THE COLUMBIA
SCIENCE & TECHNOLOGY
LAW REVIEW

VOLUME 27 STLR.ORG NUMBER 2

ARTICLE

AI SUPPRESSION: E-DISCOVERY SOFTWARE AND *BRADY*

Jason D Hartline,[*] Liren Shan,[†] Alec Sun, [‡] and Rebecca Wexler[§‖]

Prosecutors regularly rely on AI e-discovery software, known as technology assisted review (TAR) tools, to sort and prioritize digital evidence. These tools implicate constitutional concerns: they can either risk suppressing or help to surface exculpatory and impeachment evidence that prosecutors must disclose under the Brady *due process rule. Yet doctrine, agency guidance, and scholarship offer virtually no direction on their use.*

This Article examines how TAR affects Brady *compliance. Using computer science simulations on synthetic data sets, we show that TAR can either hide or help to expose* Brady *evidence, depending on how it is configured and the configurations of evidence to which it is applied. From these results we derive three TAR workflow recommendations: prosecutors should run TAR separately for*

[*] Northwestern University, Department of Computer Science, Evanston, IL, Email: hartline@northwestern.edu.

[†] Toyota Technological Institute at Chicago, Chicago, IL, Email: lirenshan@ttic.edu.

[‡] University of Chicago, Department of Computer Science, Chicago, IL, Email: sundogx@gmail.com.

[§] Columbia Law School, New York, NY, Email: rw3013@columbia.edu

[‖] Authors are listed in alphabetical order. Thank you to Hannah Bloch-Wehba, Chesa Boudin, Andrew Chin, Andrew Ferguson, Eric Fish, Clay Kaminsky, Steven Kochevar, Jennifer Mason McAward, Paul Ohm, Daniel Richman, Stuart Russell, Kelly Scribner, Robert Weisberg, Jonathan Wroblewski, and Christopher Yoo for helpful comments and conversations. This Article benefited from workshops at the University of Pennsylvania School of Law, Columbia Law School, and Cardozo Law School. Thank you to the editors of the Columbia Science and Technology Law Review. A preliminary version of this paper was presented at the 2025 ACM Symposium on Computer Science and Law.

inculpatory and Brady *evidence; TAR coding of* Brady *material should be permitted even when active searching is constitutionally contested; and procurement guidelines should favor flexible classifiers.*

Our examination of TAR also highlights unresolved tensions in Brady *doctrine: whether liability attaches when the prosecution possesses but does not know about* Brady *evidence; how* Brady *interacts with Fourth Amendment privacy protections; and whether* Brady *should be limited to preventing suppression, as current doctrine states, or expanded into a full duty to assist defense investigations. We argue that* Brady *liability should apply strictly to all* Brady *evidence in the control of the prosecution team, regardless of whether anyone on the team knows or has reason to suspect that it exists.*

I. INTRODUCTION

In one of the most well-known court cases in U.S. history, John Brady was convicted of murder and sentenced to death after prosecutors hid the fact that another man had confessed to the crime.[1] The U.S. Supreme Court held that prosecuting Mr. Brady without telling him about the other man's confession violated his Fifth Amendment right to due process.[2] Later cases have clarified the scope of what is now known as the *Brady* rule: due process requires the prosecution to disclose to the defense all material exculpatory or impeachment evidence that the prosecution lawfully "possesses."

This Article examines the risks and opportunities of using AI e-discovery software, known as technology assisted review (TAR) tools, for *Brady* compliance and argues for interpretations of the *Brady* rule that mitigate these risks and enhance these opportunities. TAR technologies increase efficiency by prioritizing documents for human review. Consider a common tool called RelativityOne. When users code documents as relevant or "positive," RelativityOne's machine learning algorithm responds by recommending documents it deems similar for review.[3] This type of workflow is sometimes called *continuous active learning* (CAL).[4] While some e-discovery software also incorporates generative AI features through Large Language Models such as OpenAI's ChatGPT,[5] as of this writing, the generative

[1] Brady v. Maryland, 373 U.S. 83, 84 (1963).

[2] *Id.* at 86.

[3] *See, e.g.*, JUST. MGMT. DIV., U.S. DEP'T OF JUST., PRIVACY IMPACT ASSESSMENT FOR THE DOJ RELATIVITYONE 1 (June 2, 2021), https://www.justice.gov/file/1148006/dl?inline= [https://perma.cc/M3RW-ACLU] ("RelativityOne leverages a 'active learning workflow' to assist users [to] better assess the relevancy of e-discovery documents, actively displaying more relevant data to users based on their actions.").

[4] *See* Gordon V. Cormack & Maura R. Grossman, *Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery*, 37 PROC. INT'L ACM SIGIR CONF. ON RSCH. & DEV. INFO. RETRIEVAL 153, 153-62 (2014), https://dl.acm.org/doi/pdf/10.1145/2600428.2609601 [https://perma.cc/UX8H-2PLN]; *see also* RELATIVITY, ASSISTED REVIEW ACTIVE LEARNING GUIDE 1, 5 (May 16, 2025), https://data.aclum.org/storage/2025/01/DHS_www_dhs_gov_data_AI_inventory.pdf [https://perma.cc/74CW-D68N] (explaining how RelativityOne implements the CAL framework); *DHS Enterprise AI Use Cases*, U.S. DEP'T OF HOMELAND SEC., https://www.dhs.gov/ai/use-case-inventory/dhs-enterprise [https://perma.cc/W8HD-4E45] ("Relativity One is a document review platform used to gain efficiencies in document review for litigation, Freedom of Information Act (FOIA) requests, and other arenas where large-scale document review and production are necessary. It is currently in use across several different DHS Components and Offices. AI techniques used: Machine Learning, Clustering, Continuous Active Learning.").

[5] *See, e.g.*, *Relativity aiR*, RELATIVITY, https://www.relativity.com/data-solutions/air/ [https://perma.cc/K2KL-SEG9] (showing how Relativity implements generative AI).

AI components supplement rather than replace the CAL framework.[6] We anticipate that this will remain the case for the foreseeable future; accordingly, this Article focuses on risks and opportunities that arise from the CAL workflow.

We begin by explaining the need for guidance regarding how prosecutors should use TAR to enhance rather than undermine *Brady* compliance. In high data-volume cases where manual review of every document is impossible, whether prosecutors will disclose *Brady* evidence to the defense depends on prosecutorial search strategies.[7] Despite these high stakes, courts, commentators, and Department of Justice (DOJ) policies have largely, if not entirely, overlooked the issue of how prosecutors should use TAR tools to maximize the discovery of *Brady* materials.[8]

Two surprising ambiguities in *Brady* doctrine contribute to the dearth of guidance. *First*, while the case law is clear that prosecutors must disclose evidence "known" to them or others acting on behalf of the prosecution team, it is unclear whether prosecutors are liable for nondisclosure of evidence that the prosecution team literally possesses or controls but that no one on the prosecution team knows about or has reason to know exists. In other words, it is ambiguous whether a *Brady* violation may be inferred from the mere existence of undisclosed *Brady* material, or whether some additional showing of knowledge is required.[9] This ambiguity matters for TAR. Depending on how TAR tools are used, they can increase the risk that *Brady* material will be overlooked by facilitating the prosecution's search for inculpatory evidence while minimizing review of all other materials, thus minimizing the possibility that prosecutors will encounter unknown *Brady*

---

[6] *See, e.g.*, Chris Haley, *Document Review or Chatbot: Which Generative AI e-Discovery Solution Is Right for You?*, RELATIVITY BLOG (Apr. 24, 2024), https://www.relativity.com/blog/document-review-or-chatbot-which-generative-ai-e-discovery-solution-is-right-for-you/ [https://perma.cc/9UKR-U4YP] (explaining how LLMs might summarize documents for human coding rather than requiring a human to review and label the original documents).

[7] Of course, prosecutorial search strategies matter less in jurisdictions that have adopted open-file discovery, where the prosecution may disclose Brady material via document dumps. *See* Ben Grunwald, *The Fragile Promise of Open-File Discovery*, 49 CONN. L. REV. 771, 789-91 (2017). However, even in open-file jurisdictions, not all cases and categories of information are necessarily subject to mandatory disclosure and there may be good reasons for prosecutors to delay disclosing certain non-*Brady* information that implicates privacy or security concerns. *Id.* at 789-90. Optimizing TAR workflows to separate the *Brady* material that must be disclosed should be helpful in these cases. Regardless, the recommendations in this Article should also help to guide defense search strategies for reviewing the document dumps.

[8] *See, e.g.*, U.S. DEP'T OF JUSTICE, JUST. MANUAL § 9-5.000 (2026), https://www.justice.gov/jm/jm-9-5000-issues-related-trials-and-other-court-proceedings [https://perma.cc/4EC6-Q5SQ] (not mentioning e-discovery software or TAR). For a discussion of the gaps in current doctrine, see Part III *infra*. As for prior legal scholarship, Andrew Ferguson has argued that prosecutorial use of data analytics more generally to select and investigate cases should affect *Brady*, but he does not discuss e-discovery software tools. *See generally* Andrew Guthrie Ferguson, *Big Data Prosecution and* Brady, 67 UCLA L. REV. 180 (2020) (not discussing e-discovery software tools' impact on *Brady*).

[9] *Cf.* BENNETT L. GERSHMAN, PROSECUTORIAL MISCONDUCT § 5:11 (2d ed. 2025) ("Clearly, knowledge is the touchstone of Brady.").

evidence that is in any way distinct from the inculpatory content. Yet the doctrine is unclear as to whether liability exists for such missed encounters.

Second, as prior scholars have identified, *Brady*'s relationship to the Fourth Amendment is under-specified.[10] When law enforcement officers execute a warrant to search for digital records, they may seize entire data repositories for later review.[11] In the process, they are likely to seize data unrelated to the warrant.[12] Once officers identify which data are outside the warrant's scope, the Fourth Amendment restricts the government from searching those out-of-scope data unless they get a new warrant. Yet the prosecution still literally (if not legally) possesses the data, so *Brady* arguably obliges them to disclose any material exculpatory or impeachment evidence contained therein.[13] The conundrum risks trapping the government between a Fourth and Fifth Amendment violation.[14] Whether the prosecution is prohibited from searching the out-of-scope data for *Brady* evidence, or required to do so, matters for determining optimal TAR configurations.

To technically analyze TAR configurations, we synthesized data representing possible distributions of inculpatory evidence, *Brady* material, and irrelevant information. We configured the synthesized data to test plausible relationships between the inculpatory evidence, the *Brady* material, the irrelevant documents and the geometry by which the TAR tools organize the data. These relationships include, for example, situations where the *Brady* material is outnumbered by and lacks common characteristics with the inculpatory evidence.[15] We then ran simulation experiments on these data using a standard CAL framework prevalent in the computer science literature.[16]

Our experiments show that, for certain configurations of evidence, running TAR tools twice—initially targeting only inculpatory evidence and subsequently targeting only *Brady* material—is preferable to running a TAR tool once while labeling both inculpatory and *Brady* material as positive. Running TAR tools twice enables more flexible classification and mitigates the risk that prosecutors might stop running the tool too soon, after it surfaces all the inculpatory evidence but before it surfaces all the *Brady* evidence. We therefore recommend dual TAR runs.

---

[10] *See generally* Effy Folberg, *Search Warrants for Digital Speech*, 22 YALE J.L. & TECH. 318 (2020) (discussing how Fourth Amendment doctrine can conflict with *Brady* obligations).

[11] FED. R. CRIM. P. 41(e)(2)(B); *see also* Bihter Ozedirne, Note, *Fourth Amendment Particularity in the Cloud*, 33 BERKELEY TECH. L.J. 1223 (2019) (discussing search and seizure of cloud databases).

[12] *Cf.* Laurent Sacharoff, *The Fourth Amendment Inventory as a Check on Digital Searches*, 105 IOWA L. REV. 1643 (2020) (discussing how law enforcement often gains access to information unrelated to their investigation by obtaining a warrant for databases).

[13] Folberg, *supra* note 10, at 339.

[14] *Id.* at 336-38.

[15] We do not test scenarios in which the *Brady* evidence is a subset of, or very significantly overlaps with, the inculpatory evidence, as such cases are not the ones where TAR tools pose a risk of suppression.

[16] *Cf.* Cormack & Grossman, *supra* note 4, at 154 (describing the use of the CAL protocol in the TAR process).

Prosecutor office policies could mandate dual TAR runs. Alternately, courts could encourage dual TAR runs (and more general optimization of TAR workflows) by clarifying that the doctrine imposes strict liability for nondisclosure of *Brady* evidence that the prosecution team literally possesses, regardless of whether anyone on the prosecution team knows or has reason to know it exists.

We find that the design and use of TAR can also have consequences in the context of the *Brady* Fourth-Fifth Amendment conundrum. Assume that analysts working for the prosecution team use TAR tools while separating data that are in- and out-of-scope for the warrant. Assume as well that they encounter some *Brady* evidence during this separation process, but the Fourth Amendment prohibits them from conducting an independent TAR run across the out-of-scope data to target additional *Brady* material. Our experiments show that, for *Brady* material that does not significantly overlap with inculpatory evidence, if analysts do not label the *Brady* material they initially encounter as positive in TAR coding, then there is an extreme risk that no further *Brady* material will be surfaced during the separation process. We therefore recommend that, if the Fourth Amendment prohibits independent TAR runs across out-of-scope data to seek *Brady* material, courts should at least construe it to allow labeling *Brady* material as positive while analysts are separating data that is in- and out-of-scope to the warrant. Next, we show that even if analysts are allowed to label *Brady* material as positive in such scenarios, developers of TAR tools can further mitigate the risk of suppressing *Brady* evidence by instantiating the CAL framework with more flexible classifiers, like multicalibration.

In short, our experiments establish a downside risk that TAR use could result in the suppression of *Brady* evidence, and an upside potential that TAR could help to surface *Brady* evidence, depending on how TAR tools are used and configured. Notably, we make no attempt to estimate the scale or frequency of this risk or potential in current practice. Our experiments do not use real-world case data or commercially-available TAR tools, and there is a need for future research to undertake experiments that do. Nonetheless, given the high stakes of *Brady* compliance, we believe that initially establishing the existence of this risk and potential suffices to support some immediate law and policy recommendations, especially those that are easy to adopt or offer a more generalizable benefit beyond regulation of TAR alone. Accordingly, we recommend encouraging dual TAR runs for inculpatory and *Brady* evidence; encouraging positive TAR coding of *Brady* material during the process of separating seized data that is in- and out-of-scope for a warrant; encouraging the use of multi-calibrated CAL algorithms; and clarifying that *Brady* imposes a strict liability rule regardless of knowledge.

Part II of this Article describes the current lack of guidance on prosecutorial use of TAR tools. Part III lays out the ambiguities in *Brady* doctrine with implications for TAR workflows and vice versa. Parts IV, V and VI present the methods, results, and limitations of our simulation experiments. Part VII offers practical recommendations that we draw from our simulation results. Part VIII concludes with broader observations about the relationship between *Brady* and technical precision tools for criminal investigations.

II. THE NEED FOR GUIDANCE ON THE USE OF TAR TOOLS IN CRIMINAL PROSECUTIONS

Courts and commentators have grappled with the use of TAR software in civil cases for some time. The first federal judicial ruling to expressly endorse TAR software for civil e-discovery came in 2012.[17] Since then, advocates have argued that these tools increase precision, recall, and efficiency compared to manual attorney document review, [18] and the prestigious Sedona Conference has recommended their use for privilege review.[19] Critics have claimed the tools lack rigorous benchmarking and evaluation;[20] are subject to gaming by litigants who wish to conceal particular items of evidence; [21] erode "lawyers' professional jurisdiction and authority;"[22] and risk undermining the objectivity and efficiency that the tools are intended to enhance. [23] Meanwhile, computer scientists have sought to develop protocols that enable opposing parties to validate the discovery process, while minimizing the privacy loss from the necessary disclosure of irrelevant documents.[24]

Despite this robust engagement with the use of TAR software in civil cases, the legal, technical, and scholarly communities have overwhelmingly failed to address the use of these technologies in criminal discovery. The DOJ's current *Justice Manual* for prosecutors includes a chapter on "Issues Related to Discovery" but never mentions TAR tools.[25] In a recent ninety-two-page synthesis of U.S. and

---

[17] *See* Da Silva Moore v. Publicis Groupe, 287 F.R.D. 182 (S.D.N.Y. 2012).

[18] *See, e.g.*, Cormack & Grossman, *supra* note 4, at 153 ("TAR methods can be both more effective and more efficient . . . .").

[19] The Sedona Conference, *The Sedona Principles, Third Edition: Best Practices, Recommendations & Principles for Addressing Electronic Document Production* cmt. 10(g), 19 SEDONA CONF. J. 1, 156 (2018) (promoting the use of these tools to "reduce privilege review burdens and expedite production" and offering a best practice guide for doing so).

[20] Daniel N. Kluttz & Deirdre Mulligan, *Automated Decision Support Technologies and the Legal Profession*, 34 BERKELEY TECH. L. J. 853, 884-85 (2019) (identifying a lack of "benchmarking tools or rigorous empirical evaluations of systems on the market").

