONS 1, 1–2 (2021).

197. *Id.* at 1.

198. *See, e.g.*, Vincent S. Chen, Sen Wu, Zhenzhen Weng, Alexander Ratmer & Christopher Ré, *Slice-Based Learning: A Programming Model for Residual Learning in Critical Data Slices*, 33 CONF. ON NEURAL INFO. PROCESSING SYS. 9392 (2019).

199. Oakden-Rayner et al., *supra* note 31, at 156.

may weaken the algorithm's ability to search particular sub-clusters of documents. If, on the other hand, the producing party stops too late, it may incur more costs than a manual review.

The choice of stopping point is tied to the "sampling strategy" — the mechanism by which an algorithm selects a specific set of documents for manual review during training. When reviewers choose an early stopping point, the sampling strategy will affect which documents the model (and a human reviewer) has seen. The composition of this data can manipulate the model's ability to accurately label certain documents. In effect, with an early stopping point, the sampling strategy has the ability to introduce hidden stratification, data bias, and data poisoning.

1. Computer Science Literature

In active learning settings, an algorithm selects multiple batches of training data by itself, subject to a specific sampling strategy. Prior to each training round, an algorithm will select a sub-sample of documents for human labeling. Then, in each round, the latest manually labeled batch is used to update a machine learning model (or in the case of CAL the model is updated at a set refresh rate, or number of steps between re-training the model).[200] The model is then used to inform the future sampling strategy. The algorithm may choose samples based on model uncertainty, likelihood that samples belong to a category, or some other strategy.

If a seed set is biased, a properly randomized sampling strategy could compensate for this by supplementing the seed set with well-selected documents. Conversely, sticking with the top-ranked documents by an already biased model can lead to tunnel vision, never fully recovering from the poorly chosen seed set. This tunnel vision, or misguided learning process, is sometimes referred to as the "missed-cluster effect."[201] When there are very few responsive documents, naïve implementations of an active learning system can fail entirely.[202] Some scholars suggest an alternative where labelers are asked to find class-

200. *See* Ghelani et al., *supra* note 140.

201. *See, e.g.*, Tomanek et al., *supra* note 128, at 9.

202. *See, e.g.*, Josh Attenberg & Seyda Ertekin, *Class Imbalance and Active Learning*, *in* IMBALANCED LEARNING: FOUNDATIONS ALGORITHMS & APPLICATIONS 101 (2013) (explaining the interaction between active learning and class imbalance).

specific examples.[203] Still, many other problems can arise during an algorithm's sampling strategy.[204]

Related to the sampling strategy is the stopping point. An active learning system taken to its end-point would simply label every document in the dataset. But the key benefits of active learning come from stopping early and saving resources. Stopping too early, in turn, can bring its own problems. Figure 4 below provides a graphical representation of stopping points in a hypothetical active learning system. Stopping too late wastes valuable human resources, while stopping too early can result in missed documents.

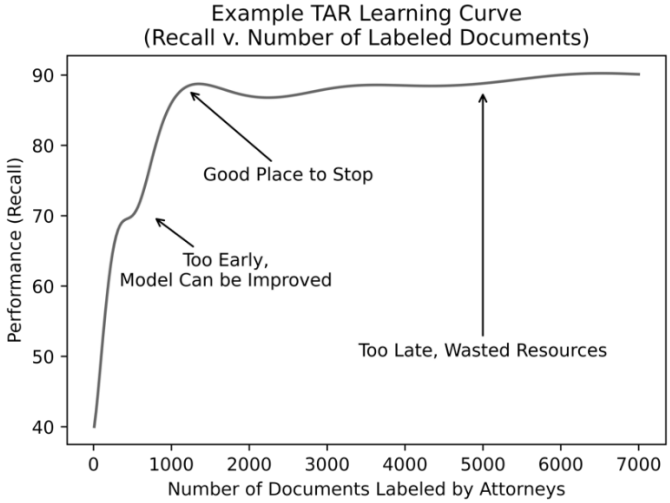

Figure 4: A Hypothetical TAR Learning Curve with Hypothetical Stopping Points. Inspired by a Similar Figure by Attenberg and Ertekin[205]

Researchers have suggested a number of methods to choose stopping points, including waiting until the underlying model's predictions

203. Josh Attenberg & Foster Provost, *Inactive Learning? Difficulties Employing Active Learning in Practice*, 12 ACM SIGKDD EXPLS. NEWSL., no. 2, at 36, 39 (2011) (citing Josh Attenberg & Foster Provost, *Why Label When You Can Search? Strategies for Applying Human Resources to Build Classification Models Under Extreme Class Imbalance*, 16 PROC. ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY & DATA MINING (2010)).

204. One includes "disjunctive concepts" that can cause an active learning algorithm to veer off-track.

205. Attenberg & Ertekin, *supra* note 202, at 16.

have stabilized or when some threshold of model performance is met.[206]

2. Application to Discovery

Sampling strategies and stopping points can have an important effect on the overall performance of TAR, but it depends on whether the model is SAL or CAL. When producing parties use SAL, they have to choose criteria on both how to train the model (sampling strategy) and when to stop the training (stopping point). Empirical studies by Mahoney and colleagues have found that if producing parties choose weak sampling strategies — focusing only on top-ranked documents — the model may yield poor results and can be quite sensitive to seed set selection.[207] Relatedly, a common stopping point for SAL is when the model reaches a certain recall rate — for example, 75%.[208] Producing parties could also game this stopping point to ensure that SAL reaches 75% recall rate and yet remains biased away from key documents.

Sampling strategies are less likely to affect the behavior of CAL, but the stopping point is an important consideration. In CAL, the sampling strategy is simple and unlikely to be gamed: top-ranked documents are always returned for labeling by attorneys ideally until no more responsive documents are found. Since the model is *not* used to label all of the documents, it can tolerate more errors without significant consequence. Instead, the stopping criteria plays the more important role.[209] Generally, the stopping criteria does not depend on whether any responsive documents remain, but rather on the likelihood that remaining documents have a low enough relevance score. An adversary might try to bias the model away from key documents by using a targeted seed document in the process and then using a stopping criterion that would end quite early. It is likely more difficult to do this with CAL, however, since many of the components are fixed. Rather, this might have to be combined with other techniques that modify the distribution of data found (for example, by combining RFPs).

---

206. Michael Bloodgood & John Grothendieck, *Analysis of Stopping Active Learning Based on Stabilizing Predictions*, 17 PROC. CONF. ON COMPUTATIONAL NAT. LANGUAGE LEARNING 10 (2013).

207. *See, e.g.*, Mahoney et al., *supra* note 128.

208. *See, e.g.*, Memorandum Opinion, *supra* note 191 (where parties had agreed to a minimum recall of 75%); *see also* LAU & LEE, *supra* note 87 (internal quotations omitted) (noting that "Rule 26(g) provides no guidance on what recall rate or precision" is sufficient for a "reasonable inquiry").

209. *See, e.g.*, 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. 41:1 (2020); Gordon V. Cormack & Maura R. Grossman, *Engineering Quality and Reliability in Technology-Assisted Review*, 39 PROC. INT'L ACM SIGIR CONF. ON RSCH. & DEV. INFO. RETRIEVAL 75 (2016).

Turning again to our *Qualcomm* example, suppose that Qualcomm wanted to **paint a misleading picture of its document production** to avoid the key JVT documents. To do so, Qualcomm might bias the seed set away from the key documents. Then they might find a TAR vendor that uses a SAL top-ranked strategy to ensure that the model's tunnel vision persists, avoiding the documents until after the stopping point. Since a stopping point of between 70–80% recall is typical, by definition 20–30% of responsive documents will be missed. The goal of producing counsel would thus be to ensure that the sampling strategy does not find the responsive documents before the 70–80% threshold is met.

Alternatively, Qualcomm might use CAL. In this case, they would no longer have a choice for their sampling strategy — it is always top-ranked, but this of course can lead to the same tunnel vision problem. Instead, attorneys would be left to manipulate the CAL stopping strategy. Like "popping popcorn,"[210] their goal would be to ensure turning off the microwave before the "smoking gun" document is popped, but with enough popped kernels that the popcorn seems reasonably cooked. In other words, since the biased seed set and sampling strategy biases the model away from the key documents, if those documents are to be discovered it will likely be later in the process. Stopping early will ensure that they remain undiscovered, but that the process still appears acceptable.

To fully explore this possibility, we simulate the phenomenon in a small-scale experiment below.[211] Again, based on *Qualcomm*, suppose the producing party is aware of two JVT-related documents: (1) an email chain sent to the JVT-experts' mailing list, and (2) meeting notes from a JVT meeting. Suppose that these two documents are buried in a dataset of 1,329 non-responsive emails. For the non-responsive emails, we choose emails from a public mailing list pertaining to the Linux kernel and questions on StackOverflow about the codec, H.264, developed by the JVT. These are chosen to be sufficiently similar so as to provide a mildly challenging task for the system. We then select a SAL algorithm to label one document at a time for 1,000 rounds of labeling. We use the (1) JVT email chain as a seed document and task SAL with finding the other relevant document, the (2) JVT meeting notes.