[21] Neel Guha, Peter Henderson, and Diego Zambrano have detailed at least six ways that attorneys in civil litigation could manipulate, or "game," such e-discovery tools to conceal particularly damning items of evidence while appearing to the technologically-unsophisticated as if they were complying with their disclosure obligations. Neel Guha, Peter Henderson & Diego A. Zambrano, *Vulnerabilities in Discovery Tech*, 35 HARV. J.L. & TECH. 581 *passim* (2022).

[22] David Freeman Engstrom & Jonah B. Gelbach, *Legal Tech, Civil Procedure, and the Future of Adversarialism*, 169 U. PA. L. REV. 1001, 1046 (2021).

[23] *Id.* at 1072-73.

[24] For instance, Jinshuo Dong, Jason Hartline, and Aravindan Vijayaraghavan, *Classification Protocols with Minimal Disclosure*, 2022 PROC. SYMP. ON COMPUT. SCI. & L. 67, 67-76, and Jinshuo Dong et al., *Error-Tolerant E-Discovery Protocols*, 2024 PROC. SYMP. ON COMPUT. SCI. & L. 24, 24-35, have developed protocols that minimize privacy losses when a party using e-discovery software validates the performance of that software to the opposing party by disclosing a subset of negative documents to opposing counsel.

[25] U.S. DEP'T OF JUST., JUST. MANUAL § 9-5.002 (2025), https://www.justice.gov/jm/jm-9-5000-issues-related-trials-and-other-court-proceedings [https://perma.cc/4EC6-Q5SQ].

international case law related to TAR tools, the word "Brady" appears nowhere, and the word "criminal" appears only once—in a comment by a foreign court.[26] To be sure, DOJ had a working group focused on e-discovery as early as 1998,[27] and in 2012 it collaborated with criminal defense counsel to publish a protocol for e-discovery in federal criminal cases.[28] That protocol—which has not been superseded—requires prosecutors to disclose voluminous evidence in an accessible format with a table of contents,[29] but it never contemplates the use of TAR software.[30] A welcome exception is the Federal Judicial Center's 2015 guide on criminal e-discovery, which insists that "[i]n voluminous e-discovery cases, parties must be able to rely on document-review software"[31] and that attorneys "should involve individuals with sufficient technical knowledge and experience to understand, communicate about, and plan for the orderly exchange of ESI [electronically stored information] discovery."[32] Even that guide, however, omits any consideration of how TAR tools might hinder or help *Brady* compliance.[33]

Moreover, distinctions between civil and criminal proceedings mean that tools and procedures developed for civil e-discovery may not work well in criminal cases. For instance, in civil cases, TAR performance is sometimes validated by giving opposing counsel a small subset of documents that the tool labeled as negative.[34] That level of transparency may be impossible in a criminal case. For example, prosecutors might seek to use TAR to surface *Brady* evidence in witnesses' prior statements, or "3500 materials." Here, the prosecution needs to identify the *Brady* evidence to disclose it sufficiently in advance of trial to comply with their constitutional due process obligations. However, the remaining non-*Brady* 3500 materials may be withheld until a later round of discovery.[35] Hence, the prosecution may not wish to reveal documents labeled negative to defense counsel to enable collaborative verification of the TAR coding results. Indeed, it will be precisely in those circumstances when prosecutors wish to withhold non-

---

[26] The Sedona Conference, *TAR Case Law Primer, Second Edition*, 24 SEDONA CONF. J. 1, 76 (2023).

[27] Hon. Eric H. Holder, Jr., *In the Digital Age, Ensuring that the Department Does Justice*, 41 GEO. L.J. ANN. REV. CRIM. PROC. iii, vii (2012).

[28] *Id.*

[29] U.S. DEP'T OF JUST. & ADMIN. OFF. OF THE U.S. CTS. JOINT WORKING GRP. ON ELEC. TECH. IN THE CRIM. JUST. SYS., *Recommendations for Electronically Stored Information (ESI) Discovery Production in Federal Criminal Cases* 10 (2012), https://www.justice.gov/archives/dag/page/file/913236/dl?inline [https://perma.cc/2EKX-PZW3].

[30] U.S. DEP'T OF JUST. & ADMIN. OFF. OF THE U.S. CTS. JOINT WORKING GRP., *supra* note 29.

[31] SEAN BRODERICK ET AL., FED. JUD. CTR., *Criminal E-Discovery: A Pocket Guide for Judges* 13 (2015).

[32] *Id.* at 44.

[33] *See generally* BRODERICK ET AL., *supra* note 31 (omitting discussion of TAR tools' utility for *Brady* compliance).

[34] *See, e.g.*, Da Silva Moore, 287 F.R.D. at 187.

[35] The Jencks Act, 18 U.S.C. § 3500 (not requiring disclosure of a prosecution witness's prior statements until after that witness has testified on direct examination at trial).

Brady material from the defense that they need to be especially meticulous about separating out the *Brady* content that must be revealed.

Further, the nature of inculpatory and *Brady* evidence may differ in ways that risk producing technical bias. Items of non-*Brady* inculpatory evidence may share a nexus of characteristics that will help a machine-learning software classify them as similar. For instance, emails related to the commission of a particular crime might all contain the same code word; occur within a limited date range; mention the same locations, names, and things; and be sent from and to the same circle of participants. As a result, if a TAR user identifies one such document as positive, then the TAR tool is likely to predict that the others will be positive too. In contrast, items of *Brady* exculpatory or impeachment evidence may be significantly different from one another and from the inculpatory evidence in a case. For example, emails documenting a defense alibi, third-party guilt, or prior statements of a prosecution witness that reveal bias may have occurred between entirely different sets of communicants in very different time periods and may share no words or technically-identifiable characteristics, either with each other or with the evidence indicating guilt.[36] As a result, TAR may be better at identifying inculpatory evidence and more likely to miss, or suppress, its exculpatory or impeachment counterpart.[37]

To be sure, *Brady* review using TAR tools has a human in the loop. The "CAL" e-discovery protocol examined in this Article presumes that a human will review documents as the TAR software selects them. And it may still violate due process if the prosecution team fails to disclose all *Brady* material, regardless of whether their TAR software flagged that material for review or erroneously passed it by.[38] In short, humans are doubly in the loop because: 1) humans continually review documents during the TAR protocol; and 2) human prosecutors are theoretically liable for *Brady* and discovery violations that arise from TAR tool use.

Some might argue that this human involvement is enough to allay concerns about *Brady* compliance without micromanaging prosecutors' use of TAR technologies. We would disagree. The mere presence of a human in the loop does not obviate the need to specify guidelines for the design and use of TAR tools in criminal cases. As a general matter, humans in the loop are often insufficient to

---

[36] *See* Andrew Guthrie Ferguson, *Real-Time Crime Centers and the* Brady *Puzzle*, 106 B.U. L. REV. (forthcoming 2026) (manuscript at *20-22) (available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6468120 [https://perma.cc/Y4A9-TFL5]) (describing the wide variety of information that can qualify as Brady material).

[37] Indeed, Andrew Chin has pointed out in conversation with this Article that the risk that inculpatory evidence may "cluster" more than *Brady* material creates a generalizable problem for prosecutorial workflows beyond TAR review and arguably requires a more generalized doctrinal response. *See generally* Andrew Chin, *Rebuilding Brady for Machine-Mediated Discovery*, UNC L. STUD. RSCH. PAPER (forthcoming 2026) (available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6432478 [https://perma.cc/E5ZK-N3L2]) (arguing that machine-mediated discovery creates systemic risks for prosecutorial Brady compliance).

[38] *See infra* Section III.A.

safeguard against AI failures. While some scholars of AI liability place faith in having a human in the loop with ultimate responsibility for human-machine decisions,[39] others have argued that such schemes use the human as a "moral crumple zone"[40] to suck liability away from developers and designers without providing the human with sufficient control to perform as a least-cost avoider for accidental harm.[41]

More specifically for *Brady* compliance, (a) humans already are not always complying with *Brady* and (b), as the simulations we provide show, the configuration of the TAR tools affects which documents even get surfaced for human review. Critics have excoriated *Brady* compliance in old-fashioned manual evidence review as ranging from ineffective to fraudulent. One federal judge on the United States Court of Appeals for the Ninth Circuit stated over a decade ago that there "is an epidemic of *Brady* violations," and cited a long string of cases documenting these transgressions.[42] Part of the problem with relying solely on humans in the loop for *Brady* compliance is that even good faith prosecutors can violate *Brady* unwittingly if they do not understand a defendant's theory of the case sufficiently to identify all the evidence that might be favorable and material to the defense.[43]

Further, deterring bad faith violations of *Brady* is an uphill battle. It is hard to hold prosecutors to account for *Brady* violations because it is difficult to discover the violations in the first place; because immunity doctrines shield individual prosecutors from being sued for *Brady* violations; and because the high burden to overturn convictions means that most will be upheld even after an appellate court finds that *Brady* misconduct occurred.[44] Few of those structural barriers have changed over the past decade, with the hopeful exception of Federal Rule of Criminal Procedure 5(f), which requires courts to issue judicial orders for *Brady* compliance that form a non-constitutional basis for sanctioning prosecutors who violate their disclosure duties.[45] While, in theory, the government should bear the

---

[39] *See generally* Rebecca Crootof, Margot E. Kaminski & W. Nicholson Price II, *Humans in the Loop*, 76 VAND. L. REV. 429 (2023) (proposing human-in-the-loop regulations as pathways to effective legal schemes).

[40] *See generally* Madeleine Clare Elish, *Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction*, 5 ENGAGING SCI. TECH. & SOC'Y 40 (2019) (discussing difficulties and ethical implications of human-in-the-loop policies due to misattribution of moral responsibility on the human).

[41] *See, e.g.*, Bryan H. Choi, *AI Malpractice*, 73 DEPAUL. L. REV. 301, 310 (2024) (examining complications of tacking ordinary legal liability rules onto AI decision-making tools because of lack of human control).

[42] United States v. Olsen, 737 F.3d 625, 626 (9th Cir. 2013) (Kozinski, *C.J.*, dissenting).

[43] Indeed, in a recent empirical study of 386 state and federal *Brady* rulings between 2004 and 2022, Jennifer Mason McAward found that forty-two percent of *Brady* violations were unintentional. Jennifer Mason McAward, *Understanding* Brady *Violations*, 78 VAND. L. REV. 875, 880, 901 (2025).

[44] *See id.* at 895-99; Jason Kreag, *The* Brady *Colloquy*, 67 STAN. L. REV. ONLINE 47, 47 (2014).

[45] Due Process Protections Act, Pub. L. No. 116-182, 134 Stat. 894 (2020).

cost of *Brady* violations via overturned convictions or otherwise, in reality, the defense almost certainly bears that cost unknowingly much of the time.

This baseline of minimal *Brady* enforcement against the human in the loop makes it all the more urgent to ensure that adding TAR software to the mix will minimize rather than exacerbate the suppression of *Brady* evidence. This Article begins that work.

III. *BRADY* PUZZLES FOR CASES WITH VOLUMINOUS DATA

How do prosecutors' *Brady* obligations apply in cases where the records are too voluminous for a human to review every document? What if the government discloses all the *Brady* material that members of the prosecution team know about, but misses a document buried deep in a hard drive that no one on the prosecution team has ever laid eyes on or has reason to suspect is there? Would failing to disclose that buried document violate *Brady*? Should it matter if the government has seized the only copy, albeit unwittingly, so there is no other way for even an exceedingly diligent defendant to find it on their own? What if the Fourth Amendment limits whether and how the government can search through data it possesses? To what extent, if any, does or should the availability of TAR search tools change the calculus?[46]

These questions are important but rarely litigated. As a result, courts have had few opportunities to weigh in and little to say about them. There are good and bad reasons for this dearth of litigation.

Starting with the good, some prosecutors' offices have voluntarily adopted policies of affirmatively searching through voluminous evidence to seek out *Brady* and other discovery materials for disclosure to the defense. For instance, the DOJ's *Justice Manual* explains that "'discoverable information,' and *the duty to search for it*, includes information required to be disclosed by Fed. R. Crim. P. 16 and 26.2, the Jencks Act, *Brady*, and *Giglio*, and additional information disclosable pursuant to [DOJ] policy."[47] The *Manual* further instructs: "Generally, all evidence and information gathered during the investigation should be reviewed, including anything obtained during searches or via subpoenas, etc."[48] Hence, DOJ policy on its face requires prosecutors affirmatively to search voluminous hard drives, email accounts, and other seized data for *Brady* materials. Active search policies like the one in the DOJ *Manual* benefit the prosecution as well as the defense. Such policies help prosecutors not only to comply with their discovery obligations but also to identify weaknesses in their own inculpatory case that might affect charging

[46] Thank you to Paul Ohm for prompting us to pose these questions.

[47] U.S. DEP'T OF JUST., JUST. MANUAL § 9-5.002 (2025), https://www.justice.gov/jm/jm-9-5000-issues-related-trials-and-other-court-proceedings [https://perma.cc/4EC6-Q5SQ] (emphasis added).

[48] *Id.*

decisions or assist with trial preparation.[49] At the same time, these policies mean that *Brady* material is less likely to be overlooked and lead to litigation that could clarify the constitutional rules.

There is also an unfortunate practical reality at play in the dearth of litigation about how *Brady* should apply to cases with voluminous data. In the all too likely scenario that some *Brady* material *is* overlooked during the search process and never seen by a human or disclosed to the defense, it will be rare for anyone to discover the omission.[50] This is especially so if the prosecution has the only copy, so the material will never turn up in even the most diligent defense investigation. Undiscovered evidence is unlikely to lead to litigation that could clarify the constitutional rules.[51]

The upshot is that current doctrine has ambiguities for cases with voluminous data. The following discussion explains two such ambiguities, and how the possibilities and pitfalls of TAR that our simulation experiments expose could factor into the doctrinal analyses.

A. The Brady Search Puzzle: Is There a Duty to Search for Evidence that No One on the Prosecution Team Knows About or Has Reason to Suspect Exists?

Establishing a *Brady* violation requires a defendant to prove three elements: "The evidence at issue must be favorable to the accused, either because it is exculpatory, or because it is impeaching; that evidence must have been suppressed by the State, either willfully or inadvertently; and prejudice must have ensued."[52] Whether *Brady* applies to documents within voluminous collections that the prosecution team literally possesses, but that no human knows about or has reason to suspect exists, concerns the "suppression" element. If, in such circumstances, a failure to disclose counts as "suppressing" the evidence, then a *Brady* violation will have occurred. If not, then it won't. Although courts have not yet resolved this question directly,[53] there are good reasons to think they will—and should—find that nondisclosure under such conditions counts as suppression, especially when

---

[49] *Cf.* United States v. Brooks, 966 F.2d 1500, 1502-03 (D.C. Cir. 1992) ("Of course the prosecutor's own interest in avoiding surprise at trial gives him a very considerable incentive to search accessible files for possibly exculpatory evidence, quite independent of *Brady*.").

[50] *See, e.g.*, JaneAnne Murray, *The Brady Battle*, THE CHAMPION (May 2013), https://www.nacdl.org/Article/May2013-TheBradyBattle [https://perma.cc/38T4-W7JV] (discussing difficulty of discovering Brady violations).

[51] *See, e.g.*, McAward, *supra* note 43, at 921-22.

[52] Strickler v. Greene, 527 U.S. 263, 281-82 (1999) (explaining how discovered evidence may lead to constitutional rules).

[53] To be clear, courts have "extend[ed] the *Brady* duty to searches for evidence." *Brooks*, 966 F.2d at 1502. But the leading rulings have done so for information that at least some human on the prosecution team writ large knew or had reason to suspect exists, not for information that no one but a computer has ever seen or could reasonably be expected to anticipate. *See* Valentin v. Mazzuca, No. 05–CV–0298(VEB), 2011 WL 65759, at *17 (W.D.N.Y. Jan. 10, 2011) (summarizing cases).

TAR tools ease the burden of sifting through the records to find *Brady* materials of which the prosecution was previously unaware.

1. Affirmative Search Duties

One reason to think that courts will, and should, find suppression in these circumstances is that the doctrine has evolved over time increasingly to impose affirmative search obligations on the prosecution team. The original *Brady* opinion merely obliged prosecutors to hand over evidence in response to a defense discovery request.[54] The Supreme Court later clarified that the disclosure obligation applies even without a request by the defense.[55] This led to some confusion and a circuit split over whether *Brady* attached, in the words of the D.C. Circuit, "to information that is available to the prosecution but of which none of the prosecutors was aware."[56]

The Court partially resolved that question in *Kyles v. Whitley*, a case often cited for its imposition of a "duty to learn" on prosecutors.[57] Mr. Kyles was convicted of murder and sentenced to death. Police had prior inconsistent statements from two eyewitnesses and other evidence indicating that a different man had committed the murder, but they never told the prosecutor, and the prosecutor, in turn, never told the defense. The Supreme Court held that this failure to disclose violated *Brady*, despite the fact that the prosecutor did not actually know about the undisclosed evidence.[58] The prosecutor had a "duty to learn" about that information from the police.

Hence, the doctrine is clear that prosecutors have some sort of affirmative duty to search for *Brady* material, including searching beyond their own files for evidence in the hands of the police or other members of the prosecution team. As articulated in the DOJ's *Justice Manual*, "It is the obligation of federal prosecutors, in preparing for trial, to seek all exculpatory and impeachment information from all the members of the prosecution team."[59] The courts have also closed the loophole of willful ignorance by clarifying that deliberate misconduct is not necessary to establish a *Brady* claim; inadvertent or negligent nondisclosure can count as

---

[54] *See Brady*, 373 U.S. at 87 ("[T]he suppression by the prosecution of evidence favorable to an accused upon request violates due process . . . .").

[55] *See* United States v. Bagley, 473 U.S. 667, 682 (1985); United States v. Agurs, 427 U.S. 97, 103-07 (1976).

[56] *Brooks*, 966 F.2d at 1502.

[57] Kyles v. Whitley, 514 U.S. 419, 437 (1995).

[58] *Id.* at 438.

[59] U.S. DEP'T OF JUST., JUST. MANUAL § 9-5.001 (2020), https://www.justice.gov/jm/jm-9-5000-issues-related-trials-and-other-court-proceedings [https://perma.cc/4EC6-Q5SQ].

suppression as well.[60] So the law incentivizes prosecutors to attempt a thorough search by making them at least theoretically liable for overlooked evidence.[61]

Nonetheless, the precise scope of this duty to learn remains ambiguous.[62] *Kyles* states "that the individual prosecutor has a duty to learn of any favorable evidence *known to* the others acting on the government's behalf in the case, including the police."[63] What does it mean for evidence to be "known"? Is information *known* if one physically possesses it but has never seen it and has no reason to suspect that it exists? Post-*Kyles* cases expounding on the "duty to learn" do not answer this question because they have generally concerned either information that *some* human on the prosecution team knew about (even if the prosecutor did not),[64] or information buried in databases that are possessed by third parties.[65] Courts and commentators often express the rule in the conglomerated phase "possession and knowledge," as in "discovery of information in the government's possession and knowledge," without contemplating information that might be physically possessed but remain unknown.[66]

Meanwhile, legal scholarship in the area has focused on determining how far the duty extends beyond a prosecutor's immediate office to the files of other government agencies or entities that may be deemed part of the prosecution team.[67] For instance, Jonathan Abel has written about whether the duty applies to police

---

[60] *See, e.g.*, CHARLES ALAN WRIGHT & ARTHUR R. MILLER, FEDERAL PRACTICE AND PROCEDURE § 256 (4th ed.) (analyzing cases that refined and debated the *Brady* rule over time, finding suppression regardless of prosecutors' intent or negligence).

[61] *Id*.

[62] *See, e.g.*, Jonathan Abel, *Cop-"Like"*, 74 STAN. L. REV. 1199, 1252 (2022) ("The sweet spot for *Brady* is evidence that members of the prosecution team know about but that a reasonably diligent defendant could not access on her own."); Jonathan Abel, Brady's *Blind Spot: Impeachment Evidence in Police Personnel Files and the Battle Splitting the Prosecution Team*, 67 STAN. L. REV. 743, 753 (2015) ("This creates a difficult line-drawing problem for which Supreme Court cases provide no definitive answer: How far does the prosecutor's "duty to learn" extend? . . . Clearly, the prosecutor's duty to learn cannot extend infinitely.").

[63] *Kyles*, 514 U.S. at 437 (emphasis added).

[64] *See, e.g.*, United States v. Price, 566 F.3d 900, 903 (9th Cir. 2009) (emphasizing the prosecutor's failure to learn information known to others acting on the government's behalf); Abel, Brady's *Blind Spot*, *supra* note 62, at 757-58 (discussing the limits of *Brady*'s "constructive knowledge" doctrine in the lower federal courts).

[65] *See generally* United States v. Gray, 648 F.3d 562 (7th Cir. 2011) (involving a database possessed by a third-party government contractor).

[66] *See, e.g.*, 22A C.J.S. *Criminal Procedure and Rights of Accused* § 387 (2026) ("However, *Brady* and its progeny do not require prosecuting authorities to disclose exculpatory information to defendants that the state does not have in its possession and that is not known to exist.").

[67] *See, e.g.*, Sarah Patterson, Note, Co-*Opted Cooperators: Corporate Internal Investigations and* Brady v. Maryland, 1 COLUM. BUS. L. REV. 417 (2021) (considering the effect of investigative outsourcing on *Brady* obligations); Joshua A.T. Fairfield & Erik Luna, *Digital Innocence*, 99 CORN. L. REV. 981, 1030-31, 1038-42 (2014) (discussing intersection of prosecutorial use of Big Data and mass surveillance tools with *Brady* obligations); Brandon L. Garrett, *Big Data and Due Process*, 99 CORN. L. REV. ONLINE 207, 211 (2014) (emphasizing the need to reconsider *Brady* obligations in the context of government data collection).

officers' social media posts.[68] Andrew Ferguson has analyzed the interaction between *Brady* obligations and prosecutors' use of vast, interconnected databases fed by multiple law enforcement and other government agencies.[69] Ferguson identifies the risk that these database systems might be designed to flag inculpatory but not exculpatory data and raises concerns that such technical bias could violate *Brady* due process requirements—concerns that we share and amplify in this Article.[70] Yet Ferguson never directly examines the question of whether or to what extent *Brady* attaches to information that no one on the prosecution team knows or has reason to know exists.[71]

The upshot is an open doctrinal question with an arc through the *Kyles* case law that bends towards increasing duties on the prosecution affirmatively to search voluminous records for *Brady* materials.

2. Document Dumps

What is colloquially called the "document dump" doctrine offers another reason to conclude that the prosecution "suppresses" *Brady* evidence if it fails to disclose evidence in voluminous records that the prosecution technically possesses but has no reason to know exists. Document dumps are when the prosecution hands over troves of data to the defense—think millions of documents—without first identifying the *Brady* evidence contained therein. Although the Supreme Court has yet to weigh in on the issue, most lower courts that considered document dumps concluded that the prosecution can comply with *Brady* solely by handing over all the evidence, without any obligation to affirmatively search that evidence for exculpatory information and flag it for the defense.[72] Put another way, a prosecutor

---

[68] Abel, *Cop-"Like"*, *supra* note 62.

[69] *See generally* Ferguson, *Real-Time Crime Centers*, *supra* note 36; Ferguson, *Big Data Prosecution and* Brady, *supra* note 8.

[70] Ferguson, *Big Data Prosecution and* Brady, *supra* note 8, at 243 (identifying the issue of "how to build systematic procedures and regulations to find information when the systems were not originally designed to flag potential *Brady* information").

[71] Instead, Ferguson's analysis focuses on evidence that at least one person on the prosecution team, broadly defined, knows or once knew about, such as witness statements given to police in alternate cases or a person's affect visible on a body camera recording that would also presumably have been observed by the police officer wearing that body camera. As a result, he presumes that, in his words, "the duty to search for *Brady* information remains clear (even in big data systems)." *Id.* at 241.

[72] *See, e.g.*, United States v. Warshak, 631 F.3d 266, 297-98 (6th Cir. 2010) (affirming that the government performed its *Brady* obligations by providing millions of pages of evidence, despite the onerous format, because the conduct was not in bad faith); United States v. Ohle, No. S3 08 CR 1109(JSR), 2011 WL 651849, at *4 (S.D.N.Y. Feb. 7, 2011) (holding that "the Government did not violate *Brady* by failing to specifically highlight . . . particular documents"); Report and Recommendation, United States v. Perraud, No. 09–60129–CR, 2010 WL 228013, at *12 (S.D. Fla. Jan. 14, 2010) (stating that the government "need not do more than it already has in directing . . . to the . . . universe of [documents]"), *approved and ratified by* 2010 WL 298601 (S.D. Fla. Jan. 20, 2010). There are two caveats: (i) the prosecution cannot pad the document dump deliberately to conceal exculpatory evidence; and (ii) the prosecution must hand over the documents in a format that makes them readily searchable by the defense. *See, e.g.*, Holder, *supra* note 27, at viii (noting

can comply with *Brady* by providing the defense with a document dump regardless of whether the prosecutor identifies or even knows about all *Brady* evidence contained therein.[73]

The document dump cases are not exactly on point for our issue because they concern documents that *were* given to the defense, not documents withheld. Nonetheless, they suggest that withholding exculpatory or impeachment documents would count as a *Brady* suppression even if prosecutors are unaware of their existence. The point is that, in the document dump cases, courts held that prosecutors complied with *Brady* because they *handed over* the *Brady* evidence, despite failing to flag it and potentially not knowing it exists. By inference, prosecutors who do not hand over such materials would risk violating *Brady*. Hence, the DOJ's *Justice Manual* advises prosecutors that: "In cases involving voluminous evidence obtained from third parties, prosecutors should consider providing defense access to the voluminous documents to avoid the possibility that a well-intentioned review process nonetheless fails to identify material discoverable evidence."[74] The strategy of over-compliance via document dump implies that narrower, more targeted disclosures might violate *Brady* by failing to disclose evidence that the prosecution has inadvertently overlooked.

To be sure, it is also possible to read the document dump cases as holding that prosecutors need *not* search through every file in their possession. Courts could arrive at this conclusion regardless of whether the defense is given access to a dump with buried exculpatory evidence or not. But ruling that *Brady* merely requires the prosecution to affirmatively search for evidence that a human on the prosecution team knows about or has reason to suspect exists, rather than to search every document known or unknown, would be a mistake. It would bless cases in which the prosecution withholds *Brady* evidence that would be impossible for the defense to access any other way. The result would undermine accuracy in investigations and re-open the loophole of willful ignorance that *Kyles* sought to plug.

Meanwhile, a minority of courts have come out the other way and ruled that, "at some point 'disclosure,' in order to be meaningful, requires 'identification' as well."[75] If more courts pick up that thread and oblige prosecutors to identify *Brady*

---

that *Brady* is generally satisfied "when the government in good faith provides ESI, voluminous as that ESI may be, in a format that readily allows defendants to search for and locate potentially exculpatory material").

[73] *Cf.* United States v. Skilling, 554 F.3d 529, 577 (5th Cir. 2009) (suggesting that, absent bad faith, the government's provision of a voluminous open file is unlikely to violate *Brady*).

[74] U.S. DEP'T OF JUST., JUST. MANUAL, § 9-5.002 (2025), https://www.justice.gov/jm/jm-9-5000-issues-related-trials-and-other-court-proceedings [https://perma.cc/4EC6-Q5SQ]; *see also id.* ("[I]n cases involving a large volume of potentially discoverable information, prosecutors may discharge their disclosure obligations by choosing to make the voluminous information available to the defense.").

[75] United States v. Salyer, Cr. No. S–10–0061 LKK [GGH], 2010 WL 3036444 at *6 (E.D. Cal. Aug. 2, 2010); *see also* United States v. Salyer, Cr. No. S–10–0061 LKK [GGH], 2011 WL 1466887 at *4-6 (E.D. Cal. Apr. 18, 2011) (describing a "'too much' problem," with too much information from too many different sources, exacerbated by technological difficulty of navigating digital files); Hilary Oran, *Does* Brady *Have Byte? Adapting Constitutional Disclosure for the*

content in document dumps disclosed to the defense, then prosecutors will have to affirmatively search their files for such content before they can comply with their discovery obligations.

TAR tools might change the calculus. Courts' endorsement of document dumps has relied in part on an assumption that the defense is as (if not more) capable than the prosecution of searching through the documents for exculpatory information.[76] So, the more TAR helps defendants the less likely it is that courts will require additional assistance from the prosecution. If prosecutors' offices can access systemically better TAR tools than criminal defense teams, courts might shift the burden of search duties to the prosecution as the least-cost avoider. In contrast, if TAR tools are equally available to the defense, then this would support the existing dump endorsements.

3. Lost or Destroyed Evidence

On the other hand, some existing doctrine suggests that no due process violation occurs if the prosecution fails to disclose *Brady* materials that it literally possesses, even exclusively so, but of which it is entirely unaware.

Specifically, in *Arizona v. Youngblood*, the Supreme Court held that *Brady* was not violated by the prosecution's negligent failure either to test potentially-exculpatory blood and semen stains on a victim's clothing, or to preserve that evidence for defense testing.[77] *Youngblood* fueled a line of doctrine holding that, so long as the prosecution does not act in bad faith, losing or destroying potentially-exculpatory evidence does not violate *Brady*.[78] The line between *Youngblood*'s bad faith requirement and *Kyles*'s "duty to learn" runs through knowledge of the exculpatory nature of evidence. If at least some person on the prosecution team already knows that evidence is exculpatory, then the government is liable for

---

Digital Age, 50 COLUM. J.L. & SOC. PROBS. 97, 126-28 (2016) (discussing *Blankenship* approach to evaluating "data dumps" under *Brady*, requiring identification of "any known favorable evidence within . . . electronic materials" in addition to disclosure).

[76] *See, e.g.*, *Skilling*, 554 F.3d at 577 (noting the "the equal access that Skilling and the government had to the open file"); *Ohle*, No. S3 08 CR 1109(JSR), 2011 WL 651849 at *4 ("Both the Government and defense counsel had equal access to this database. Thus, the defendants were just as likely to uncover the purportedly exculpatory evidence as was the Government."); *cf.* United States v. Rubin/Chambers, Dunhill Ins. Servs., 825 F. Supp. 2d 451, 456 (S.D.N.Y. 2011) (noting that, unlike in *Salyer*, the defense team was not small) (citing *Salyer*, Cr. No. S–10–0061 LKK [GGH], 2010 WL 3036444 at *6-7); *Warshak*, 631 F.3d at 297-98 (noting the lower court's observation that the defense team's motion practice showed it was capable of navigating discovery).

[77] Arizona v. Youngblood, 488 U.S. 51, 58 (1988) ("[U]nless a criminal defendant can show bad faith on the part of the police, failure to preserve potentially useful evidence does not constitute a denial of due process of law.").

[78] *See generally* Evan S. Glasner, Youngblood *in Practice: How the Bad Faith Standard Preserves Wrongful Convictions and Creates Perverse Incentives*, 75 RUTGERS L. REV. 1307, 1309 (2023) (exploring the practical reality and challenges of the "nearly impossible to meet" *Youngblood* bad faith standard).

nondisclosure. However, if the evidence is merely *potentially* exculpatory, then the *Youngblood* bad faith standard arguably controls.

Courts could conceivably characterize evidence that the prosecution possesses but of which it is unaware as potentially exculpatory, thus falling on the *Youngblood* side of the line. In this case, the prosecution's negligent failure to unearth exculpatory evidence buried in voluminous records would not trigger a due process violation. Scholars have criticized the *Youngblood* doctrine as creating "perverse incentives for law enforcement with regard to evidence collection and retention."[79] If the doctrine is extended to overlooked evidence, it would offer little incentive for prosecutors to optimize TAR workflows to surface *Brady* content.

4. Balancing

Finally, courts might adopt a balancing approach whereby the easier it is for the prosecution to search voluminous records, the more the courts will impose an affirmative duty to do so. Pre-*Kyles*, a number of federal circuit courts held that *Brady* disclosure duties do not apply when "the possibility of exculpatory materials [is] purely speculative."[80] The D.C. Circuit proposed something of a balancing test: "As the burden of the proposed examination rises, clearly the likelihood of a pay-off must also rise before the government can be put to the effort."[81] *Kyles* cut short that line of reasoning by imposing a non-balancing duty to learn of evidence known to other members of the prosecution team. But for unknown evidence, the balancing option remains.

Once again, TAR tools might change the calculus. If TAR workflows reduce the overall effort the prosecution must undertake affirmatively to search for information, then a balancing analysis might still require the prosecution to search for even those *Brady* materials that are "purely speculative." Note, there are reasonable scenarios where TAR will lower the overall cost of review while increasing the marginal cost of finding *Brady* material. In other scenarios, TAR will reduce the overall search cost and the marginal cost of finding *Brady*. These potential economic changes might need to go into the calculus of how the courts oversee the handling of *Brady* documents.

In sum, current doctrine is ambiguous about whether *Brady* disclosure duties apply to information that the prosecution literally possesses but that no one on the prosecution team knows about or has reason to suspect exists. The simulation experiments in this Article illustrate why it is important for courts to clarify this question. Specifically, our simulations reveal that TAR configurations and workflows can be optimized to either surface or suppress *Brady* material. If *Brady* imposes strict liability for nondisclosure of evidence that is possessed but unknown, the doctrine is likely to incentivize the former. If *Brady* imposes a mere negligence

---

[79] *Id.* at 1309.

[80] *Brooks*, 966 F.2d at 1503-04 (discussing cases that "sensibly warn against reliance on utter speculation").

[81] *Id.* at 1504.

standard in these circumstances, the doctrine is likely to incentivize the latter. At the same time, the potential for TAR to facilitate affirmative searching for unknown *Brady* material should ease any concerns courts may have that a strict liability rule would impose too onerous a burden on the prosecution. More generally, our simulation experiments show how search methodologies matter for *Brady* compliance in cases where there is too much evidence for the prosecutor to manually review every item.[82] Yet, it would arguably be unwise for courts or legislators to micromanage search methodology directly because rapid technological development could quickly obviate specific rulings. Clarifying instead that *Brady* imposes strict liability would be a future-proof, technologically-neutral approach for optimizing disclosure of *Brady* materials.