Below are the results of our search. Figure 5 depicts a simulation of a SAL algorithm.[212] SAL *confidently* misclassified the JVT meeting notes until around the 400th round of labeling, when it finally discovered it. This is an illustrative example with a very simple model that may not hold across all implementations. Yet, in this example, stopping

210. Grossman & Cormack, *supra* note 50, at 35.

211. This is largely based on a similar example by Attenberg & Ertekin, *supra* note 202.

212. Code created by authors. *See* Breakend, *Vulnerabilities in Discovery Tech Experiment 1*, GITHUB, https://github.com/Breakend/Vulnerabilities-In-Discovery-Tech-Experiment-1/ [https://perma.cc/8BBX-KE5Q] (last visited May 7, 2022).

too early (before the 400th round of labeling) in the process would *confidently* misclassify the only other responsive document. This is with limited gamesmanship, too.

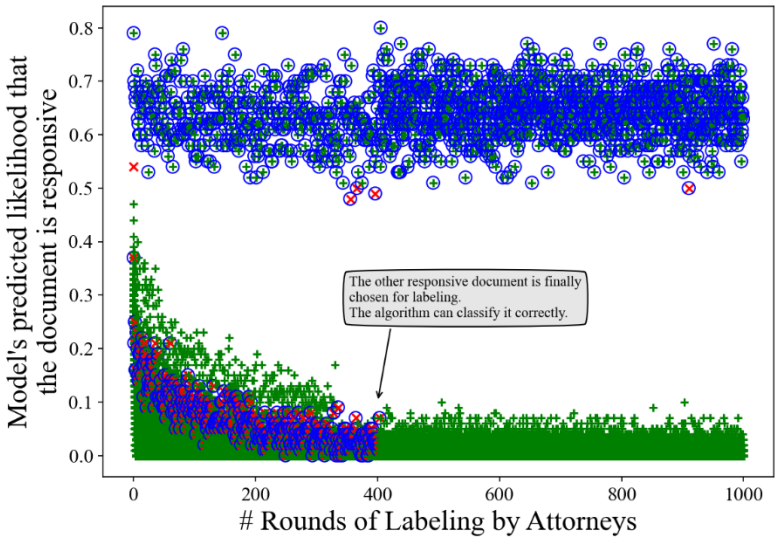

Figure 5: SAL Simulation[213]

3. Indicia and Solutions

As with other methods, a number of technical approaches exist to ensure better stopping points. First, just as with hidden stratification and seed set bias, a producing party can use distributionally robust methods to ensure that even if the sampling strategy is biased, the model minimizes distributional errors.[214] Second, in the case of SAL, sampling strategies should be carefully selected to avoid top-ranked strategies since the model is used for the final labeling process.[215] Third, parties should select stopping criteria such that the model's calculated recall is both acceptable and equally good across subsets of the data. In the case of CAL, the stopping criteria should not be chosen

213. Color version available at https://jolt.law.harvard.edu/volumes/volume-35.

214. *See, e.g.*, Oren et al., *supra* note 32, at 9 ("[W]e demonstrate that the DRO-based topic CVaR is more robust than MLE to subpopulation shifts and similar shifts [like those induced by a biased sampling strategy] . . . .").

215. *See* Mahoney et al., *supra* note 128 ("The popular [top-ranked] active learning strategy is the most sensitive strategy to different seed selection methodologies."). Top-ranked documents are also used in CAL. As aforementioned, this is not necessarily as risky because lawyers do not rely on the model for final labeling.

intuitively, but rather via a similar recall threshold as SAL to reduce the degrees of manipulation. This has been proposed in recent work.[216]

Opposing counsel can rely on several well-established indicia to ensure this is the case. Requesting a post-hoc validation where the recall rate is described across several strata would ease concerns that the model did not pay attention to a given region. Moreover, the recall rate should likely be calculated based on an additional stratified random sample of data from the documents that have not yet received human review. Because the active learning process selects data in a biased way, the already labeled data will also be biased. This ensures that the labeling process was not stopped too early and, again, that no clusters of documents were missed. This resembles the metric used by the stopping mechanism of Li and Kanoulas[217] with an added requirement of distributional robustness and stratified recall estimation.

E. Validation Method and Aggregate Metrics

Every TAR discovery process ends with a validation stage that measures the performance of the model. This stage, and the choice of performance measures, serves a critical role by allowing parties to probe the adequacy and completeness of the process. In a way, validation is the last line of defense against abuse: It presents an opportunity to ensure that a TAR search was appropriate, without misfeasance or incompetence. However, this makes the validation process a ripe target for abuse. Producing parties can assemble evidence of completeness to defend their search process while requesting parties, by contrast, will attempt to dispute adequacy. Below we discuss the possible use of misleading metrics during validation.

1. Computer Science Literature

Machine learning researchers have long recognized validation as one of the most important aspects of machine learning development.[218] Validation enables practitioners to determine the adequacy of a model, and thus informs decisions on whether to deploy a model to users.[219]

---

216. *See generally* Li & Kanoulas, *supra* note 209 (deciding the stopping point of TAR by jointly training a ranking model and conducting "greedy" sampling to effectively retrieve relevant documents).

217. *See id.* at 41:1.

218. *See, e.g.*, Karan Goel, Nazneen Rajani, Jesse Vig, Zachary Taschdjian, Mohit Bansal & Christopher Ré, *Robustness Gym: Unifying the NLP Evaluation Landscape*, 2021 PROC. CONF. N. AM. CHAPTER ASS'N FOR COMPUTATIONAL LINGUISTICS 42, 43 (stressing the importance of systematically evaluating machine learning models).

219. IAN GOODFELLOW, YOSHUA BENGIO & AARON COURVILLE, DEEP LEARNING TEXTBOOK 416 (2016).

Validation is typically a multi-stage process.[220] First, practitioners identify one or more evaluation metrics (e.g., accuracy) to measure performance. The choice of metrics depends on the specific context and use of the algorithm, type of task, and nature of the dataset.[221] Second, practitioners apply the model to a collection of data samples known as a "test set."[222] Third, practitioners compute the selected metrics for the model's predictions over this test set.[223] Finally, practitioners analyze the results of these metrics to determine if the model's performance is satisfactory. If practitioners find that performance is unsatisfactory, they will continue to improve the model and repeat the validation process after additional changes.

A validation process faces a series of challenges.[224] As an initial matter, machine learning systems almost never achieve perfect performance.[225] Evaluating adequacy thus requires practitioners to determine the extent and types of errors that can be tolerated in deployment.[226] Moreover, the multitude of available evaluation methods[227] creates a "paradox of choice," making it difficult to choose a particular method. Choosing the wrong validation procedure can be costly — models that perform well under one procedure are sometimes identified as inadequate under different procedures.[228] Finally, the validation method is only as good as the ground truth labels provided by humans. Recent work in machine learning has demonstrated that labels in commonly used datasets are incorrect due to either ambiguity of the label[229] or lack of sufficient background knowledge on the part of labelers.[230] These errors not only hurt true algorithm performance, but also give a false sense of security in accuracy when used in a test set.

---

220. *Id.* at 418.

221. *See id.* at 418–19 (detailing circumstances in which different metrics may be appropriate).

222. *See infra* Section VI.C.

223. MEHRYAR MOHRI, AFSHIN ROSTAMIZADEH & AMEET TALWALKAR, FOUNDATIONS OF MACHINE LEARNING 4 (2d ed., MIT Press 2018).

224. *See* Goel et al., *supra* note 218, at 2.

225. *See* Goodfellow et al., *supra* note 219, at 417 ("Keep in mind that for most applications, it is impossible to achieve absolute zero error.").

226. *See* Chen et al., *supra* note 198, at 1.

227. *See* Goel et al., *supra* note 218, at 42.

228. *See, e.g.*, Yonatan Belinkov & Yonatan Bisk, *Synthetic and Natural Noise Both Break Neural Machine Translation*, 2018 INT'L CONF. ON LEARNING REPRESENTATIONS (finding that state-of-the-art translation models produce poor scores when evaluated on noisy texts that humans easily comprehend); Jia & Liang, *supra* note 158 (proposing a new evaluation scheme for evaluating reading comprehension models and finding that on average, existing models experience a drop in F1 score from 75% to 36%).

229. *See* Mitchell L. Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto & Michael S. Bernstein, *The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics in Line with Reality*, 2021 PROC. CHI CONF. ON HUM. FACTORS COMPUTING, May 8–13, 2021, at 1, 11.

230. *See, e.g.*, Freitag et al., *supra* note 38, at 2.

As part of the effort to improve and standardize validation, scholars have begun to identify common errors in validation and "best practices" to combat these errors. For instance, scholars have noted that reliance on aggregate metrics can lead practitioners to overlook model failures on critical subsets of the data.[231] Some scholars thus counsel that validation should probe for errors in data with common attributes, and — where possible — use metrics over "slices" of the dataset corresponding to distinct samples.[232] Moreover, scholars caution that assertions about the statistical significance of results and measurements should be made with care.[233]

2. Application to Discovery

The TAR validation process depends on metrics that focus on the entire dataset, rendering it subject to potential abuse. Specifically, the process produces two key metrics: an estimate of recall and an estimate of precision. A TAR search is often considered adequate if the estimated recall[234] of produced documents is at least 70%.[235] However, machine learning research suggests that such aggregate calculations of recall can be misleading when they do not account for specific "slices" of the data. If a producing party only validates using a recall computed over the entire dataset, the TAR process may appear facially adequate despite the existence of underlying failures.

Consider the potential problem of "**Obfuscation via Global Metrics**." Suppose that Qualcomm suspects that its TAR model has differential performance over different types of documents. For instance, while the model performs well for technical reports, it struggles for shorter emails, which often contain spelling mistakes and abbreviations.[236] If most of the relevant documents are technical reports, reporting a "global" recall would allow Qualcomm to hide deficiencies. For example, suppose the actual set of responsive documents for a set of RFPs consists of 800 reports and 200 emails. Suppose too that

231. *See, e.g.*, Chen et al., *supra* note 198 (noting the possibility of reduced performance for images containing cyclists in vision systems for autonomous driving).

232. *See id.*

233. Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald & Dan Jurafsky, *With Little Power Comes Great Responsibility*, 2020 PROC. CONF. ON EMPIRICAL METHODS NAT. LANGUAGE PROCESSING 9263, 9263 (2020).

234. Recall is defined as the proportion of relevant documents in the corpus that are successfully identified in production.

235. Grossman & Cormack, *supra* note 33, at 473, 481–82.

236. Model performance can vary significantly based on the nature of the text being processed. *See* Lichao Sun et al., Adv-BERT: BERT Is Not Robust on Misspellings! Generating Nature Adversarial Samples on BERT (Feb. 27, 2020) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/2003.04985.pdf [https://perma.cc/Q9JF-CT6T] (finding NLP models experience significant performance reduction when the underlying text contains typos).

Qualcomm identifies 700 of these reports and 50 of the emails. Qualcomm's validation stage might report that its TAR algorithm achieved a recall of 75%, painting a misleading picture of the algorithm's performance over the email search.[237] This could offer Qualcomm a significant advantage in litigation, as evidence in emails may differ from evidence in technical reports.

To be clear, it is the aggregate metrics used at validation that can cause a problem, not any underlying feature of the TAR search. Validation enables a producing party to certify that its production was reasonable. Gaming of validation and performance thus diminishes a party's opportunity to contest adequacy.

A recent case presents a potentially direct and dramatic example of obfuscation via global metrics. In *Epic v. Apple*, plaintiffs complained that Apple reported only aggregate metrics to mask TAR's potential under-performance.[238] Plaintiffs argued in a Joint Letter Brief Regarding Validation Protocol to the court that 2.2 million out of 3.8 million documents that Apple produced in discovery were all versions of *a single automatically generated email*. Including these 2.2 million documents in the validation set, plaintiffs argued, would result in a misleadingly high recall estimate. Specifically, plaintiffs highlighted that through a global validation statistic "Apple could exceed overall recall of 75% even while achieving recall of just 57% across the population of responsive documents *other than* the 2.2 million iTunes Content Manager emails."[239] This problem emerged even though the agreed-upon protocol stipulated that no document set comprising more than 2% of the total set of documents due to duplication or automatic generation shall be considered in the calculation of recall.[240]

Setting aside the problem of aggregate metrics, there may be other sources of abuse during validation. For instance, a producing party could mislabel documents collected as part of the "evaluation set," thus producing misleading statistics.[241] Attorneys may also count privileged documents when measuring recall, resulting in unrepresentative metrics. Other scholars have noted that problems may arise when attorneys attempt to quantify the amount of statistical certainty they possess in

237. *Supra* note 234. Thus, recall = (700 + 50) / (800 + 200) = 750 / 1000 = 0.75.

238. Joint Letter Brief Regarding Validation Protocol, *supra* note 7, at 3.

239. *Id.* (emphasis in original).

240. Joint Stipulation and [Proposed] Order Re: Electronically Stored Information at 3, Epic Games, Inc. v. Apple Inc., No. 20-cv-05640 (N.D. Cal. 2020).

241. *See, e.g.*, Remus, *supra* note 27, at 1707. Unsurprisingly, labeling errors for evaluation data can make validation processes misleading and suspect. *See* Curtis G. Northcutt, Anish Athalye & Jonas Mueller, Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks 1 (Apr. 8, 2021), (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/2103.14749.pdf [https://perma.cc/HU6Y-57PW].

estimates of recall.[242] Here, attorneys may apply dubious mathematical techniques to present their validation measures as statistically significant.

3. Indicia and Solutions

As with other forms of abuse, there are several indicia that requesting parties can look at to detect faulty validation procedures. First, requesting parties can ask producing parties to compute recall over specific subsets, or slices, of the data in addition to computing over the entire dataset.[243] In the Qualcomm example above, Qualcomm could be required to compute and then report to Broadcom a recall score for each type of document, such as memoranda or emails. In the example given, these scores would be 87.5% over memoranda and 25% over emails.[244] By asking a producing party to compute slice specific metrics, a requesting party is better equipped to identify algorithmic failures.[245] However, this approach could be problematic because TAR practitioners frequently operate in a low prevalence environment — a relatively small fraction of the documents in the corpus are ever actually relevant, raising challenges with validation. In addition, control sets must be large enough to capture a statistically significant number of relevant documents. If a particular subgroup is too small, it would be infeasible to accurately measure recall.

Second, requesting parties can also follow the common machine learning practice of "error analysis."[246] By analyzing the individual errors a machine learning model makes, researchers are able to determine if there are systematic underlying faults in the machine learning system. In the course of validating TAR, producing parties may identify relevant documents mistakenly marked as unresponsive. Requesting

242. *See, e.g.*, Lilith Bat-Leah, *There Is No One-Size-Fits-All Sample Size Appropriate for TAR Validation (Part I)*, JD SUPRA (Oct. 23, 2019), https://www.jdsupra.com/legalnews/there-is-no-one-size-fits-all-sample-46673/ [https://perma.cc/KBK2-M9KJ]; *Considering the Impact of Richness on Control Set Metrics*, FRONTEO (Nov. 7, 2019), https://legal.fronteousa.com/2019/11/07/considering-the-impact-of-richness-on-control-set-metrics/ [https://perma.cc/F9MM-Q2L8].

243. This would follow the practices recommended by machine learning researchers. *See* Chen et al., *supra* note 198, at 1.

244. Recall that in this example, Qualcomm successfully identified 700 out of the 800 relevant memoranda, and 50 out of the 200 relevant emails.

245. *See supra* Parts IV.A–D (discussing hidden stratification, stopping points, and data poisoning).

246. *See, e.g.*, Vikas Solegaonkar, *Error Analysis in Neural Networks*, TOWARDS DATA SCI. (Feb. 6, 2019), https://towardsdatascience.com/error-analysis-in-neural-networks-6b0785858845 [https://perma.cc/UE36-488M] (defining error analysis as the process of analyzing misclassified samples in order to identify the root causes of model errors); Besmira Nushi, *Responsible Machine Learning with Error Analysis*, MICROSOFT (Feb. 18, 2021, 8:00 AM), https://techcommunity.microsoft.com/t5/azure-ai-blog/responsible-machine-learning-with-error-analysis/ba-p/2141774 [https://perma.cc/39DX-KHG8] (describing a toolkit to aid practitioners in the process of conducting error analysis).

parties could inspect these documents to discern underlying errors. If the documents differ significantly from those produced — or contain valuable information not present elsewhere in the production — then the requesting party may have reason to believe that the TAR process was inadequate. Having the ability to inspect TAR's errors thus better equips requesting parties to contest the adequacy of production.[247] Indeed, the DOJ has already adopted this practice.[248]

F. Role of Proprietary Datasets

Finally, we discuss opportunities for abuse that may arise for repeat players who are well positioned to leverage proprietary datasets. In short, repeat players may be able to select algorithms and parameters that can disadvantage opposing parties. Benchmark datasets serve an important role in validating machine learning algorithms. Often, the only way to test the accuracy of an algorithm is to run it on a dataset that has already been manually labeled. The problem we address here is that, typically, repeat players are the only actors with access to high quality benchmark datasets. This, in turn, could increase abuse by providing those repeat players with information on the limitations of TAR algorithms, while simultaneously depriving less sophisticated actors of the means to contest dubious protocols.

1. Computer Science Literature

Datasets are the primary method by which engineers evaluate machine learning algorithms and understand the models they produce.[249] Modern models can consist of millions of numeric parameters[250] and

247. This practice is also recommended by Maura R. Grossman and Gordon V. Cormack. *See* Grossman & Cormack, *supra* note 33, at 436.