B. The Brady *Encounter Puzzle: A Fourth-Fifth Amendment Conundrum for Over-Seized Data*

This Subpart describes another unsettled doctrinal issue that has previously been discussed in the scholarly literature: whether and how *Brady* applies to over-seized data from hard drives, email accounts, and other digital repositories.[83] Once again, how courts ultimately resolve this doctrinal ambiguity will have implications for TAR workflows, and TAR capacities may influence judicial reasoning about resolving the ambiguity.

Here is an example of over-seized data creating ambiguities for *Brady* compliance. In November 2003, the U.S. Army Criminal Investigation Division obtained a warrant to seize hard drives from Taxes International, an accounting service, pursuant to an investigation into one of its accounting clients.[84] The hard drives contained data pertaining to "many other accounting clients" who were not (at that time) suspected of criminal conduct.[85] Nevertheless, the government imaged the entire hard drives so that forensic experts could search them off-site in a controlled environment.[86] The warrant authorizing this broad seizure included a non-exhaustive list of "digital search protocols" that government agents could use to find data that was in-scope for the warrant.[87] Circa December 2004, "the agents had finished identifying and segregating the files within the November 2003 warrant's scope."[88]

---

[82] As Andrew Chin astutely states, "tool-mediated review forces courts to confront a question *Brady* doctrine has largely avoided: what counts as a constitutionally adequate process for finding favorable evidence inside the State's own corpus." Chin, *supra* note 37 at *5.

[83] *See generally* Folberg, *supra* note 10 (exploring tension between "the government's *Brady* obligations to conduct exhaustive searches of data" and Fourth Amendment limitations on digital disclosures).

[84] *See* United States v. Ganias, 824 F.3d 199, 201 & n.2 (2d Cir. 2016).

[85] *Id.* at 202-03 (internal quotation marks omitted).

[86] *Id.* at 201-02 & n.4.

[87] *Id.* at 201 & n.3.

[88] *Id.* at 205 (internal citations omitted).

At this point the remaining, out-of-scope data entered a legal limbo that has yet to be resolved more than two decades later. The Fourth Amendment bars the government from searching such data unless it obtains a new warrant authorizing additional review. How long the government may retain the data, and what it can or must do with the data in the meantime, remain open questions.[89] Most pertinent for this Article, if any of that data contains *Brady* material, then possessing it puts the prosecution between a rock and a hard place. If government agents search it for purposes of complying with the *Brady* rule, they violate the Fourth Amendment. If they do not, they risk violating the Fifth.

Some courts have further tightened the screws on this conundrum by specifying exhaustive search protocols in the initial warrant.[90] Search protocols may be "designed exclusively to search for incriminating evidence,"[91] making *Brady* compliance especially difficult. Restrictions can range from limiting the time period within which the government is allowed to search for data, to specifying the forensic software that the government must or must not use to conduct its search, to walling off case agents from the search process, to destroying or returning out-of-scope data for which the government lacks probable cause.[92] These conditions can preclude discovery and disclosure of *Brady* evidence because *Brady* material may be unrecognizable until later stages of an investigation or by those not intimately familiar with the details of the case.

There are at least three ways that current law governing this conundrum could be resolved. At one extreme, courts could decide that *Brady* does not apply at all to material outside the scope of the warrant because such material is not lawfully within the government's possession. Accordingly, if prosecutors determine that data is out-of-scope, then they should seek to "unsee" that data, even if it contains exculpatory or impeachment evidence. The unseen data would not be subject to *Brady* disclosure because it would be legally de-possessed.

At the other extreme, *Brady* might apply to everything that the government has seized, regardless of whether it falls outside the scope of a warrant. In this scenario, depending on how courts resolve affirmative search duties for unknown *Brady* evidence, prosecutors might be obliged to sift through everything. If they found inculpatory evidence in the out-of-scope data, the Fourth Amendment exclusionary rule might prevent them from introducing that evidence at trial. But if they found

---

[89] *See, e.g.*, Lindell v. United States, 82 F.4th 614, 621 (8th Cir. 2023) (holding that the government cannot retain the nonresponsive data indefinitely); *Ganias*, 824 F.3d at 215 (discussing retention for purposes of authentication and defense access).

[90] *See, e.g.*, United States v. Comprehensive Drug Testing, 579 F.3d 989, 995 (9th Cir. 2009) (en banc) (endorsing warrant's grant of broad seizure subject to certain procedural safeguards and restrictions).

[91] Folberg, *supra* note 10, at 340.

[92] *See, e.g.*, *Comprehensive Drug Testing*, 579 F.3d at 1000-03, *revised and superseded*, 621 F.3d 1162 (9th Cir. 2010) (en banc); *see also* Paul Ohm, *Massive Hard Drives, General Warrants, and the Power of Magistrate Judges*, 97 VA. L. REV. ONLINE 1, 11-12 (2011) (defending such search protocols); Orin Kerr, Ex Ante *Regulation of Computer Search and Seizure*, 96 VA. L. REV. 1241, 1251-58 (2010) (describing *ex ante* search protocols and challenging their constitutionality).

exculpatory or impeachment evidence in the out-of-scope data, then *Brady* would require them to disclose that evidence to the defense.

A third plausible interpretation is as follows: Government analysts must disclose *Brady* material that they encounter in the course of separating files that are in- and out-of-scope for the warrant but are barred from running a separate, independent search targeting solely *Brady* material in files that have already been deemed out-of-scope. We call this interpretation: the *Brady* "encounter rule."

Once again, the simulation experiments in this Article illustrate why it is important for courts to clarify whether and how *Brady* applies to over-seized data that falls outside the scope of a warrant. For instance, simulations S3 vs. S4 (Figures 5.3 & 5.4) and S5 vs. S7 (Figures 5.5 & 5.7) reveal that, for certain configurations of evidence, optimally surfacing *Brady* material will require running a distinct TAR search for that material, as opposed to a search that seeks *Brady* and inculpatory materials simultaneously. If, on the one hand, courts decide that *Brady* does not apply at all to seized material outside the scope of a warrant, then the Fourth Amendment may prohibit the government from running a distinct TAR search for *Brady* content on the out-of-scope data. If, on the other hand, courts decide that *Brady* applies fully to data seized outside the scope of a warrant, then the doctrine will both permit and encourage prosecutors to run independent TAR searches for *Brady* material on all data that the prosecution literally controls. The possibility of designing a TAR search to target solely *Brady* content might encourage courts to adopt this latter, broader interpretation of *Brady*'s application because it means that the broad interpretation would not require giving the prosecution a free pass to rummage unconstrained through the out-of-scope data.

Meanwhile, simulations S1 (Figure 5.1) and S2 (Figure 5.2) show that, if independent TAR runs are prohibited, then for certain configurations of evidence,[93] optimally surfacing *Brady* material will require at a minimum that the human TAR user labels *Brady* materials as "positive" alongside inculpatory evidence (and, as shown by simulation S6, these searches will benefit from a multicalibration tool). Therefore, if courts adopt the "encounter rule" interpretation described above, then prosecutors seeking to maximize discovery of *Brady* materials should still make sure to label as positive any *Brady* evidence encountered during the process of separating data that is in- and out-of-scope for the warrant. The possibility of incorporating TAR labeling of *Brady* material into the separation procedure might also encourage courts to adopt the "encounter rule" interpretation because it may result in surfacing more *Brady* content during initial review while maintaining more Fourth Amendment protections than the broadest application of *Brady* would afford.

The following Parts present our simulation experiments.

[93] Specifically, configurations other than where the *Brady* material is a subset of or very significantly overlapping the inculpatory evidence.

IV. METHODS

This Part presents the experimental methodology for a simulation-based study of various configurations of algorithms for technology assisted review (TAR) on synthetic data sets. There are three main components of the simulation study: the high-level TAR framework, the way the algorithm geometrically represents and separates the data, and how the synthetic data itself is generated.

First, the TAR framework studied is *continuous active learning* (CAL). This framework takes the perspective that all positive documents will be manually reviewed and that the job of the TAR tool is to organize the search through documents to minimize the manual review of negative documents. In this organized search, the TAR tool will iteratively suggest the most relevant remaining documents to manually review, and the process will terminate when the manual review of newly-suggested documents is overwhelmingly negative.

Second, we will consider TAR based on machine learning algorithms that geometrically understand the data in two numerical dimensions. Most of the paper will restrict attention to a canonical approach, from many decades of the literature in machine learning, of separating the data with linear classifiers.[94] In two dimensions, a linear classifier is defined by a straight line that separates the data into points on either side of the line. In this setting, the CAL framework facilitates the search, over all possible lines, for the one that makes the most accurate classification. Figure 4.1 depicts one-dimensional and two-dimensional data and a linear classifier. While algorithms in practice may use more sophisticated methods than linear classification and generally operate in higher-dimensional geometries, the issues that arise in two-dimensional linear classification are broadly representative (see further discussion in Section IV.C).

Third, the synthetic data sets will be generated to highlight different regimes for linear classification in two dimensions. To create these regimes, we use a standard approach: the data points are grouped around a few central locations in two-dimensional space, with each group having a normal distribution (a.k.a. bell-shaped) around its center. These groupings reflect documents with similar characteristics. For example, the relevant documents may be in one grouping and the irrelevant documents may be in another grouping, reflecting the assumption that they have different characteristics. Machine learning with normally-distributed data is standard in the literature,[95] though the fine details of this aspect of our data-generating process are unlikely to have a big impact on any of the results.

Our study of canonical algorithms (i.e., linear classification) on stylized synthetic data (two-dimensional, with different groups from different normal distributions), is consistent with the literature on machine learning simulations with

---

[94] *See generally* TREVOR HASTLE, ROBERT TIBSHIRANI & JEROME FRIEDMAN, THE ELEMENTS OF STATISTICAL LEARNING: DATA MINING, INFERENCE, AND PREDICTION ch. 4 (Springer, 2d ed. 2009).

[95] *See generally* RICHARD O. DUDA, PETER E. HART & DAVID G. STORK, PATTERN CLASSIFICATION ch. 2, 3 & 10 (Wiley-Interscience, 2d ed. 2000).

synthetic data and is likely demonstrative of practical issues that arise, though we will not provide a formal discussion.

A. Simulation Philosophy

The aim of these simulations is to provide qualitative understanding of the risks of various approaches for configuring TAR software in various data regimes. To achieve this qualitative understanding, the simulations will present simple canonical algorithms on stylized synthetic data sets where distinct behaviors arise and where outcomes can be easily interpreted. While the simulations will be for specific algorithms on specific datasets, the stylized nature of the study will allow broader conclusions.

This approach contrasts with two alternate approaches. We do not run commercially-available TAR tools on real data from court cases. While such a study of real-world algorithms and data would be of interest for evaluating the severity of current risks, the blackbox nature of such algorithms and the complexity of the data would not afford the same qualitative conclusions. Without understanding the structural differences in the data, we would be unable to attribute performance differences to canonical structural challenges in machine learning. Moreover, a study of the TAR tools on data from specific cases may not be representative of the data from other cases. Real algorithms may employ heuristics that are tuned to current data sets which might improve performance but then make it hard to attribute risks to specific features of the technology, and then not allow for concrete recommendations for algorithms. Finally, focusing on canonical models and algorithms from the machine-learning literature gives results that are less sensitive to current trends.

We do not conduct theoretical analyses with formal proofs that the behaviors we exhibit broadly occur, or theoretical characterizations of the regimes in which certain behaviors occur. Performing theoretical analyses of the questions this paper addresses would likely require an over-simplified model to be analytically tractable and, from such a model, broad conclusions about realistic scenarios might not be any more attainable.

B. Data and Algorithms

To understand how the use of TAR tools may interact with *Brady* obligations, we overview the data model and algorithmic framework. There is a universe of documents U, a subset P of inculpatory documents which strengthen the prosecution, a subset D of exculpatory or impeachment documents which strengthen the defense (a.k.a. *Brady* material), and the remaining documents $U \setminus P \setminus D$ are irrelevant.

The legal team uses a TAR tool to facilitate the manual review of documents. These tools understand the search by example, i.e., they recommend the review of specific documents and the legal team must label each such document reviewed as either "positive" or "negative." The tool uses this feedback to focus the search on

documents more similar to those labeled as positive. Let $S \subset U$ be the set of documents that the legal team would label as positive in manual review. For example, the prosecution should certainly label the inculpatory documents in P as positive so that the tool targets its search on more documents like these documents.

There are other ways to use the tool. For example, to comply with the *Brady* search rule, the prosecution may run the TAR tools twice. Once with $S = P$ to find the inculpatory documents, and again with $S = D$ to find *Brady* material. This is because effective tools for searching only for inculpatory documents $S = P$, as we will see, minimally encounter the *Brady* data not labeled as positive in the search, $D \cap (U \setminus S)$. However, the law may not always permit running TAR tools twice. In certain applications of the *Brady* encounter rule, the prosecution is only authorized to search for $S = P$, but if the prosecution encounters any D documents during this search, the prosecution must disclose those documents to the defense team. We will recommend, as discussed further below, a different interpretation of the *Brady* encounter rule where a search is authorized for $S = P \cup D$. In other words, when using the TAR tools, both inculpatory documents P and *Brady* documents D are labeled as positive.

Given the universe of documents U and a subset of positive documents $S \subset U$, the *continuous active learning* (CAL) framework is instantiated with a batch size k, a family of classifiers, and a priority function. It works in batches as follows. The initial batch of k documents is selected from all documents U randomly or using a keyword search. Subsequent batches are selected by (a) manually labeling the newly-selected documents, (b) finding the best classifier for all the labeled documents, (c) prioritizing the remaining unselected documents according to how likely they are to be positive, and (d) selecting the top k prioritized documents to manually label for the next batch. The process terminates after several rounds in which no further documents in positive set S are discovered.

For example, Cormack and Grossman use linear classifiers and the support vector machine (SVM)[96] algorithm in step (b) and prioritize unselected documents in step (c) by their distance to the boundary between positive and negative documents according to the selected classifier.[97] Alternatively, probabilistic predictions can be used instead of a classifier in the CAL framework where the forecast probability that a document is positive is its priority in step (c).[98] Recall, the goal of CAL is to automate the manual review of the positive documents while minimizing the manual review of the irrelevant documents, rather than the standard goal of machine learning which is to quickly identify a good classifier.[99]

---

[96] *See infra* Section IV.C for further discussion of SVMs.

[97] Cormack & Grossman, *Evaluation of Machine-Learning Protocols*, *supra* note 4 at 156.

[98] A traditional binary classifier labels documents as either positive or negative. A probabilistic prediction instead gives a probability that each document is positive.

[99] According to the standard machine-learning goal, algorithms would prioritize the manual labeling of points near the boundary rather than the ones that are most likely to be positive. For more details, see the literature on active learning. *See generally* BURR SETTLES, ACTIVE LEARNING LITERATURE SURVEY, COMPUTER SCIENCES TECHNICAL REPORT 1648 (Univ. Wis.-Madison 2009).

A key challenge of using the same TAR tool for selecting inculpatory documents P and selecting exculpatory/impeachment documents D is that these sets of data may have very different properties. This Article considers a number of risks for *Brady* violations associated with using TAR tools in criminal proceedings. These risks include the following.

- The instantiation of the TAR algorithm has *asymmetric labeling* if documents from D are not labeled as positive. In this case, the CAL framework will not be properly utilized to uncover documents from D.[100]
- The data has *asymmetric rate* if the rate of positives for P, i.e., $|P|/|U|$ is significantly higher than the rate of positives for D, i.e., $|D|/|U|$. If the data has this property, then few points from D will be in the initial selection of CAL, and the rate of discovery of D may be significantly slower than that of P.[101]
- The data and TAR instantiation have *incompatible classification* if there does not exist a classifier that correctly classifies $S = P \cup D$ as positive. This risk is especially problematic when P and D can be individually identified to a similar accuracy, i.e., when the classification tasks are not asymmetric (defined next).
- The data and TAR instantiation have asymmetric classification if application of TAR tools to search alone for the inculpatory documents $S = P$ is much more effective than searching alone for the exculpatory/impeachment documents $S = D$. In other words, running CAL to select $S = P$ and or to select $S = D$ have very different accuracies. This asymmetry could be caused from the set of classifiers not being sufficiently rich to identify D. This risk could be especially severe in cases where there is considerable diversity in D.[102]

The challenge of incompatible classification is one that is resolved by simply searching separately for $S = P$ and $S = D$. In some legal environments (discussed later), executing a separate search might be undesirable, thus it may be desirable to have methods for a combined search that are nearly as effective as a separate search.

---

[100] Asymmetric labeling could occur if the prosecution forgets or chooses not to label D, is barred from labeling D by the Fourth Amendment, or is simply less effective at identifying D than P because, for example, it is unaware of the defense's theory of the case and is therefore poorly positioned to anticipate how a particular item of evidence would support the defense's argument.