248. *See* DOJ Antitrust TAR Model Agreements, *supra* note 89 (outlining how DOJ representatives will review the validation sample and responsiveness predictions); *see also* City of Rockford v. Mallinckrodt ARD Inc., 326 F.R.D. 489, 494, No. 17-cv-50107 (N.D. Ill., Aug. 7, 2018) (ordering a similar validation review); Winfield v. City of New York, No. 15-CV-05236, 2017 WL 5664852, at *11 (S.D.N.Y., Nov. 27, 2017) (ordering a producing party to share non-responsive documents).

249. For an example of this practice, *see, e.g.*, Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li & Li Fei-Fei, *Imagenet: A Large-Scale Hierarchical Image Database*, 2009 IEEE CONF. ON COMPUT. VISION & PATTERN RECOGNITION, June 2009 (presenting the Imagenet benchmark dataset consisting of 14 million annotated images and 20,000 labels, updated on Imagenet website, https://image-net.org/about.php [https://perma.cc/J435-GZRK]).

250. *See, e.g.*, Jacob Devlin, Ming-Wei Chang, Kenton Lee & Kristina Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding 3 (May 24, 2019) (unpublished manuscript) (on file with the arXiv), https://arxiv.org/pdf/1810.04805.pdf [https://perma.cc/3ENN-7QNC] (presenting "BERT," a widely used language model containing 110 million distinct parameters); Tom B. Brown et al., Language Models are Few-Shot Learners 1 (July 22, 2020) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/2005.14165.pdf [https://perma.cc/6N8R-YNXC] (presenting GPT-3, a state-of-the art language model for many tasks which contains 175 billion parameters).

are thus too complex to analyze directly. However, computer science scholars have shown that these models can be examined via their performance on pre-existing datasets. By constructing specialized datasets — and evaluating the predictions of models on these datasets — engineers can better understand the strengths and weaknesses of different algorithms.[251]

Suppose, for instance, that engineers wish to evaluate a new machine learning system for identifying pictures of cats. They manually curate a collection of 10,000 images, labelling each according to whether it contains a cat. This manually-curated dataset is now the "benchmark." By running the machine learning system over the benchmark images and analyzing the predictions, engineers can understand the flaws, errors, accuracy, and completeness of their system. By comparing performance of their new system to their older systems, they can evaluate whether their newer methods have resulted in improvements.

In order to test new machine learning algorithms, some organizations have developed public benchmarks — specially-designed datasets collected and made freely available for use.[252] Public benchmarks offer a common standard upon which engineers can measure performance of different algorithms, thereby creating consensus as to which methods are preferred. Practitioners are often required to justify algorithmic choices via performance on benchmarks,[253] making public benchmarks akin to an informal regulatory mechanism. Public leaderboards that rank machine learning algorithms incentivize academic and industrial research labs to compete for best performances as measured against the benchmark.[254]

---

251. *See* Laurel Orr et al., Bootleg: Chasing the Tail with Self-Supervised Named Entity Disambiguation 19 (Oct. 23, 2020) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/2010.10363.pdf [https://perma.cc/4YUK-4MP8] (discussing that a model's ability to perform certain "reasoning patterns" is evaluated by analyzing performance over benchmark datasets for that purpose).

252. *See, e.g.*, Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy & Samuel R. Bowman, GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding 1 (Feb. 22, 2019) (unpublished manuscript) (on file with arXiv) https://arxiv.org/pdf/1804.07461.pdf [https://perma.cc/6LPM-UXJZ].

253. *See, e.g.*, Pengcheng He, Xiaodong Liu, Jianfeng Gao & Weizhu Chen, *Microsoft DeBERTa Surpasses Human Performance on the SuperGLUE Benchmark*, MICROSOFT RSCH. BLOG (Jan. 6, 2021), https://www.microsoft.com/en-us/research/blog/microsoft-deberta-surpasses-human-performance-on-the-superglue-benchmark/ [https://perma.cc/3L9T-VD6F].

254. *See* Kyle Wiggers, *AI Models from Microsoft and Google Already Surpass Human Performance on the SuperGLUE Language Benchmark*, VENTUREBEAT (Jan. 6, 2021, 11:04 AM), https://venturebeat.com/2021/01/06/ai-models-from-microsoft-and-google-already-surpass-human-performance-on-the-superglue-language-benchmark/ [https://perma.cc/SU8Z-Q4US].

2. Application to Discovery

In the TAR context, parties frequently subject to discovery requests (repeat producers) are advantaged in two distinct ways. First, repeat producers have access to data from prior litigation, enabling them to build better proprietary benchmarks. Producing parties can use these proprietary benchmarks to understand the characteristics of their own TAR algorithms (potentially allowing them to game TAR). Second, existing public benchmarks for TAR suffer from a range of weaknesses. These advantages could make it difficult for requesting parties to hold repeat producers algorithmically accountable, as existing benchmarks may be incapable of demonstrating faults in particular TAR algorithms.

Every time a producing party applies TAR in response to a discovery request, it creates (1) a training set that is used to develop the TAR model, and (2) an evaluation set that is used to validate the results of the model. As repeat producers are frequently involved in litigation — and each suit involves different types of claims — they can accumulate a collection of training and validation datasets corresponding to different discovery requests.[255] By studying how different types of algorithms perform on past datasets, repeat players can better understand where certain algorithms "miss" documents.

For example, Qualcomm could have abused discovery through a mechanism we call **Choosing Algorithms that "Fail Silently."** Suppose Qualcomm discovered in previous litigation that a particular tokenization setting performs well over longer documents but makes critical mistakes over shorter direct messages (say, over an internal company communication platform). Suppose also that most public benchmarks contain only longer documents. This creates an asymmetry of information: Only Qualcomm may be aware of the deficiencies of their tokenization strategy. This information asymmetry, in turn, presents an opportunity for gaming. When asked to perform discovery over these messages, Qualcomm could opt to use the deficient algorithm. As the algorithm is more likely to miss incriminating documents, Qualcomm will deprive the requesting party of potent evidence, thereby gaining an advantage in litigation.

To be clear, the repeat producer advantage stems from the opportunity a producing party has to "practice" abuse on prior datasets and determine the best techniques it can apply in litigation. The advantage does *not* stem from directly applying models learned during previous

255. *See* Engstrom & Gelbach, *supra* note 19, at 1017–18.

productions. As every production centers on different claims, prior models and datasets are rarely directly relevant to current cases.[256]

Moreover, this kind of gaming could be effective only because existing public benchmarks for TAR suffer from several problems: small sizes, an underrepresentation of documents that have been used in prior litigation, and a lack of new document types used in litigation.[257] The lack of public benchmarks helps facilitate TAR abuse and makes it easier for existing vendors to market their products on the basis of potentially weak in-house evaluations.[258]

3. Indicia and Solutions

Benchmark problems can be addressed, of course, by developing improved benchmarks. Although several initiatives have sought to do so (e.g., TREC Workshops), progress is uncertain.[259] In the short term, requesting parties can take the following steps. First, they can ask producing parties to compare the parameters of their proposed TAR protocol to previously executed productions. Evidence that a party is gaming may be found in their inconsistent use of protocols across different productions. Second, requesting parties can ask producing parties to explain their methodology for concluding that certain algorithmic settings are superior. At the very least, the requesting party can contest the process by which the producing party arrived at certain conclusions, thereby preventing the types of bad science or gaming discussed above.

Requesting parties will have to deal with the potential counterargument that benchmarks are largely unhelpful for TAR as every production implicates a unique set of issues. Here, requesting parties can argue that even if machine learning models are specific to datasets and issues, properties of the algorithms used to train those models

256. A notable exception is when producing parties apply TAR to label documents on the basis of privilege. Here, documents privileged in prior litigation are likely to remain privileged for future litigation. *See* Peter Gronvall, Nathaniel Huber-Fliflet, Jianping Zhang, Robert Keeling, Robert Neary & Haozhen Zhao, *An Empirical Study of the Application of Machine Learning and Keyword Terms Methodologies to Privilege-Document Review Projects in Legal Matters*, 2018 IEEE INT'L CONF. ON BIG DATA 8.

257. The most "real world" benchmark we could identify were the tasks used in the 2008–2010 TREC challenges, which involved "mock" complaints and documents collected in connection with litigation involving tobacco companies. *See* Douglas W. Oard, Bruce Hedin, Stephen Tomlinson & Jason R. Baron, *Overview of the TREC 2008 Legal Track*, 2008 TEXT RETRIEVAL CONF., Nov. 2008, at 24, https://trec.nist.gov/pubs/trec17/papers/LEGAL.OVERVIEW08.pdf [https://perma.cc/N8ZX-HEP3].

258. Daniel N. Kluttz & Deirdre K. Mulligan, *Automated Decision Support Technologies and the Legal Profession*, 34 BERKELEY TECH. L.J. 853, 884 (2019).

259. *See* Adam Roegiest et al., *TREC 2015 Total Recall Track Overview*, 24 TEXT RETRIEVAL CONF., Nov. 2015, at 1–3, https://trec.nist.gov/pubs/trec24/papers/Overview-TR.pdf [https://perma.cc/TZ47-N463]; Maura R. Grossman et al., *TREC 2016 Total Recall Track Overview*, 25 TEXT RETRIEVAL CONF., Nov. 2016, at 1–3, https://trec.nist.gov/pubs/trec25/papers/Overview-TR.pdf [https://perma.cc/WYN6-RDSN].

generalize across different datasets. If an algorithm produces subpar models on a dataset, then there are likely specific qualities of that dataset which lead to poor performance. We should expect that the same algorithm — when applied to other datasets — will produce similarly poor models. Thus, benchmarks provide value as exemplars of where algorithms may struggle.

Moreover, benchmarks enable public scrutiny and thus provide for some measure of contestation. They allow consumers of TAR tools — both requesting parties and producing parties — to validate the performance of different vendors and algorithms. Their public nature would diminish the likelihood of false misrepresentations, and instead serve to improve trust in TAR. These benefits would go beyond short-term considerations regarding the adequacy of a particular production, and instead strengthen the longer-term viability of TAR as a whole.

* * * * *

Taking all six mechanisms together, it appears that TAR abuse is possible and can weaken the discovery process. Many of the mechanisms can be triggered without a high level of technical sophistication. Lawyers can induce biased seed sets, data poisoning, and hidden stratification with early choices in the TAR process, such as choosing a bad sample of documents, using the same algorithm to search for multiple RFPs, employing outdated OCR models, forgetting to de-duplicate a database, and choosing an early stopping point. In all of these instances, lawyers do not need complicated tools or a team of engineers to exploit a vulnerability. Figure 6 summarizes the chronology of TAR abuse.

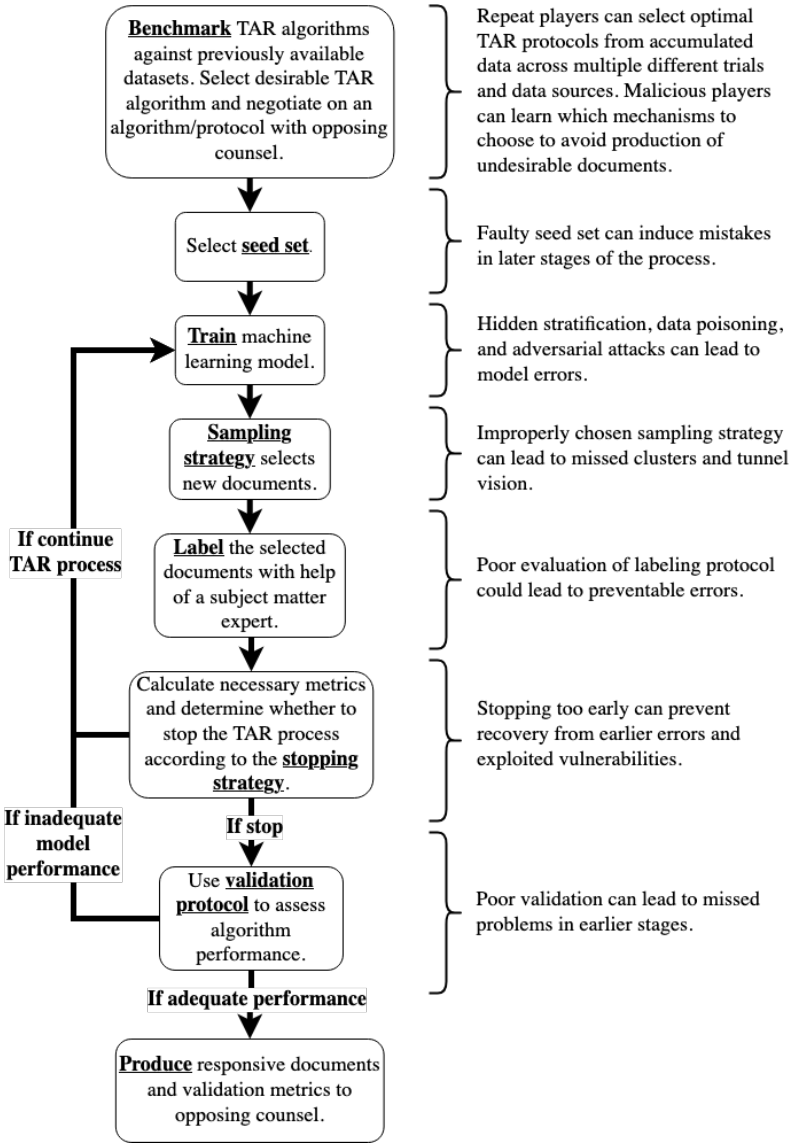

Figure 6: A Diagram of the TAR Process and Where Abuse Mechanisms are Introduced

A deep exploration of the system shows that, as theorized, TAR can face the potential problems of scalability and propagation, a false sense of security, and low visibility. TAR abuse is thus possible and brings a set of complications missing in traditional discovery.

V. EVALUATING TAR ABUSE: POSSIBLE BUT PREVENTABLE

In this Part we explore the possibility of TAR abuse, the risks it poses, and potential prevention through metrics and review processes. We make the argument in two steps:

In Section V.A. we argue that many of the abuse mechanisms discussed above may be deterred by the threat of sanctions because the abuses require intentional malfeasance that goes beyond gamesmanship. The discovery system *already* accounts for the possibility of such intentional abuse and attempts to deter it with sanctions. Moreover, intentional abuse can be ameliorated by existing best practices.

In Section V.B., we highlight how TAR may thrive in the non-sanctionable context of gamesmanship. This development may therefore call for changes to the discovery rules and how we police the system. In sum, we arrive at a middle-ground conclusion: TAR abuse is possible and sloppiness likely — but it is also less of a danger than expected.

A. Existing Sanctions and Counter-Moves Limit the Risks of TAR Abuse

Even if TAR abuse is possible, some of the abuse mechanisms are difficult to execute and are likely to be deterred by existing sanctions. Any common-sense theory of discovery abuse and deterrence would predict that the more sanctionable behaviors are less likely to occur. In order to determine which TAR abuse mechanisms present new risks, then, we first disaggregate them into methods that are (a) arguably sanctionable, or (b) mere gamesmanship. Categorizing TAR abuse in this manner can help us determine whether new tools are needed to police such actions. With these categories in mind, we then conclude that the dangers of TAR abuse are preventable.

1. Many Forms of TAR Abuse Are Already Sanctionable

As discussed above, discovery abuse can encompass different levels of discovery misfeasance, from intentionality to gamesmanship. With that in mind, it appears that many of the six mechanisms of TAR abuse seem to entail either intentionality or negligence while others can be mere gamesmanship. And this distinction matters because the rules already account for (and punish) mechanisms that are intentional or negligent[260] — no new rules or changes are likely needed. We should worry, however, about mechanisms that present new and more dangerous forms of gamesmanship. Therefore, it is important to disaggregate

260. *See supra* notes 92–97.

TAR abuse into behaviors that are sanctionable and behaviors that comprise acceptable gamesmanship.

As an initial matter, this Section will discuss TAR abuse that is intentional or negligent. Some variants of data poisoning, for instance, require that attorneys either unnecessarily duplicate documents or make technical alterations to document features. Consider one hypothetical in the example discussed above. A Qualcomm engineer actively ensures the overrepresentation of irrelevant documents that nonetheless contain the key phrase "JVT." The engineer would have to act with full knowledge of the consequences or with gross negligence, such that a court could infer intentionality, exposing the engineer to sanctions. Similarly, the canonical demonstration of an adversarial example involves intentionally altering the subfeatures of a document. This behavior resembles the non-TAR *Guarisco* case discussed above, where there was evidence of alteration of a photograph.[261] Both of these examples present the kind of intentional sabotage that falls firmly in the sanctionable category.

Yet, a few particular examples of the mechanisms we discuss are much more difficult to detect than traditional sanctionable conduct. For example, the manipulation of a machine learning model's vocabulary size or tokenization mechanism is difficult to discover. Including or excluding certain tokens from the vocabulary may be a perfectly reasonable design decision. Consider our (perhaps extreme) hypothetical above, where a Qualcomm TAR vendor tailors the algorithm's vocabulary to the optimal composition to avoid the "JVT" emails. Although sanctionable in spirit, it would be difficult for opposing counsel to prove negligence or bad faith.