[101] Asymmetric rate could occur, for instance, if the government seizes hard drives or email accounts primarily because they are likely to contain inculpatory evidence, yet those drives or accounts also happen to contain a smaller quantity of exculpatory or impeachment evidence.

[102] D may be more diverse than P if inculpatory evidence tends to be more homogenous in characteristics than the exculpatory and impeachment evidence. For instance, repeated communications among a group of co-conspirators within a limited time frame may be highly inculpatory and may share characteristics of repeated names and dates. In contrast, exculpatory evidence indicating lack of *mens rea*, an alibi, coercion, or third-party guilt might appear in communications with different, unrelated individuals or in a different time frame. Similarly, impeachment evidence undermining the credibility of a prosecution witness could arise in an entirely different set of communications or time frames. The exculpatory and impeachment evidence would thus share fewer characteristics and appear more diverse than the inculpatory evidence.

A natural approach is to attempt to learn two classifiers at once. When both P and D can be easily searched for separately, then a combination of classifiers should be able to find $S = P \cup D$ together. In fact, this approach is challenging even for the simple model of linear separation that is considered in this Article. In the separable case, while finding P and D separately is computationally easy, finding $P \cup D$ is computationally intractable. In the inseparable case, both problems are considered hard, and while there are many heuristic approaches that work well for linear separation (such as support vector machines, SVMs), there are not generally methods for combining by taking the union of two classifiers.

We propose using multicalibration to extend the flexibility of classifiers used in the CAL framework with the multicalibrated partitioning algorithm of Gopalan et al.[103] This algorithm identifies partitions of the data by their features, using a technique similar to "boosting,"[104] and gives predictions that are statistically accurate on each partition, i.e., predicting p if approximately a p-fraction of the documents in the partition are positive. When P and D are individually separable, the multicalibration approach will probably find $S = P \cup D$ (see Appendix C).

The concern of asymmetric classification, where classifying the exculpatory/impeachment documents is harder than classifying the inculpatory documents, is motivated from the potentiality that the exculpatory/impeachment documents are inherently more diverse than the inculpatory documents. The simulations model diversity as groups with different centers, and documents in each group drawn from a normal distribution around its center. With this model the approach of running CAL independently for the inculpatory and exculpatory/impeachment documents will fail because even running CAL to search for the exculpatory/impeachment documents $S = D$ will fail. Multicalibration, which allows multiple linear classifiers to be combined, is a viable remedy for asymmetric classification.

We evaluate the severity of risks of compliance with *Brady* empirically on simulated data. The simulated dataset allows us to create scenarios where there may be greater risk and evaluate that risk. These evaluations demonstrate the plausibility of the risks. We consider the TAR framework instantiated with linear classifiers, SVM classifiers, and with the multicalibration.

Full details of the algorithms considered are given in Appendix A and a complete discussion of the experimental setup is given in Appendix B.

C. Simulation Generality

The simulation results presented in Part V are broadly representative of outcomes that could be expected from the TAR framework of *continuous active*

---

[103] Parikshit Gopalan et al., *Omnipredictors*, *in* 13TH INNOVATIONS THEORETICAL COMPUT. SCI. CONF. (ITCS 2022), 215 LEIBNIZ INT'L PROC. INFORMATICS 79, § 5 (Mark Braverman ed., Schloss Dagstuhl–Leibniz-Zentrum für Informatik 2022).

[104] *See* Yishay Mansour & David McAllester, *Boosting Using Branching Programs*, 64 J. COMPUT. & SYS. SCI. 103, 103-12, § 3 (2002).

learning (CAL). The simulations predominantly rely on the geometric compatibility of the classes of data and the classifiers used to classify the data, i.e., whether the data is separable into its classes via the classifiers or whether it is inseparable.

The paper focuses on linear classifiers. More sophisticated classifiers, such as support vector machines and deep neural networks, are built on linear classification.[105] The difference between vanilla linear classifiers and both support vector machines and neural networks is the latter project the data into higher dimensional space where the data—that originally was not separable by a linear classifier—is separable (if the projection is good).

Our simulations confirm this perspective. We will compare linear classifiers under separability and inseparability in two dimensions. As described above, data that looks inseparable in two dimensions might or might not become separable in the high-dimensional geometry of a support vector machine. In settings where the data looks separable to SVMs, their performance will resemble the performance of linear classifiers in the two-dimensional separable case. In settings where the data looks inseparable to SVMs, their performance will be more like the performance of linear classifiers in the two-dimensional inseparable case. By studying both the separable and inseparable cases we have studied all cases.

As an illustration of how inseparable data in low dimensions can become separable in high dimensions, consider Figure 4.1. The top row of Figure 4.1 (subfigures a and b) depicts one dimension on which the data is linearly inseparable. The best linear separator is the one depicted (in subfigure b) which classifies as negative the two positive points on the left and all of the negative points; while the positive points on the right are correctly classified as positive. For example, if the positive points on the left are exculpatory/impeachment and the positive points on

---

[105] Support vector machines (SVMs) are a standard ML approach to classification. *See generally* Corinna Cortes & Vladimir Vapnik, *Support-Vector Networks*, 20 MACH. LEARNING 273 (1955). SVMs are commonly used for text classification where the text is projected into high-dimensional numerical space and then the SVM framework is applied to find a linear classifier. *See generally* Thorsten Joachims, *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*, *in* 10th EUR. CONF. ON MACH. LEARNING (ECML 1998), 1398 LECTURE NOTES COMP. SCI. 137 (2005); THORSTEN JOACHIMS, LEARNING TO CLASSIFY TEXT USING SUPPORT VECTOR MACHINES: METHODS, THEORY AND ALGORITHMS (Springer, 1st ed. 2002); Ioannis Tsochantaridis et al., *Support Vector Machine Learning for Interdependent and Structured Output Spaces*, 21 PROC. INT'L CONF. ON MACH. LEARNING 104 (2004). For deep neural networks, see generally Rie Johnson & Tong Zhang, *Effective Use of Word Order for Text Categorization with Convolutional Neural Networks*, 2015 PROC. CONF. N. AM. CHAPTER ASS'N COMPUT. LINGUISTICS 103. For transformers, see generally Zichao Yang et al., *Hierarchical Attention Networks for Document Classification*, 2016 PROC. CONF. N. AM. CHAPTER ASS'N COMPUT. LINGUISTICS 1480; Jacob Devlin et al., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, 2019 PROC. CONF. N. AM. CHAPTER ASS'N COMPUT. LINGUISTICS 4171 (demonstrating mapping documents non-linearly into a latent feature space and then applying a linear classifier). For both support vector machines and deep neural networks, the projection into a high dimensional space aims to render the data separable. As an example, Figure 4.1 contrasts the same data in one dimension (where it is inseparable) and two dimensions (where it is separable).

the right are inculpatory, then such a classifier fails to discover the exculpatory/impeachment documents. The dimension considered in this data is

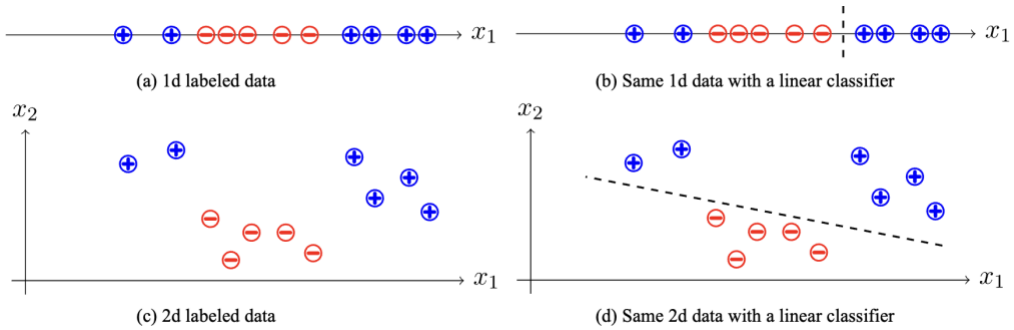

Figure 4.1: 1d and 2d binary classification data and a linear decision boundary. The 1d data is not linearly separable (two points are misclassified); while the 2d data is linearly separable. Note that the x_1 feature is the same in the 1d and 2d data.

feature x_1. Consider including a second feature x_2 and obtaining the data shown in the second row of the figure (subfigures c and d). As shown (subfigure d) the positive and negative points are now linearly separable. Thus, in two dimensions, a linear classifier can successfully classify both the exculpatory/impeachment documents on the left and the inculpatory documents on the right as positive (and the remaining documents as negative).

Two dimensions is especially convenient for depicting situations where the geometry of classifiers and classes of data match (the separable case) or mismatch (the inseparable case). In any application of e-discovery tools, either the geometry matches or it does not match, with the possibility that enhancing the geometry can make it match better. Thus, the focus of the simulations on highlighting the differences between the separable and inseparable case is fully general.

V. TECHNICAL SIMULATIONS

In this Part, we conduct simulations of the CAL algorithm with various configurations on data sets with various structural features. The key variables in the simulations are:

- **labeling: asymmetric or symmetric**: In combined execution,[106] is the labeling of documents raised for manual review:
  - **asymmetric** with only the inculpatory (P) labeled as positive ($S = P$), or
  - **symmetric** with both the inculpatory (P) and *Brady* (D) documents labeled as positive ($S = P \cup D$)?
- **data: rare, individually inseparable, inseparable, or separable**: Is the *Brady* data D
  - **rare** enough that the initial step of CAL is unlikely to discover it,

[106] It is implicit in a separate execution of TAR tools that inculpatory and *Brady* documents are labeled symmetrically.

  - **individually inseparable**,[107]
  - **inseparable** in the combined search for $S = P \cup D$, or
  - **separable** in the combined search for $S = P \cup D$?

  See detailed descriptions in Appendix B.
- **mode: combined, multicalibrated, individual, or SVM**: Is the search for positive documents $P \cup D$ performed in CAL as
  - **combined** into a single execution searching for $S = P \cup D$ with a linear classifier,
  - executed via a more flexible **multicalibration** approach that aggregates multiple linear classifiers,
  - parallel **individual** searches for the inculpatory ($S = P$) and *Brady* ($S = D$) documents, or
  - executed via the **support-vector machine** method (i.e., a kernel embedding into a high-dimensional space where different kinds of documents can be more easily separated)?

  See detailed descriptions in Appendix A.

Each category is ordered in decreasing risk of omitting *Brady* materials. We view the structure of the data as a given, and the specifics of the configuration of CAL as a choice of the prosecution or advice of the court.

A. Main Comparisons

We conduct four main simulation studies investigating the risks outlined in Section IV.B. The four studies aim to illuminate *Brady* risks under *asymmetric labeling*, *asymmetric rate*, *incompatible classification*, and *asymmetric classification*. In our descriptions of these studies, below, we clarify the key parameters above, listing the ones that vary in bold.

Our first set of simulations aims to understand the risks of asymmetric labeling in a single combined execution of the TAR tools, i.e., where only the inculpatory documents are labeled as positive ($S = P$), and we compare this to the case of symmetric labeling where both inculpatory and exculpatory/impeachment documents are labeled as positive ($S = P \cup D$). Given the extreme risk of this configuration of the TAR tools, we evaluate them only on the easy separable data set.

S1. **asymmetric labeling**, separable data, combined mode (Figure 5.1):
Consider the issue of **asymmetric labeling**. We see, even with the easiest data configuration, that asymmetric labeling, i.e., not labeling the exculpatory/impeachment documents as positive, results in significant failures in identifying *Brady* material.

S2. **symmetric labeling**, separable data, combined mode (Figure 5.2):

[107] The individually inseparable data set is described in more detail in the discussion of the third set of simulations.

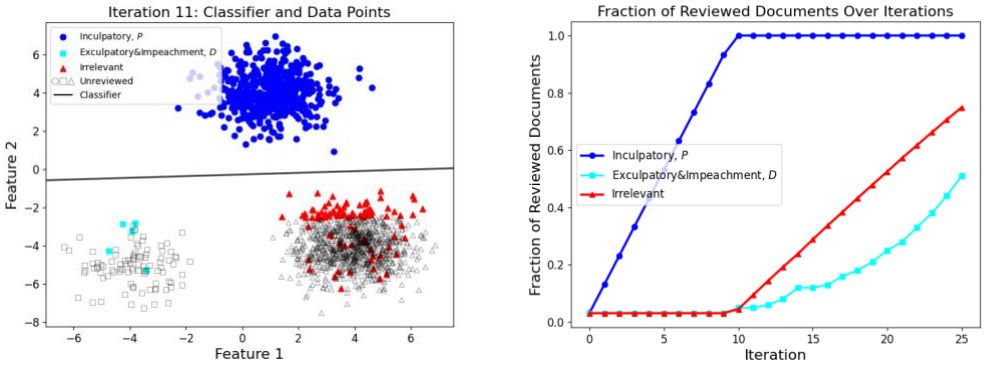

Figure 5.1: Simulation S1. Search of separable data with asymmetric labeling $S = P$. The left column shows the data flagged for manual review by CAL (colored). The right column shows the fraction of each data set flagged by CAL for manual review as a function of the iteration of CAL. Brady documents D are not discovered.

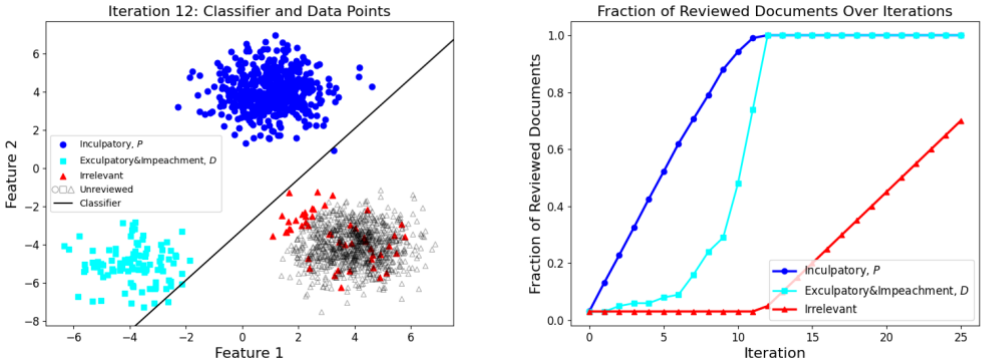

Figure 5.2: Simulation S2. Search of separable data with symmetric labeling $S = P \cup D$, i.e., the benchmark good case. Inculpatory and *Brady* documents $P \cup D$ are discovered. Asymmetric rate results in relatively slower discovery of the *Brady* documents versus the inculpatory documents.

> With symmetric labeling, when the responsive documents ($P \cup D$) are linearly separable, the TAR tools effectively discover both inculpatory and exculpatory/impeachment documents with minimal manual review of negative documents. Simulation S2 is the benchmark good case for a combined execution of CAL. Asymmetric rate results in relatively slower discovery of the exculpatory/impeachment documents versus the inculpatory documents.

Our second experiment, fixing symmetric labeling, aims to understand the risks from asymmetric rate, i.e., when the exculpatory/impeachment documents D are much less frequent than the inculpatory documents. The first step of continuous

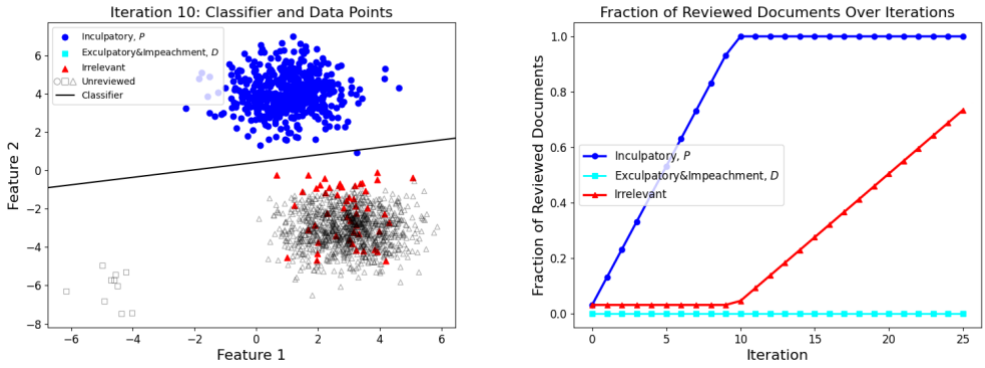

Figure 5.3: Simulation S3. Search of rare data in combined execution mode for $S = P \cup D$. *Brady* documents D are not discovered at all.