Some other mechanisms — versions of adversarial examples, hidden stratification, aggregate metrics, and stopping points — seem to fall in the gamesmanship category. Consider the use of poor text-recognition software that introduces spelling mistakes into a dataset or model. If the attorney does this intentionally, she is subject to sanctions. But if she does this negligently or merely fails to supervise a sloppy process, it may not be sanctionable. Even more, the hidden stratification problem of stacking multiple requests (for documents A and B) into a single model seems to be acceptable gamesmanship. A producer can stack requests without any intention to bias a model or hide documents, even if they are aware that this is not an optimal search. Similarly, producers can choose stopping points that are early in the process (failing to produce relevant documents) with no intention to hide specific documents. Indeed, this resembles the gamesmanship of producing documents at a difficult time or place. Moreover, a producer is not obligated

261. *See supra* note 96 and accompanying text (discussing Guarisco v. Boh Bros. Constr. Co., 421 F. Supp. 3d 367, 380 (E.D. La. 2019)).

to produce the best available metrics — only reasonable ones.[262] A party can engage in non-sanctionable gamesmanship by producing only aggregate metrics, as long as it does not deliberately misrepresent a search. Finally, the structural advantages that repeat producers derive from benchmarks are not abusive in a recognizable way.

As discussed above, whether the six mechanisms are arguably sanctionable or mere gamesmanship determines whether new tools are needed to deter such actions. For mechanisms that require negligence or malice, such actions should be partially deterred by existing sanctions and potential liability. Thus, our categorization of TAR abuse is consequential and will determine what kind of reforms are necessary.

2. TAR Attacks Are Difficult to Complete, Leading to Partial Attacks that Opposing Counsel Can Counteract

The potential for TAR abuse may be lower than expected because it is often sanctionable, difficult to complete, and can be counteracted by opposing counsel, especially due to TAR's increased transparency regime. We discuss these three arguments in turn.

First, for intentional data poisoning, biased seed sets, and adversarial examples, there is nothing unique about TAR and our current rules may sufficiently deter this misfeasance. Just as altering an image is sanctionable spoliation, so is deliberately altering the words of a seed set document. The rules already account for such misbehavior.[263] The current status quo represents a balance between two pulls, adversarial discovery as broad as possible but with sufficient safeguards that can deter malfeasance. For these behaviors, there is arguably no reason to single out TAR abuse.

Second, the mechanisms of intentional abuse may be easy to trigger at first — but they are more difficult to *complete* successfully. In other words, many of these mechanisms can become what we call "partial attacks" where they fail to completely sabotage the discovery process. When it comes to biased seed sets, data poisoning, and hidden stratification, a malicious attorney may be able to steer the algorithm away from relevant documents but fail to do so completely. This may cause some delay, but opposing counsel may ultimately find the relevant information through further searches or depositions. Indeed, that is exactly what happened in the *Qualcomm* case, where Broadcom used search terms and depositions to uncover discovery abuse.[264] Partial

262. *See, e.g.*, Grossman & Cormack, *supra note* 141, at 313.

263. *See supra* Part III (discussing FRCP 26(g) and 37 and ABA rules).

264. *See* Qualcomm Inc. v. Broadcom Corp., No. 05cv1958-B, 2010 U.S. Dist. LEXIS 33889, at *16–17 (S.D. Cal. Apr. 2, 2010).

attacks are especially likely if the producing party is employing CAL, which depends less on seed sets or model specifications.[265]

But even without CAL, other mechanisms invite only partial attacks. For instance, stacking multiple requests in a single model — leading to hidden stratification — will work mostly when the requests are different in a significant way. Most cases, however, may employ related RFPs that are unlikely to stump the algorithm. This is also true for data poisoning or adversarial examples that at first may be easy to trigger but may later be corrected by the algorithm. Even if the mechanisms are successful, opposing counsel can uncover some of these problems before the end of discovery through depositions, discovery on discovery, or other methods discussed below.

Third, opposing counsel have an array of defense techniques, due to the indicia discussed above or ex ante protocol provisions, that can counteract abuse. For instance, a well-negotiated protocol — with randomized seed set selection, word recognition models for misspellings, de-duplication, splitting data into sub-clusters, machine learning error analysis, etc. — could prevent the easiest forms of abuse. Counsel could also insist on more aggressive monitoring, potentially requesting an audit of the model or better metrics over specific subsets of data.

While each of these arguments are unlikely to prevent all forms of TAR abuse on its own, the combination of all three will cover most mechanisms of TAR abuse. The key question going forward is what happens if one of these three arguments — deterrence, partial attacks, and defense and transparency techniques — are either not available or not possible in a specific case. It is in those situations that TAR abuse is highly dangerous and maybe even likely.

B. TAR, Gamesmanship, and New Sanctions?

The most likely forms of TAR abuse probably fall under the gamesmanship category, where the possibility of sanctions is removed. That is true for versions of data poisoning, adversarial examples, hidden stratification, aggregate metrics, and stopping points. Parties may deliberately play with these hyper-technicalities to weaken their discovery searches and avoid sanctions. In this Section, we address the consequences of TAR gamesmanship, including information asymmetries, producing party advantage, and distributive concerns. We also consider whether the potential for TAR gamesmanship should change our sanctions regime so that more behaviors are punishable.

265. *See supra* Part IV.A.2. (discussing TAR vulnerabilities of SAL and CAL emerging in training process).

1. Information Asymmetries and Moral Hazard

Intentional gamesmanship in TAR can exacerbate the classic concerns of information asymmetry and moral hazard.[266] Many of the problems highlighted above stem from the fact that (a) the producing party has superior information about the TAR pipeline, including the seed set, training process, and incriminating documents themselves; and yet, (b) discovery asks that the producing party act as an agent for the plaintiff in finding the relevant documents. However, due to confidentiality and related concerns, the producing party has residual discretion to manage the process. That discretion, in turn, allows the producing party to hide valuable information. TAR supercharges this information asymmetry and moral hazard by adding a series of new (and potentially opaque) steps that can be gamed.

It seems likely that errors in the TAR process predominantly work in favor of producing parties. That is certainly so with intentional manipulation, which allows the producing party to game the TAR process in a targeted way to hide incriminating evidence. But requesting parties (often plaintiff's attorneys) could also try to game the system to impose costs and acquire privileged information from the producing party (often the defendant's attorneys). As discussed above, a plaintiffs' attorney may negotiate for the inclusion of advantageous seed set documents to take advantage of TAR, or convince a judge to order higher levels of transparency by arguing that TAR is opaque and necessitates closer supervision. Lawyers have complained about plaintiffs' attorneys exploiting the TAR process to increase costs.[267] With all of that said, it is much harder for a requesting party to game the system because they have no control over TAR.

The existence of information asymmetries, moral hazard, and producing party advantage has several implications for discovery. For one, the producing party has weaker incentives to get the process right. Because the cost of false negatives falls exclusively on the requesting party (plaintiffs), the validation process should prioritize avoiding false negatives over false positives. This could have implications for how we calculate recall. Again, this emphasizes that further research is needed to improve ex post metrics like recall or precision.

2. Distributive Concerns

TAR gamesmanship also highlights distributive concerns — the existence of sophisticated producing parties that litigate against less sophisticated requesting parties. Although we highlight a series of indicia

266. We thank Julian Nyarko for some of the specific language here.
267. Payne & Six, *supra* note 18.

and potential responses to abuse, many of these counter-techniques require a high level of resources or technical capacity. Take, for instance, responses to biased seed sets, including algorithmic robustness, optimization approaches, or even testing of algorithms. This is also true for responses to hidden stratification or data poisoning, which may require partitioning data into sub-clusters. Requesting parties would have to negotiate all of these provisions ex ante in the discovery protocol, or ask for specific slices of data in ex post validation. This requires a high level of technical sophistication which may be out of reach for some plaintiffs' attorneys.

Despite this distributive concern, there are reasons to believe this problem is either small or can be alleviated. As an initial matter, the most complex cases will often involve sophisticated counsel on both sides. That is true for antitrust cases, where wealthy competitors are often plaintiffs. And even consumers are often represented by deep-pocketed plaintiffs' firms that employ high-level experts. Setting aside these cases, even in small consumer vs. corporate cases, plaintiffs' attorneys can benefit from existing protocols that are adaptable, like the one used in the *Broiler Chicken* case.[268] Requesting parties can entirely borrow and adapt these protocols to their specific case at low cost.

3. Sanctions for Gamesmanship?

More directly, if TAR opens up an array of new gamesmanship tools, the key question then becomes whether those tools should be sanctioned. In the next Section, we propose a set of potential reforms. Note that any reforms will hinge on further empirical research to determine two underlying variables: the existing degree of disruption and levels of abuse. We simply lack sufficient data to understand whether producing parties often abuse the TAR process or not. We also don't yet know whether TAR gamesmanship is usually counteracted in depositions (as in the partial attacks discussed above). All of this means that the extent of discovery abuse, and the remedies or sanctions available to temper it, are still very much an open question. We therefore conclude that it is too early to tell whether new rules are needed.

* * * * *

All of this suggests that while TAR introduces the real danger of sanctionable abuse, most of the action will be in non-sanctionable gamesmanship. Still, we may wonder, is the system better off with TAR? The answer seems to be yes. Even with the potential for abuse, the algorithmic turn might result in greater transparency for the

268. *See supra* note 88 and accompanying text.

discovery system and prove to be more effective than manual or keyword searching.[269] TAR's emphasis on cooperation and transparency increases the ability of opposing counsel to discover problems in the system. Moreover, searches are likely to be more accurate than in analog discovery. All of this means that even less sophisticated parties will be better off.