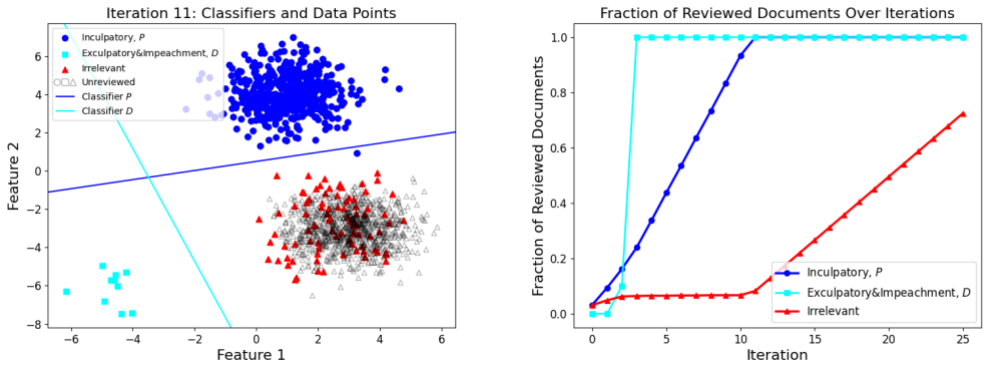

Figure 5.4: Simulation S4. Search of rare data in individual execution mode for $S = P$ and $S = D$. Shortly after the first inculpatory document is discovered, they are all discovered.

active learning is to randomly label documents. If *Brady* documents are very rare then this step may discover inculpatory documents but not discover any *Brady* documents. Without any *Brady* documents in this initial round, it is possible and even likely that they are never discovered.

S3. symmetric labeling, rare data, **combined mode** (Figure 5.3):

We see that when the *Brady* data is rare relative to the inculpatory data that it may never be discovered in a combined search for $S = P \cup D$.

S4. symmetric labeling, rare data, **individual mode** (Figure 5.4):

We see that when the *Brady* data is rare relative to the inculpatory data, individual searches for $S = P$ and $S = D$ discover the *Brady* data as soon as the random selection discovers the first *Brady* documents.

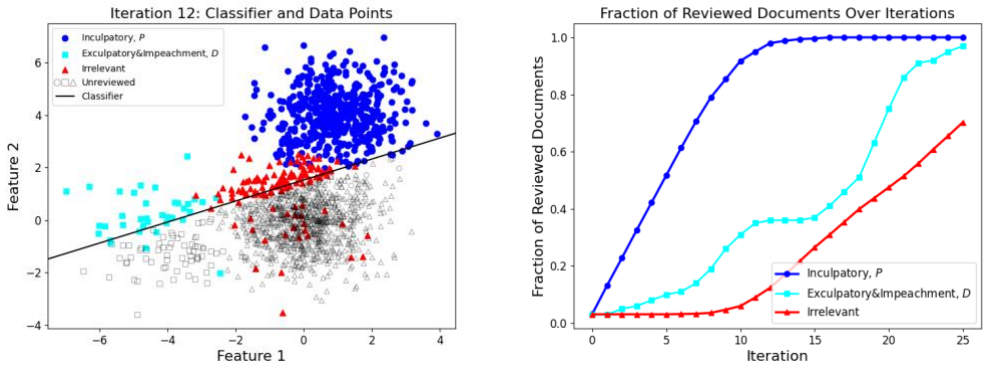

Figure 5.5: Simulation S5. Search of inseparable data in combined execution mode for $S = P \cup D$. Inculpatory documents P are fully discovered well before *Brady* documents D and a considerable number of irrelevant documents are reviewed.

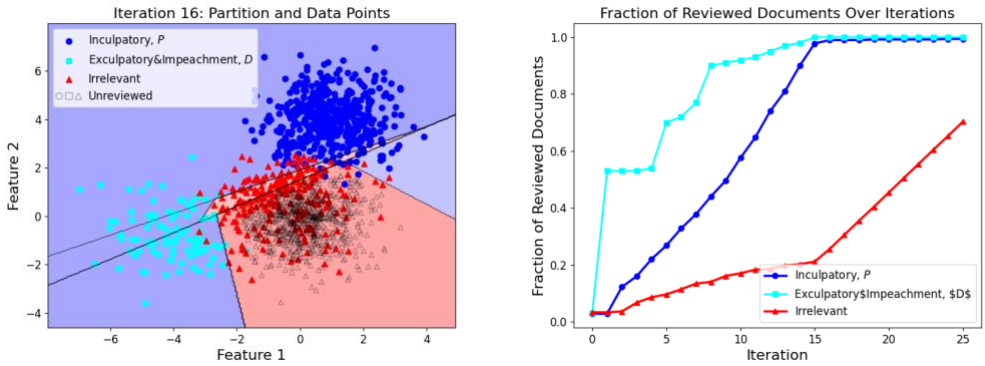

Figure 5.6: Simulation S6. Search of inseparable data with multicalibrated linear classifier and $S = P \cup D$. Inculpatory documents P are fully discovered well before *Brady* documents D and a considerable number of irrelevant documents are reviewed.

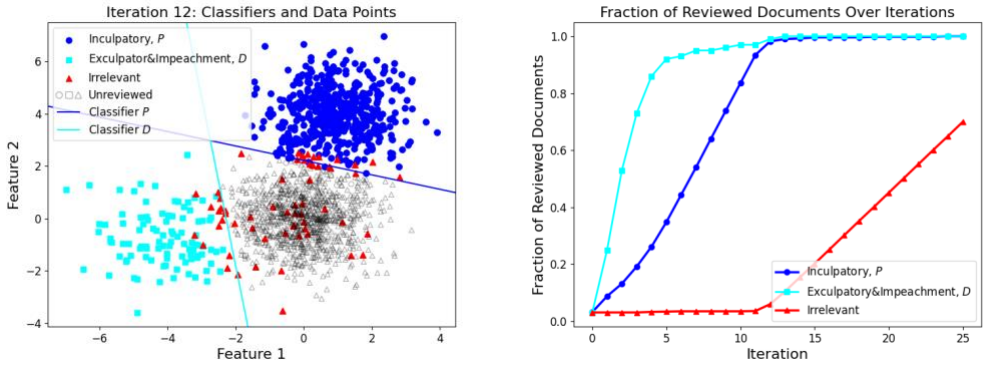

Figure 5.7: Simulation S7. Search of inseparable data in individual execution mode for $S = P$ and $S = D$. Inculpatory and *Brady* documents $P \cup D$ are discovered with the smaller set of *Brady* documents discovered first. Compare to Figure 5.2.

Our third experiment, again fixing symmetric labeling but focusing on inseparable data, aims to understand incompatible classification, i.e., where the inculpatory and *Brady* documents can be identified alone, but not in a combined

search. In these simulations we compare the combined search with linear classifiers, multicalibration, and parallel individual searches.

S5. symmetric labeling, inseparable data, **combined mode** (Figure 5.5):

> We see that when the positive data is inseparable, there are risks of either (a) stopping early (when all of the inculpatory documents (P) are discovered but most of the *Brady* documents (D) are undiscovered or (b) manually reviewing a large amount of irrelevant documents ($U \setminus P \setminus D$).

S6. symmetric labeling, inseparable data, **multicalibrated mode** (Figure 5.6):

> We see that when the positive data is inseparable, combined execution of the TAR tool with the multicalibrated linear classifier finds the less frequent *Brady* documents D before the more frequent inculpatory documents P, but the review of irrelevant documents is significant.

S7. symmetric labeling, inseparable data, **individual mode** (Figure 5.7):

- We see that when the positive data is inseparable, separate executions of the TAR tools with independent stopping conditions, i.e., one execution with $S = P$ and one execution with $S = D$, discovers both the inculpatory and exculpatory/impeachment documents with similar effectiveness.

Our fourth simulation study evaluates *asymmetric classification*, the possibility that it is easier to classify the inculpatory documents P than the exculpatory/impeachment documents D. Specifically, this study considers the data where the inculpatory documents P are separable while the exculpatory/impeachment documents D are inseparable. While D does not exhibit structure that allows it to be separated from the other documents $U \setminus D$, we assume it is comprised of groups of documents, each of which can be separated from the other documents. While each group is self-similar, documents between groups are not similar. Exculpatory/impeachment documents may exhibit such structure more so than the inculpatory documents, hence the inclusion of this study.

S8. symmetric labeling, individually inseparable data, **individual mode** (Figure 5.8):

> As expected, inseparability of $S = D$ implies that even separate executions of CAL on $S = P$ and $S = D$ fails to discover all the exculpatory/impeachment documents D without significant review of negative documents $U \setminus P \setminus D$.

S9. symmetric labeling, individually inseparable data, **multicalibrated mode** (Figure 5.9):

> With multicalibration, because the exculpatory/impeachment documents comprise groups that themselves are linearly separable, multicalibration finds them with similar performance to the case of

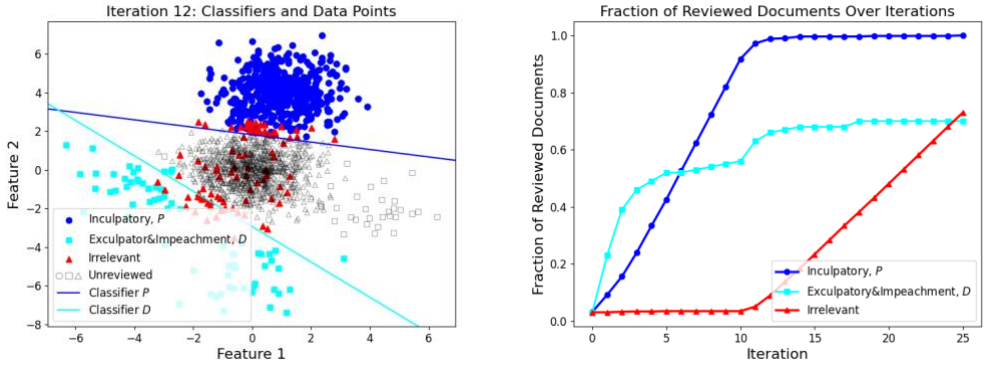

Figure 5.8: Simulation S8. Search of individually inseparable data in individual execution mode for $S = P$ and $S = D$. Higher-frequency inculpatory documents P are discovered before lower-frequency exculpatory/impeachment documents D with minimal review of negative documents.

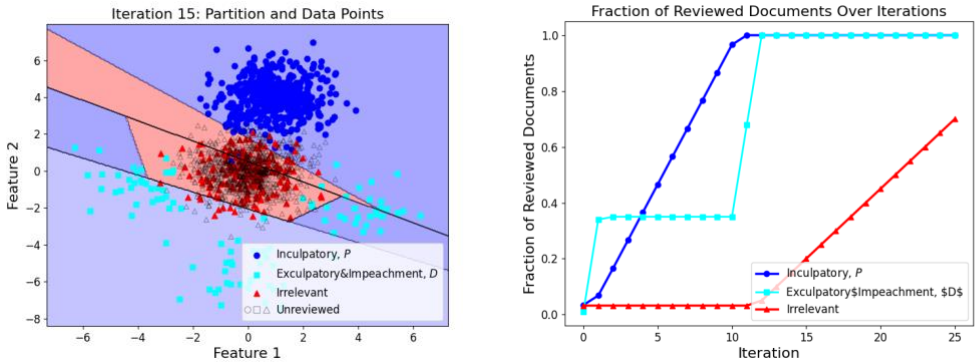

Figure 5.9: Simulation S9. Search of individually inseparable data with multicalibrated classifier and $S = P \cup D$. Higher-frequency inculpatory documents P are discovered before lower-frequency exculpatory/impeachment documents D with minimal review of negative documents.

> inseparable data. However, the discovery of exculpatory/impeachment documents plateaus as the inculpatory documents are discovered, which imparts a risk of stopping early without discovering all of the exculpatory/impeachment documents.[108]

A fifth and final study is to benchmark the previous two studies of incompatible classification and asymmetric classification against support vector machines. In those studies, we showed that individual execution model (where CAL is instantiated separately to search for inculpatory $S = P$ and exculpatory/impeachment $S = D$ documents) and multicalibration, respectively, were good methods for reducing the risk of leaving *Brady* materials undiscovered. In our experimental setup, support vector machines project the two-dimensional feature space into higher dimensions where the data is separable. Thus, we see that

[108] Multicalibration can also be run individually on $S = P$ and $S = D$ and doing so removes the plateau in the discovery of the exculpatory/impeachment documents. Details of this simulation are omitted.

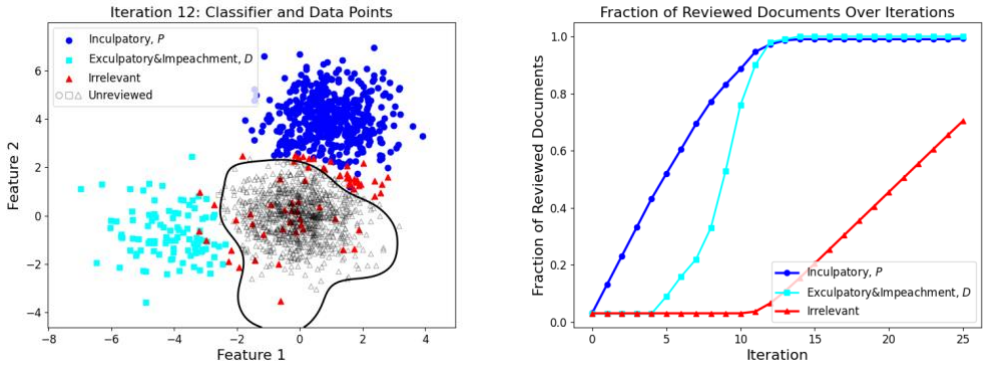

Figure 5.10: Simulation S10. Search of inseparable data with support vector machine classifier and $S = P \cup D$. Higher-frequency inculpatory documents P are discovered before lower-frequency exculpatory/impeachment documents D with minimal review of negative documents. Compare to Figure 5.2.

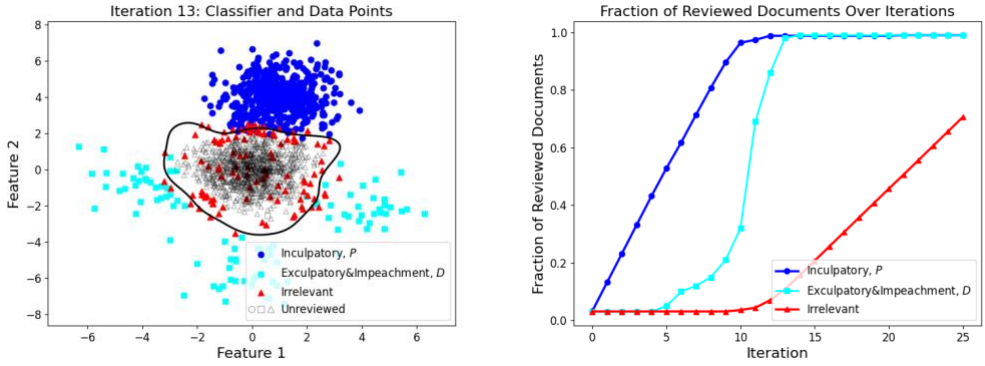

Figure 5.11: Simulation S11. Search of individually inseparable data with support vector machine classifier and $S = P \cup D$. Higher-frequency inculpatory documents P are discovered before lower-frequency *Brady* documents D with minimal review of irrelevant documents. Compare to Figure 5.2.

flexible methods like support vector machines can potentially lessen *Brady* risks as well. Compare these simulations to the good case of simulation S2.

S10. symmetric labeling, **inseparable data**, support-vector-machine mode (Figure 5.10):

> We see that when the positive data is inseparable, combined execution of the TAR tool with a support-vector-machine behaves similarly to combined execution of a linear classifier on separable data (Simulation S2).

S11. symmetric labeling, **individually inseparable data**, support-vector-machine mode (Figure 5.11):

> While the exculpatory/impeachment documents D are not linearly separable in two dimensions, they are in the higher dimensional projection of the support vector machine; thus, we see comparable

performance to combined execution of a linear classifier on separable data (Simulation S2).

B. General Observations

The simulation studies exhibit a few trends across various configurations and data regimes. First, when a combined search for $S = P \cup D$ is separable, then linear classification succeeds, but when there are significantly fewer *Brady* documents than inculpatory documents, i.e., $| D | \ll | P |$, the *Brady* documents D will be discovered more slowly than the inculpatory documents P. We see such outcomes in Simulations S2, S10, and S11. The latter two are the support vector machine approach which, in the case of our simulations, successfully separates the data even though it is not separable in two dimensions. The reason for the lower rate on the smaller set of *Brady* documents is that in a combined search the documents closest to the decision boundary are prioritized and there are more such documents from the larger inculpatory data set P.

On the other hand, both multicalibration (on $S = P \cup D$, in S6 and S9) and parallel individual search (on $S = P$ and $S = D$, in S4 and S7)[109] find the smaller set of *Brady* documents D faster than the larger set of inculpatory documents P. In fact, documents are found at a similar rate, it just takes longer to find the larger set of documents, and thus the search for the smaller set completes relatively faster. The reason that documents are found at a similar rate is because both individual search and multicalibration consider the decision boundaries separately. This behavior contrasts to combined search where, in the good case that the data is separable (discussed above), the smaller set of *Brady* documents D are found at a slower rate.

While multicalibration robustly succeeded in finding the *Brady* documents D, it did so at the expense of requiring more manual review of irrelevant documents. See Simulations S6 and S9.

VI. LIMITATIONS

This Article has presented an empirical study that exhibits risks and opportunities for TAR software and *Brady* compliance. As an empirical study, the exhibited phenomena occur in the settings described and are not known to occur in other settings. The simulation environments studied are stylized environments where outcomes can be easily visualized and understood. Specifically, the study largely considered linear classifiers in two-dimensional feature space. Though this space is very specialized, we expect our results to generalize, and the tradeoffs we have identified are illustrative of tradeoffs in practice.

Our understanding of asymmetric and incompatible classification is the most limited. It generally depends on the data of any given court case whether modern methods such as support vector machines or deep neural networks would result in

[109] Note that individual search when the exculpatory/impeachment data is individually inseparable ultimately fails to find all the exculpatory/impeachment data because it is not separable (Simulation S8), but the same relative rates are observed initially.

the separable case or the inseparable case. If the data is separable, then our main advice of using multicalibration or individually running the TAR tools on $S = P$ and $S = D$ will have limited benefit.

On the other hand, our main findings that asymmetric labeling and asymmetric rate pose significant risks are likely to persist for most applications of continuous active learning for technology assisted review. For the former, we advise courts to require the labeling of all *Brady* material as positive in an execution of the tool. For the latter, we recommend taking measures to ensure that *Brady* material is discovered in early rounds of the continuous active learning framework. One approach to ensure this condition is met is to run an independent search for the *Brady* material $S = D$.

VII. LESSONS FOR COURTS AND PROSECUTORS

Our experimental findings point to three practical recommendations for prosecutors using TAR, as well as doctrinal implications.

First, prosecutors should run TAR separately for inculpatory and *Brady* evidence. From the simulations of asymmetric rate S3 and S4 (Figures 5.3 and 5.4), we see that there is a serious risk of not discovering *Brady* materials when they are much less frequent than inculpatory documents and are unlikely to be discovered in the initial round of manual review of the continuous active learning framework. However, individual search for $S = D$, i.e., running TAR separately for *Brady* materials, can alleviate this risk.

More generally we have seen that even in the best case, there are small risks from a combined search to simultaneously identify the inculpatory and *Brady* documents $S = P \cup D$ and no significant extra manual review of negative documents by parallel individual searches for $S = P$ and $S = D$. Moreover, in bad cases, the singular run of TAR has significant risk of either foregoing discovery of *Brady* documents or requiring manual review of a large number of negative documents. Independent runs of TAR avoid this risk.

These findings have implications for how courts and prosecutors should interpret the *Brady* doctrine. Practically speaking, prosecutors should be advised that the best way to ensure their *Brady* obligations are met is by independent runs of their TAR software for $S = P$ and $S = D$. Doctrinally, courts should incentivize independent runs by clarifying that *Brady* liability attaches strictly to all *Brady* materials in the prosecution's literal possession, regardless of whether any human on the prosecution team knows or has reason to know that the materials exist. If optimally-configured TAR could have surfaced *Brady* materials, the prosecution should be liable for their nondisclosure. Adopting this strict liability interpretation of *Brady* would further encourage the development of TAR workflows optimally-configured to expose *Brady* evidence.

Second, failure to label *Brady* materials as positive during TAR review substantially increases the risk that such materials will never be seen by a human. From the simulations of asymmetric labeling S1 and S2 (Figures 5.1 and 5.2) we

see that there is considerable risk of *Brady* documents remaining undiscovered when using effective automated methods, such as continuous active learning, which focus manual review on documents deemed most likely to be relevant to the prosecutor's search criteria.

Once again, these findings have implications for how courts and prosecutors should interpret the *Brady* doctrine. Regarding law enforcement over-seizure of data beyond the scope of a warrant, courts may adopt the "encounter rule." Recall that this would require prosecutors to disclose documents they encounter while separating in- and out-of-scope materials but prohibit them from running independent searches of out-of-scope data, even for *Brady* content. In this case, our findings suggest that courts should construe the Fourth Amendment at least to permit TAR labeling of *Brady* materials as positive during the separation process, regardless of whether the reviewed materials are themselves in- or out-of-scope for the warrant. In other words, the Fourth Amendment should permit TAR labeling of *Brady* materials as positive even if it prohibits an independent search for $S = D$. Doing so would strike a middle ground between full permission to troll through whatever out-of-scope data remains unseen after segregation and full abdication of *Brady* liability for such data.

Third, we have seen that multicalibration allows for multiple distinct groups of documents to be identified with a single execution of continuous active learning. This allows all responsive documents to be identified in the inseparable case of S6 (Figure 5.6) and the individually inseparable case of S9 (Figure 5.9). Thus if, in the *Brady* encounter rule context, it is forbidden to run the tools twice, the risks can be mitigated by using multicalibration.

The risk of asymmetric and incompatible classification are also effectively mitigated by modern machine learning tools such as support vector machines and deep neural networks (S10 [Figure 5.10] and S11 [Figure 5.11]). These approaches often allow data that might be inseparable by less flexible tools to be separated. Thus, if prosecutorial procurement guidelines and search protocols in warrants for digital evidence prioritize tools that are flexible, then these risks are minimal. It is beyond the scope of our study to understand whether modern methods are sufficient to render the data separable in any given criminal case.

A quick note is in order regarding the practical recommendations that we derive from our simulation experiments. Our experiments establish that, for certain configurations of evidence, TAR tools' technological design and user workflow can either risk suppressing or help to surface *Brady* materials. We think the configurations we study will appear in some real-world cases in which inculpatory evidence is more numerous and shares more characteristics (like names, date ranges, and keywords) than certain *Brady* material (like evidence of alibis, third-party guilt, or witness bias). But we make no claim about how frequently these configurations of evidence appear in real-world cases. As a result, we are unable to predict the real-world magnitude or probability of the risks and possibilities we identify. We hope and anticipate that future scholarship will study the effects we identify in real-world data.

Nonetheless, we believe that the mere existence of these risks and possibilities justifies adopting our practical recommendations now. These recommendations strictly dominate alternative approaches because they are each low cost to adopt. Running separate TAR searches for *Brady* materials; permitting and/or requiring coding of *Brady* material during a post-seizure separation of data that is in- and out-of-scope for a warrant; and favoring flexible classifiers in procurement should be sufficiently easy to implement that it makes sense to do so prophylactically while awaiting further studies of real-world case data and commercially-available TAR tools.

Further, while our simulation experiments inspired our doctrinal recommendations, we believe these recommendations offer more generalizable good and should thus be adopted regardless of how probable or not our experimental findings are to reflect real-world circumstances. Clarifying that *Brady* applies to all information literally possessed by the prosecution team, regardless of knowledge, would incentivize thorough *Brady* searches, whether incorporating TAR or not. To be sure, some *Brady* evidence will remain undiscovered even when search methods have been optimized to the full extent possible. And, practically speaking, if unknown, overlooked evidence is never discovered, the defense will bear the cost of that error unwittingly. Nonetheless, in those cases in which previously unknown, overlooked evidence is subsequently discovered, the prosecution should be on the hook for a *Brady* violation. *Brady* doctrine should assign the cost of such unavoidable errors to the prosecution because it is the least—and perhaps sole—cost avoider and because allocating the risk of error to the prosecution will not lead to wrongful convictions.

Similarly, clarifying whether and how *Brady* applies to seized data that falls out-of-scope of a warrant has importance beyond TAR. Our recommendation that the Fourth Amendment should at a minimum permit TAR labeling of *Brady* materials during the separation process can be generalized to broader principles: It may be excessive for *Brady* to permit unfettered access to data that law enforcement has already deemed out-of-scope for a Fourth Amendment warrant. But since law enforcement is already permitted to review some out-of-scope data for the purpose of separating it from evidence lawfully seized, they should at a minimum craft that review process in a manner most likely to surface *Brady* materials.

VIII. CONCLUSION

Zooming out, we observe a broader tension in constitutional criminal procedure: the more precise law enforcement investigations become—i.e., the better they are at collecting solely evidence that is the target of the government's investigation—the less *Brady* material the government may collect, identify, and disclose. There is thus an inherent tension between the Fourth Amendment (and privacy interests at large) and the usefulness of *Brady* (and other statutory discovery disclosure obligations) for surfacing exculpatory and impeachment evidence. This tension may have been dampened in the past by technological limits on investigative precision. As technologies like TAR tools increase investigative

precision, we anticipate that this tension will grow and ultimately require courts to reckon with two competing visions of the *Brady* rule: should *Brady* merely focus on keeping the prosecution's hands clean, i.e., free from deliberate or negligent concealment of evidence favoring the defense, in which case it might not matter if government searches laser-focus on inculpatory evidence? Or should *Brady* also serve as a vehicle for accurate fact-finding? If the latter, then perhaps the rule should require more expansive government investigations with a better chance of sweeping up exculpatory and impeachment evidence, not only during review of the prosecution's own files but also for searches of the world at large.[110]

Finally, this Article raises the possibility that *Brady*—and perhaps other areas of law—could and should shift focus from search *results* to search *methods*. In the case discussed here of prosecutorial use of TAR tools, perhaps prosecutors should disclose, and defendants should demand to know, how the tools were used, including details discussed here such as whether dual TAR runs were completed and whether *Brady* documents were labelled positive during tool review. The general principle is trans-substantive. For instance, defendants challenging judicial reliance on comparable cases for sentencing guidance may wish to contest not merely the similarity of prior cases that a judge used as precedents but also the methods by which the judge identified those cases and did not identify others.[111] New technologies may both increase the need for, and facilitate, ex-post auditing of search methods, thereby opening new avenues for adversarial challenges and legal oversight that previously focused on search results alone.

---

[110] An intermediate step would be to expand existing definitions of the prosecution's case files to other data sources that currently have ambiguous *Brady* status. For instance, for a fascinating discussion of the interstitial *Brady* status of data streams from Real Time Crime Centers and an argument to expand *Brady* disclosure duties to the "types of evidence" from these sources that prosecutors "routinely examine . . . for prosecution purposes," *see* Ferguson, *Real-Time Crime Centers*, *supra* note 36 (manuscript at *43, 45-46).

[111] *See, e.g.*, Woongjae Kim & Hai Jin Park, *The Impact of AI on Courts and Criminal Adjudication in South Korea*, *in* OXFORD HANDBOOK ON AI AND CRIMINAL JUSTICE (Brandon Garrett ed., forthcoming 2026) (on file with authors) (discussing how judges find comparable precedents to guide sentencing decisions in South Korea).

CAL Framework (Classify + Prioritize)

Input: Document universe U; seed set $L_1 \subseteq U$ with labels; batch size k; iterations T (or a stopping rule); training algorithm $Train$; priority rule $Prioritize$.

Output: Labeled set L_{T+1} (or L_{final} upon stopping).

1. Initialize L_1 to be the seed set.
2. For $t = 1,2, \ldots, T$ (or until stopping) do:
   (a) **Classify:** obtain a classifier $s_t \leftarrow Train(L_t)$
   (b) **Prioritize:** choose a batch $B_t \leftarrow Prioritize(s_t, U \setminus L_t, k)$ of k unlabeled documents.
   (c) **Review:** obtain labels for all documents in B_t and update $L_{t+1} \leftarrow L_t \cup B_t$.
3. Output L_{T+1} (or the labeled set at termination).

Figure A.1: CAL framework

APPENDIX A: ALGORITHMS

This section describes the algorithms used throughout the paper. In Section A.1, we formalize the Continuous Active Learning (CAL) framework. In Section A.2, we discuss how to instantiate CAL using the multicalibration approach. In Section A.3, we describe how to benchmark the performance of CAL with individual mode against CAL with combined mode.

1. Continuous Active Learning

Suppose we are given a universe of documents *U* and a subset of positive (target) documents $S \subset U$. Continuous Active Learning (CAL) is an iterative framework that selects documents for review in batches, updating its selection strategy based on the labels of previously-reviewed documents. CAL has two components: (i) a *classifier* that is trained on the labeled documents, and (ii) a *priority* algorithm that selects the next batch of unlabeled documents for review using the classifier.

Inputs. CAL is parameterized by:

- A batch size *k*.
- A seed set $L_1 \subseteq U$ of initially-reviewed documents (e.g., from keyword search or random sampling).

- A *classifier Train*(·) which, given the current labeled set L_t, returns a classifier s_t: $U \to R$.
- A *priority algorithm Prioritize*(·) which, given a classifier s_t and the unlabeled pool, selects the next batch of k documents.
- A stopping condition that determines when to stop the process.

Instantiations. The CAL framework above becomes a concrete algorithm once we specify (1) how we train a classifier-selection algorithm on the reviewed set L_t and (2) how we use this classifier to prioritize the next batch. In this paper, we use the following instantiations:

CAL + Linear. In each iteration, we train a linear classifier on the reviewed set L_t, which induces a separating hyperplane H in the original feature space. We prioritize each unlabeled document by its signed Euclidean distance to H (equivalently, its classifier score) and select the top k documents. CAL + Linear is used in the **combined** and **individual** modes.

CAL + SVM. In each iteration, we train a support vector machine (SVM) on L_t. SVMs can be viewed as first projecting documents into a higher-dimensional feature space (via a kernel or an explicit feature map) and then fitting a linear classifier in that projection space, often making data that was not separable in the original space separable. We prioritize each unlabeled document by the SVM decision value (equivalently, signed distance to the classifier in the projection space), and select the top k documents. CAL + SVM is used in the **support-vector machine** mode.

CAL + Multicalibration. In each iteration, we compute a multicalibrated partition of U using the reviewed set L_t. We *score* each subset in the partition by the empirical fraction of positively-labeled documents in the training set L_t. Documents are prioritized based on the score of the subset they are in, and we select unlabeled documents from the highest-scoring subset(s) until k documents are chosen. We describe multicalibration and the partition algorithm in detail in Section A.2. CAL + Multicalibration is used in the **multicalibrated** mode.

2. The Multicalibration Algorithm

We first provide some background on understanding multicalibration. Then we show how to compute a multicalibrated partition of the document space using an algorithm from previous work.[112]

a. Background on Multicalibration

Let D denote the distribution of documents. The e-discovery problem can be viewed as a Boolean classification problem where P and D have label 1 and all other documents have label 0. Define the *ground-truth* predictor $f^*(x) :=$

[112] *See* Gopalan et al., *supra* note 103, § 5 (using the multicalibration algorithm).

Multicalibrated Partition Algorithm

Input: All documents U, a training set $\{(x, y(x))\}_{x \in X}$ where $X \subset U$, and a target partition size m.

Output: A partition of all documents $\mathcal{S} = \{S_1, S_2, \dots\}$ not exceeding size $2m$.

1. Initialize the partition $\mathcal{S}_0 = [U]$ consists of a single subset with all documents.

2. Iteratively run the following steps until converge:

   (a) Split: Suppose the weak learner finds a hypothesis $c: S_i \to \{0,1\}$ for some subset $S_i \in \mathcal{S}_t$ in the current partition such that $|S_i \cap X| \geq (\alpha/m) \cdot |X|$ and $\left(\mathbf{E}_{D_i}[c(x)] - \mathbf{E}_{D_i}[c(x)|y = 1]\right) \cdot \mathbf{Pr}_{D_i}[y = 1] \geq \alpha$. Then, the subset S_i is split into two new subsets $A_0 = S_i \cap c^{-1}(0)$ and $A_1 = S_i \cap c^{-1}(1)$, where the hypothesis c labels as positive (A_0) and negative (A_1) respectively.

   (b) Merge: If the number of subsets in the current partition $\mathcal{S}_t$ exceeds $2m$, then the subsets are merged as follows. Divide $[0,1]$ into m equal sized intervals $\{I_j\}$, where $I_j = [(j-1)\ /\ m, j\ /\ m]$ for $j = 1,2, \dots, m-1$ and $I_m = [(m-1)\ /\ m, 1]$. For each $j = 1, \cdots, m$, all subsets with empirical positive portion $\sum_{x \in S_i \cap X} y(x)\ /\ |S_i \cap X|$ in the interval I_j are merged into a single subset.

3. Output the final partition $\mathcal{S}_T$.

Figure A.2: Multicalibrated Partition Algorithm

$E_{y \sim D}[y|x] \in [0,1]$ as the conditional expectation of the label for x. If we are able to efficiently learn an approximation to f^*, this can be used to solve the e-discovery problem, as we can simply choose all documents that have a high probability of having label 1. Gopalan et al. provide a method for learning an approximation to f^*, using *multicalibration*. Specifically, they construct a partition where subsets of the partition approximately consist of documents that have the same prediction from f^*.[113]

The notion of calibration has been well-studied in the prediction literature. A predictor f is *calibrated* if for every prediction v in its range we have

$$\mathbf{Pr}[y = 1 | f(x) = v] = v,$$

[113] *See id.*

in other words, the prediction $v = f(x)$ can be interpreted as a probability that x has label 1, and this probability is correct in expectation over documents x that have the same prediction v. By itself, calibration is a very weak property in terms of accuracy, so the notion of multicalibration was introduced that requires f to be calibrated over each subgroup in a large collection of subgroups.