VI. SAFEGUARDING TAR & DISCOVERY: BEST PRACTICES, METRICS, AND BENCHMARKS, NOT TRANSPARENCY

This Part proposes several ways to avoid TAR abuse and to improve the current system. Specifically, the most immediate, short-term changes should be towards negotiated protocols, including better metrics, disclosure provisions, and a good faith requirement. In the long term, we flag potential updates to the Sedona Principles with new insights from machine learning research. The goal is to create standardized benchmarks that build on the Sedona standards and better, more cost-effective metrics to verify the TAR process. Before that discussion, however, we first address debates over transparency.

A. The End of Process Transparency and the Rise of Algorithmic Transparency

Transparency has likely reached its limits (under any reasonable cost-benefit analysis) as a solution to TAR problems. Law firms are increasingly arguing that costly ex ante negotiations and transparency obligations are extinguishing the benefits of TAR. In a recent blog post, two attorneys argued that all the baggage added to TAR — including transparency and sharing obligations — has "weaponized [it] to the detriment of both litigants and courts."[270] The attorneys counseled that "[u]sing TAR . . . frequently requires an additional level of transparency, resulting in heavily negotiated, fear-based protocols that can be as expensive as they are cumbersome — without any sort of guarantee of the promised increase in accuracy or decrease in costs."[271] Even more, the attorneys cite a wealth of sources warning about TAR transparency and cooperation. John K. Rabiej, in an email to the Chair of the Advisory Committee on Civil Rules, stated that parties may choose to avoid "TAR rather than incur the costs of extended negotiations and satellite litigation."[272] Additionally, Partner Gareth Evans noted that "the rate of adoption for TAR was slower than initially predicted, in

269. We thank David Engstrom for some of the specific language here.
270. Payne & Six, *supra* note 18.
271. *Id.*
272. *Id.*

part because of fear related to the amount and nature of 'transparency.'"[273]

One need not automatically agree with defense law firms to understand that the decision to use TAR depends on a cost-benefit analysis. The more courts or parties increase the costs of using TAR by imposing burdensome transparency obligations, the less appealing the process. Not only do transparency obligations increase costs, they may also force defendants to reveal more internal documents than they would during manual review.[274] There is suggestive evidence in the case law that courts are having difficulty managing satellite litigation over TAR. In *Rio Tinto PLC v. Vale S.A.* for example, Judge Andrew Peck cautioned that holding TAR to a higher standard than keywords or manual review could dissuade its use, as the costs of TAR-related motion practice would exceed any savings.[275] Courts have encouraged further cooperation, perhaps cognizant that they may lack authority to order certain types of disclosure.[276] Defendants may also be deterred from using TAR by plaintiffs' transparency requirements. In some cases, plaintiffs might insist that defendants produce all documents marked by the algorithm as relevant without any sort of ex post human review.[277]

For these reasons, our proposals below reject traditional transparency and, instead, emphasize the idea of "algorithmic transparency" through better metrics. Even where we propose further disclosure requirements, we also provide less costly alternatives.

B. Short Term: Updating Protocols with Better Metrics and Disclosures

Lawyers should embrace a list of best practices in their negotiated protocols to avoid common forms of abuse. Currently, attorneys in complex litigation borrow from protocols used in previous cases, including the *Broiler Chicken* protocol.[278] We believe some common

273. *Id.*

274. Remus, *supra* note 27, at 1716–17 (noting how non-responsive documents could "include information that reveals unethical or criminal activity by a party, embarrasses an officer or employee, or aids the requesting party in an unrelated cause of action").

275. Rio Tinto PLC v. Vale S.A., 306 F.R.D. 125, 129 (S.D.N.Y. 2015).

276. *See, e.g.*, Aurora Coop. Elevator Co. v. Aventine Renewable Energy-Aurora W., LLC, No. 12CV230, 2015 WL 10550240, at *2 (D. Neb. Jan. 6, 2015).

277. *See* U.S. DEP'T OF JUST., PREDICTIVE CODING MODEL AGREEMENT, *supra* note 89, at 2 (requiring for seed set generation that "[s]earch terms, manual review, or other analytical tools (e.g., email threading) will not be used to collect documents, or to eliminate documents from the collection prior to deduplication or the application of the predictive coding algorithm").

278. *Supra* note 88; *see, e.g.*, James J. Hefferan, *A Game of Chicken? Setting Forth a Detailed TAR Review Protocol*, LEXOLOGY (Feb. 21, 2018), https://www.lexology.com/library/detail.aspx?g=586123a2-2221-429c-bdac-d8323634ef7a [https://perma.cc/6X2F-2MZP] (describing the Broiler Chicken protocol as a "useful framework" that future parties may rely on).

provisions provide a foundation for technical procedures that can uncover manipulation. Among a number of requirements, a good protocol should provide: de-duplication of documents; disclosure of any exceptions where documents were not processed electronically; disclosure of culling parameters (e.g., if documents are excluded on the basis that they are "Windows Operating System files"); disclosure of the TAR vendor and software name; post-hoc validation of recall via random sampling; and a 70–80% recall threshold as "consistent with, but not the sole indicator of," an adequate review.[279] All of these provisions are consistent with our recommendations.

But new protocols could also adopt (a) new metrics, (b) additional disclosure provisions, (c) methods robust to the vulnerabilities we identify here, and (d) "good faith" requirements that cover TAR gamesmanship. The most important protocol update would be in the metrics context. Rather than calculating just simple recall measures (as the protocol currently provides), parties should produce metrics that measure how well distributed the sampling process was among clusters of documents (disclosing the possibility of data bias) and recall rates broken down by clusters of documents.[280] These metrics would flag the vulnerabilities discussed in Part IV. For further details, we provide an extensive discussion on metrics above.

Updates to the protocol should also extend disclosure provisions. In an ideal setting, opposing counsel could take a TAR system and validate it on similar data in their own environment. But disclosures can play a similar role in understanding and evaluating the proposed protocol. While protocol disclosure requirements are highly specific for preprocessing and keyword searches, similar disclosures for TAR are more flexible, requiring only a "general description" of the TAR process.[281] An updated protocol should expand these disclosure requirements to expose potential weaknesses in the TAR process. For example, while the *Broiler Chicken* protocol requires disclosure of "stop words" excluded from keyword searches,[282] there is no similar disclosure requirement of the machine learning algorithm's vocabulary or preprocessing methods.

A potential alternative to disclosure requirements would be to allow an independent expert auditor — perhaps in a special master capacity — to have access to the TAR system and data. The auditor would be bound by non-disclosure requirements with respect to the data itself,

279. Order Regarding Search Methodology for Electronically Stored Information, *supra* note 84, at *2.

280. We recognize that breaking down recall into sub-clusters may be infeasible in some settings, for example where richness is low. *See, e.g.*, Bommannavar et al., *supra* note 153, at 3–4.

281. Order Regarding Search Methodology for Electronically Stored Information, *supra* note 84, at *2.

282. *Id.*

but could nonetheless help a judge determine whether the system performed a reasonable inquiry. Such independent audits have grown in popularity in other machine learning contexts.[283]

The TAR methods themselves should be updated to the most robust versions of machine learning algorithms. Many of the vulnerabilities we describe have proposed solutions in the machine learning literature, as we have described in Part IV. Both parties could agree to a good faith effort to use robust learning methods to build trust in the short term.

Cognizant of the transparency problems discussed above, if costs are too high of a barrier to these disclosure proposals, an alternative might be a "good faith" requirement. Specifically, producers can be required to use algorithms that are robust to several vulnerabilities.

C. Long Term: Sedona Working Group on Benchmark Methods and Additional Research

We suggest three long-term objectives to build trust in TAR systems: (1) the convention of a new working group to identify pre-defined robust TAR protocols and metrics, (2) third-party benchmarks to evaluate TAR protocols, and (3) additional research into improved cost-effective metrics and protocols.

First, we call for the convening of a new working group to study and articulate best practices for TAR benchmarks, protocols, and metrics. Going beyond the Sedona Working Group, a new body should include independent experts in machine learning research, members of the judiciary, vendors of TAR software, and attorneys with expertise in representing both sides of the discovery process. TAR and machine learning experts would help identify effective, yet efficient, standards for algorithms and metrics, as well as setting roadmaps for future innovation. Attorneys and judges would help identify how these technical standards fit into the discovery process.

The new Working Group should study formal mechanisms for evaluating TAR algorithms and vendors, including better public benchmarks. The reforms should aim to reduce TAR negotiation costs while ensuring the uniformity, fairness, and robustness of protocols. We also call for research in cost-effective metrics and algorithms that can detect and prevent the vulnerabilities identified here.

The success of the original Sedona Working Group and the Sedona Principles suggests that a new effort — focused explicitly on the challenges of TAR — would be advantageous in several ways and encourage uniformity. As discussed above, TAR protocols can vary

283. *See, e.g.*, Inioluwa Deborah Raji & Joy Buolamwini, *Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products*, 2019 PROC. AAAI/ACM CONF. ON A.I. ETHICS & SOC'Y 429, https://dl.acm.org/doi/pdf/10.1145/3306618.3314244 [https://perma.cc/6SJS-3UFL].

significantly across litigation. This lack of uniformity increases opportunities for errors and leaves less-sophisticated parties subject to the vulnerabilities we have identified. A new set of best practices would address this problem by promoting standardization, enabling parties to better judge and identify deviations from accepted protocols. New best practices would also diminish any unfair advantages to sophisticated parties. Furthermore, best practices would empower one-shot litigants to identify abuse and evaluate TAR protocols used by their opponents. By leveraging technical expertise for standard setting, best practices ensure that the attorneys need not cede professional jurisdiction during discovery proceedings.

Second, an independent body should develop a new suite of benchmarks to evaluate TAR systems. This could be studied as part of the new Working Group. To improve our understanding of TAR algorithms there is no better alternative to new benchmarks. A range of scholars, from the medical field[284] to criminal law, [285] are convening on benchmarks as a solution to many algorithmic problems. Benchmarks improve accountability and allow the public to verify the robustness of different algorithms. They provide a mechanism to evaluate the benefits and drawbacks of different methodologies.

In the context of TAR, a new set of benchmarks should be composed of documents and information requests from actual prior litigation. The benchmarks should be frequently updated to reflect changing trends in the types of documents encountered in discovery. The new benchmark creators could draw on the work of government organizations that have taken pro-benchmark steps in analogous legal contexts. For instance, a sub-unit of the Department of Commerce runs a "Facial Recognition Vendor Test" consisting of benchmarks for facial recognition algorithms.[286] These benchmarks measure which algorithms are most suited for use by the government and explicitly check for robustness to dataset bias and hidden stratification. Other scholars have recently called for similar benchmarks in the context of medical algorithms used by the Bureau of Veterans Affairs.[287]

Discovery would benefit even more from such government-run benchmarks. The government has access to documents previously accumulated in litigation and FOIA searches. These internal documents

284. Kluttz, *supra* note 258, at 885.

285. Saul Levmore & Frank Fagan, *Competing Algorithms for Law: Sentencing, Admissions, and Employment*, 88 U. CHI. L. REV. 367, 380 (2021).

286. MEI NGAN & PATRICK GROTHER, NAT'L INST. STANDARDS & TECH., U.S. DEP'T OF COM., FACE RECOGNITION VENDOR TEST (FRVT) PERFORMANCE OF AUTOMATED GENDER CLASSIFICATION ALGORITHMS (2014), https://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8052.pdf [https://perma.cc/RK5W-3VJB].

287. Mark Krass, Peter Henderson, Michelle M. Mello, David M. Studdert & Daniel E. Ho, *How US Law Will Evaluate Artificial Intelligence for Covid-19*, 372 BRIT. MED. J., Mar. 15, 2021, at 1, 3.

could be used to mimic real litigation settings without privacy violations. They would also set easy-to-understand standards for which algorithms are acceptable, while encouraging competition among vendors to improve their products.

Third, we encourage more research into both TAR systems and cost-effective ways to evaluate them. We acknowledge that stratified recall evaluation may not be possible in low richness settings. The best preventive mechanisms and benchmarks are not foolproof, require technical sophistication, and can be costly. The benchmarking process itself may provide more confidence that a system works as expected, even when it is difficult to evaluate. Nonetheless, we urge researchers to develop new mechanisms and metrics that are cost effective and practical for use by law firms. Benchmark creators can in turn incorporate more cost-effective metrics into evaluation protocols.

These long-term efforts can provide just some potential steps. In general, we urge more dialog and continuous effort to build more trust and bring more efficiency to the TAR process.

VII. CONCLUSION

In this Article we identified several vulnerabilities in the TAR process that could be used to game the system. While there is a wealth of new abusive mechanisms that exploit TAR vulnerabilities, there are also solutions. To ensure that TAR is utilized effectively and safely, we suggest technical safety mechanisms as well as the updating of current procedures. Abuse of the TAR process is not inevitable. However, with appropriate metrics, benchmarks, and oversight, TAR can equalize the playing field and give more transparency into the discovery process. We urge the discovery community to work toward this goal.

APPENDIX

Please see https://breakend.github.io/TARProtocols/ for a website containing a dataset of TAR protocols, including brief descriptions of all documents in the dataset. The dataset will be updated as new protocols are encountered. We encourage readers to submit additional protocols to the dataset as a resource for researchers, judges, and attorneys. Below is a table of sources in the dataset as of Feb 2, 2020, along with associated links.

Case Name	Links
In re Broiler Chicken Antitrust Litigation	Order Regarding Search Methodology For Electronically Stored Information
Prescott v. Reckitt Benckiser LLC	Joint Status Report Re Technology Assisted Review
Elmgart, Oskar And Nicole J Elmgart v. Ocean Prime, LLC, Columbus Property Management, Inc., The Moinian Group, Centennial Elvator Industries, Inc. And Cba Consultants, Inc.	Proposed ESI Protocol For Defendants' Discovery
850 Third Avenue Owner, LLC v. Discovery Communications, LLC	ESI Protocol Order
Yahoo! Inc. Private Information Disclosure Cases	Joint Status Conference Statement
Emerald Transformer Western States LLC v. Clean Harbors, Inc. and Clean Harbors Disposal Services, Inc.	Subpoena Duces Tecum
Guerbet Ireland Unlimited Company And Liebel-Flarsheim Company LLC v. SPECGX LLC.	E-Discovery Plan
Mani Jacob and Lesleena Mars v. Duane Reade, Inc. and Duane Reade Holdings, Inc.	Stipulation and Scheduling Order Regarding Technology Assisted Review
Rio Tinto PLC v. Vale S.A., et al.	Opinion & Order Titled Predictive Coding a.k.a. Computer Assisted Review a.k.a. Technology Assisted Review (TAR) - Da Silva Moore Revisited

Judith Cole, Louise Michael, and David Johnson v. Keystone RV Company	Stipulation Regarding Discovery of Electronically Stored Information and Order
In re: 3M Combat Arms Earplug Product Liability Litigation	Pretrial Order No. 12 Protocol Relating To Use Of Technology Assisted Review ("Tar Protocol") Pretrial Order No. 44 Supplemental Protocol Relating To Use Of Technology Assisted Review ("Supplemental Tar Protocol")
Joshua Sitzer And Amy Winger, Scott And Rhonda Burnett, and Ryan Hendrickson v. The National Association Of Realtors, Realogy Holdings Corp., Homeservices Of America, Inc., BHH Affiliates, LLC, HSF Affiliates, LLC, The Long & Foster Companies, Inc., Re/Max LLC, and Keller Williams Realty, Inc.	Order Regarding Stipulated Technology Assisted Review Protocol
Livingston v. City of Chicago	Plaintiffs' Motion For Compliance With The Court-Ordered Esi Protocol Or, In The Alternative, For Entry Of A Protocol For Technology Assisted Review [Proposed] Protocol Relating To The Use Of A Continuous Active Learning Tool ("Cal Protocol") Memorandum Opinion And Order
In re Peanut Farmers Antitrust Litigation	Validation Protocol Order

Epic v. Apple	Joint Letter Brief Regarding Validation Protocol Joint Stipulation And Order Re: Validation Protocol
In re: Valsartan, Losartan, And Irbesartan Products Liability Litigation	Protocol Regarding Validation Of Technology-Assisted Review ("Tar")
Port of Vancouver USA v. HDR Engineering, Inc. and Smith-Monroe Gray Engineers, Inc.	Agreement Regarding Discovery Of Electronically Stored Information And [Proposed] Order
In re Diisocyanates Antitrust Litigation	Special Master's Report and Recommendation
Da Silva Moore v. Publicis Groupe & MSL Group	Order and Opinion and ESI order
Bessemer System Federal Credit Union v. Fiserv Solutions, LLC	Stipulated Order on Production of Documents and Electronically Stored Information
DOJ Antitrust Division Sample Predictive Coding Agreement	Department of Justice, the Antitrust Division, Predictive Coding Model Agreement
In re: Actos (Pioglitazone) Products Liability Litigation	Case Management Order: Protocol Relating to the Production of Electronically Stored Information ("ESI")
St. Gregory Cathedral School, ADK Quarter Moon, LLC, Lexmi Hospitality, LLC, and Shri Balaji, LLC v. LG Electronics	Joint Protocol And Order Relating To The Use Of Predictive Coding For Production Of Electronically Stored Information

Matthew Edwards, Georgia Browne, and Torah Montessori School v. National Milk Producers Federation	Joint Stipulation And Order Re The Use Of Predictive Coding Technology