We now rigorously define multicalibration. We will compute a partition $S = \{S_1, \ldots, S_m\}$ of the domain of documents based on the current pool of labeled documents. In order to learn a good approximation to the ground-truth predictor f^*, the goal is for these subsets S_i to group together documents that have similar predictions. Let D_i denote the distribution D conditioned on $x \in S_i$. Letting C be a family of Boolean functions, we say that S is *α-multicalibrated* for C if for every $i \in [m]$ and $c \in C$, the expectation of $c(x)$ does not change too much conditioning on the label:

$$\left|\mathbf{E}_{D_i}[c(x)] - \mathbf{E}_{D_i}[c(x)|y = 1]\right| \cdot \mathbf{Pr}_{D_i}[y = 1] \leq \alpha.$$

This condition implies that the decision boundary of c, which separates those documents x for which $c(x) = 0$ from those for which $c(x) = 1$, cannot separate any subset S_i of the partition into two subsets such that the average labels differ significantly between these two subsets. Hence if a partition is multicalibrated, then documents in the same subset of a partition must have similar probabilities of having positive labels.

b. Achieving Multicalibration Using Boosting

Gopalan et al. describe how to compute a multicalibrated partition using a machine learning technique called boosting of weak agnostic learners.[114] The pseudo-code of the multicalibrated partition algorithm is shown in Figure A.2. The main step of the algorithm is the *split* procedure: while there exists a hypothesis c that has nontrivial correlation with the labels of the documents in a subset S_i of the training set, then we split S_i into two new subsets according to the binary labels of c, one subset containing all documents x in S_i for which $c(x) = 0$, and one subset containing all documents x in S_i for which $c(x) = 1$. The reasoning behind this split operation is that the hypothesis c, having non-trivial correlations with the labels, is a signal for document labels and hence it makes sense to split the subset based on $c(x)$. If at any point the *split* procedure produces more than $2m$ subsets in the partition, the *merge* procedure is used to trim the number of subsets down to at most m by merging subsets whose empirical fraction of positively labeled documents lie within the same interval of width $1 / m$. A main result of Gopalan et al. is that the algorithm in Figure A.2 outputs a partition that is multicalibrated.

[114] *See generally id.* (using boosting of weak agnostic learners to compute a multicalibrated partition).

CAL-Multiplicative Weights Algorithm

Input: All documents U, A seed set $\{(x, y(x))\}_{x \in X}$ where $X \subset U$, number of iterations T, batch size k, and multiplicative weight factor $\eta \in (0,1)$.

Output: A set of labeled documents.

1. Initialize the training set L_1 to be the seed set. Initialize two weights $W_P \leftarrow 1$ and $W_D \leftarrow 1$ corresponding to P and D respectively.

2. For $t = 1,2, \ldots, T$ do:

   (a) Train two SVM classifiers $C_P^{(t)}$ and $C_D^{(t)}$ on L_t that separate P from $U \setminus P$ and D from $U \setminus D$ respectively. Note that L_t consists of documents with three possible labels, P, D, and $U \setminus (P \cup D)$, so $C_P^{(t)}$ is trained with only P documents being positively labeled and $C_D^{(t)}$ is trained with only D documents being positively labeled.

   (b) Do the following k times to select the next batch of documents to label:

      i. With probability $\frac{W_P}{W_P + W_D}$, label the document most likely to be *P* or *D* (depending on which classifier is being trained) according to $C_P^{(t)}$, which is the document closest to the decision boundary on the positive side. If the document is labeled negative, update $W_P \leftarrow W_P \cdot exp(-\eta)$.

      ii. Else (with probability $\frac{W_D}{W_P + W_D}$) label the document most likely to be *P* or *D* (depending on which classifier is being trained) according to $C_D^{(t)}$, and if it is labeled negative, update $W_D \leftarrow W_D \cdot exp(-\eta)$.

   Add the k newly-labeled documents to the training set L_t to get the new training set L_{t+1}.

3. Output the set of all labeled documents L_{T+1}.

Figure A.3: CAL-Multiplicative Weights

3. Continuous Active Learning with Two Parallel Classifiers

Recall that CAL with individual mode executes two parallel searches for the inculpatory (*P*) and exculpatory/impeachment (*D*) documents. In practice, if we have a clear stopping condition then these two searches can be run with the same stopping condition and separate batches. However, it is difficult to compare the

empirical performance of such a parallel approach with a sequential one in which documents are reviewed one at a time. In Algorithm A.3 we provide an implementation of CAL with individual mode that reviews documents sequentially. This allows us to create comparable plots of reviewed documents over time in Part V.

Algorithm A.3 combines two parallel classifiers for *P* and *D* respectively in a way that maximizes the rate of discovery of positive documents from either *P* or *D* across the two executions by employing the *multiplicative weights* method from algorithmic game theory. Two weights $W_P = 1$ and $W_D = 1$ are initialized whose ratio represents the relative likelihood that a document selected using the SVM classifier C_P is positively labeled compared with a document selected using C_D. At every step, to select the next document to label, with probability $\frac{W_P}{W_P+W_D}$ we select it using the classifier C_P and with probability $\frac{W_D}{W_P+W_D}$ we select it using the classifier C_D. If the newly-reviewed document is indeed positive, then we leave the weights unchanged. However, if for example we chose a document to review according to C_P and it was negative, then we multiply the weight W_P by a fixed multiplicative factor $exp(-\eta) < 1$. Since W_P is now smaller, the next document is more likely to be selected according to the classifier C_D instead of C_P. Likewise, if the newly-reviewed document was chosen according to C_D and negative, we decrease the weight W_D.

The intuition behind why the multiplicative weights method works to maximize the rate of discovery of positive documents is explained by the following story. If we had simply labeled a fixed fraction of documents for *P* and *D* in each batch, for example $\frac{k}{2}$ documents from *P* and $\frac{k}{2}$ from *D*, then if the relative sizes of *P* and *D* are unbalanced (asymmetric rate) then we will reach a point where nearly all of *D* will be reviewed before all of *P*, at which point the documents we select from the *D* batches are necessarily negative. Starting from this point, negative documents will be labeled at the same rate as positive documents from *P*, which leads to a substantial portion of negative documents reviewed before discovering all of *P*. By using multiplicative weights, the algorithm prioritizes discovering positive documents from *either D* or *P* before negative documents. To see why, note that for example once nearly all of *D* is discovered, the classifier C_D will suggest more and more negative documents to review. Each such negative document will decrease the weight W_D exponentially, resulting in the algorithm switching to almost exclusively review documents according to C_P as $\frac{W_P}{W_P+W_D} \to 1$.

APPENDIX B: SIMULATED DATASETS

We evaluate TAR algorithms on simulated datasets. Each dataset consists of a universe of documents *U*, which includes three types of documents, inculpatory documents *P*, exculpatory/impeachment documents *D* (also referred to as *Brady* Material), and remaining irrelevant documents $U \setminus (P \cup D)$.

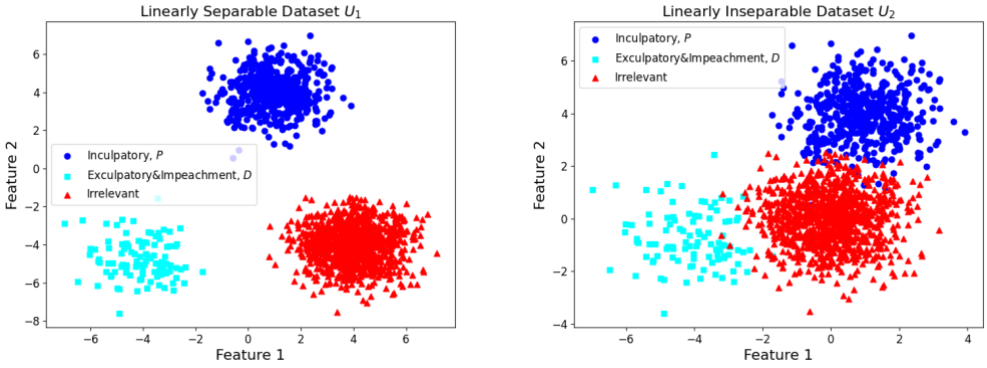

Figure B.1: Simulated datasets U_1 and U_2. The left plot shows the dataset U_1, which is linearly separable. The right plot shows the dataset U_2, which is linearly inseparable, making classification more challenging. These datasets illustrate different levels of difficulty in the discovery process.

Linearly Separable Data. The first dataset U_1 simulates a scenario with an asymmetric rate for *Brady* Material. The inculpatory and *Brady* documents $P \cup D$ and the irrelevant documents $U \setminus (P \cup D)$ can be perfectly separated by a linear classifier. However, the rate of inculpatory documents $|P| / |U|$ is significantly higher than the rate of *Brady* documents $|D| / |U|$. Specifically, the dataset U is generated as follows:

- Inculpatory documents P: 500 samples drawn from a two-dimensional Gaussian distribution with mean (1,4) and variance 1.
- *Brady* documents D: 100 samples drawn from a two-dimensional Gaussian distribution with mean (–4, –5) and variance 1.
- Irrelevant documents $U \setminus (P \cup D)$: 1000 are samples drawn from a two-dimensional Gaussian distribution with mean (4, –4) and variance 1.

Linearly Inseparable Data. The second dataset U_2 simulates the scenarios that no linear classifier can perfectly classify the inculpatory and *Brady* documents $P \cup D$ from the irrelevant documents $U \setminus (P \cup D)$. The rate of inculpatory documents $|P| / |U|$ is significantly higher than the rate of *Brady* documents $|D| / |U|$. Specifically, the dataset U is generated as follows:

- Inculpatory documents P: 500 samples drawn from a two-dimensional Gaussian distribution with mean (1,4) and variance 1.
- *Brady* documents D: 100 samples drawn from a two-dimensional Gaussian distribution with mean (–4, –1) and variance 1.
- Irrelevant documents $U \setminus (P \cup D)$: 1000 samples drawn from a two-dimensional Gaussian distribution with mean (0, 0) and variance 1.

Linearly Separable with Rare Brady Data. The third dataset U_3 simulates a scenario with scarce *Brady* Material. The inculpatory and *Brady* documents $P \cup D$ and the irrelevant documents $U \setminus (P \cup D)$ can be perfectly separated by a linear classifier. However, the rate of *Brady* documents $|D| / |U|$ is so small that the initial

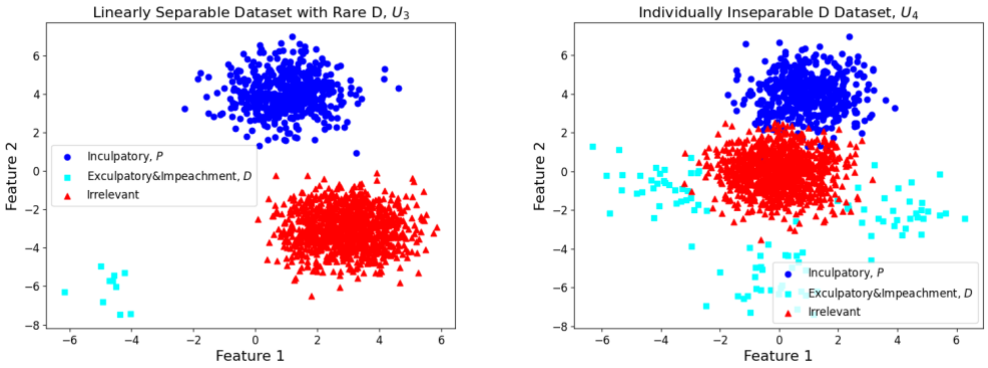

Figure B.2: Simulated datasets U_3 and U_4. The left plot shows the dataset U_3. In U_3, exculpatory/impeachment documents D are extremely rare, so the initial iteration of CAL is unlikely to discover them, even though $P \cup D$ is linearly separable from $U \setminus (P \cup D)$. The right plot shows the dataset U_4. In U_4, exculpatory/impeachment documents D are diffuse (multi-modal), so no single linear separator perfectly separates D from $U \setminus (P \cup D)$.

step of CAL is unlikely to discover D. Specifically, the dataset U is generated as follows:

- Inculpatory documents P: 500 samples drawn from a two-dimensional Gaussian distribution with mean (1, 4) and variance 1.
- *Brady* documents D: 10 samples drawn from a two-dimensional Gaussian distribution with mean (–5, –6) and variance 0.5.
- Irrelevant documents $U \setminus (P \cup D)$: 1090 are samples drawn from a two-dimensional Gaussian distribution with mean (3, –3) and variance 1.

Individually Inseparable Dataset. The fourth dataset U_4 simulates the scenarios in which the *Brady* set D is spread across several modes, making D and $U \setminus (P \cup D)$ linearly inseparable. Specifically, the dataset U is generated as follows:

- Inculpatory documents P: 500 samples drawn from a two-dimensional Gaussian distribution with mean (1, 4) and variance 1.
- *Brady* documents D: 100 samples drawn from a three-component mixture of Gaussians in two dimensions, with equal mixing weights. The three components are centered at (–4, –1), (0, –6), (4, –2) with the same variance 1.
- Irrelevant documents $U \setminus (P \cup D)$: 1000 samples drawn from a two-dimensional Gaussian distribution with mean (0, 0) and variance 1.

We implement the original CAL algorithm and our CAL Multicalibration algorithm on these simulated datasets and evaluate their performance in Part V.

APPENDIX C: THEORETICAL ANALYSIS OF MULTICALIBRATION

In this section, we analyze the partition produced by multicalibration under the assumption that the positives comprise $k \geq 2$ disjoint clusters $P_1, \ldots, P_k$, each linearly separable from the negatives $N = U \setminus (\bigcup_{j=1}^{k} P_j)$. For any subset $X \subseteq U$, we show that the final partition given by training set X has small *total impurity* on

X, which means it nearly separates positives $X \cap (\bigcup_{j=1}^{k} P_j)$ from negatives $X \cap N$. As a consequence, labeling each cell by its within-cell majority yields a small 0-1 loss on X. Thus, when X is large enough, we expect this partition closely aligns with the population separation and separates positives $\bigcup_{j=1}^{k} P_j$ from negatives N. In this case, the CAL algorithm will discover almost all positives before reviewing too many negatives. We now discuss some scenarios where this theorem applies. When P and D are each linearly separable from $N = U \setminus (P \cup D)$, the assumption holds with $k = 2$. When P is linearly separable from N and $D = \bigcup_{j=1}^{k'} D_j$ where each D_j is linearly separable from N, the assumption holds with $k = k' + 1$.

Consider any subset $X \subseteq U$ and any partition S. For each part $S \in S$, we define its impurity on X as

$$Imp(S; X) = min\{\mathbf{Pr}_{x\sim X}[f(x) = 1 | x \in S], \mathbf{Pr}_{x\sim X}[f(x) = 0 | x \in S]\}.$$

Define the total impurity of S with respect to set X by

$$Imp(S; X) := \sum_{S\in S} \mathbf{Pr}_{x\sim X}[x \in S] \cdot Imp(S; X).$$

Intuitively, the impurity of S is equal to the 0-1 loss of labeling each part S by its majority.

Theorem C.1 (Near-pure partition via multicalibration). *Consider a set U containing k positive clusters $P_1, \cdots, P_k$, and a negative cluster N, satisfying each P_j is linearly separable from N. Let $X \subseteq U$ be any subset. Let S be the partition of the multicalibration algorithm on X with linear weak learners and target parameter $\alpha = 4\varepsilon(1-\varepsilon) \,/\, k$. Then the total impurity of S on X is at most*

$$Imp(S; X) \leq \varepsilon.$$

Consequently, labeling each part $S \in S$ by its majority yields a classifier with 0–1 loss at most ε on X. Equivalently, S almost separates $X \cap (\bigcup_{j=1}^{k} P_j)$ from $X \cap N$.

Proof. We first show that if any part has a high impurity, then there exists a linear classifier h with a high covariance with the true label f, $\mathrm{Cov}(h, f) \geq \alpha$.

Consider any part that contains a set $X' \subseteq X$. Let $\hat{P}_j = P_j \cap X'$ for $j = 1, \ldots, k$ and $\hat{N} = X' \cap N$. Let $p_j = \mathbf{Pr}\left[x \in \hat{P}_j | x \in X'\right]$ for $j = 1, \ldots, k$, and $q = \mathbf{Pr}\left[x \in \hat{N} | x \in X'\right]$, so $\sum_{j=1}^{k} p_j + q = 1$. We have the true label $f(x) = 1$ on any P_j and $f(x) = -1$ on N.

By the separability assumption, for each j there exists a linear classifier h_j with $h_j = 1$ on P_j and $h_j = -1$ on N. Let $p' = 1 - q - p_j$ and $z = \mathbf{E}\left[h_j \mid x \in X' \setminus \left(P_j \cup N\right)\right] \in [-1, 1]$. Then, we have for each j:

$$\begin{aligned}
&\mathbf{E}_{x\sim X'}\left[h_j(x) f(x)\right] \\
&\quad = \mathbf{Pr}[\hat{P}_j] + \mathbf{Pr}[\hat{N}] + \\
&\quad\quad \mathbf{Pr}[X' \setminus \left(P_j \cup N\right)]\, \mathbf{E}_{x\sim X'}\left[h_j(x) \middle| x \in X' \setminus \left(P_j \cup N\right)\right]
\end{aligned}$$

$$= \mathrm{p_j} + \mathrm{q} + \mathrm{p'z},$$

$$\mathbf{E}_{x\sim X'}\big[h_j(x)\big]$$

$$= \mathbf{Pr}[\hat{P}_j] \;-\; \mathbf{Pr}[\hat{N}] \;+$$

$$\mathbf{Pr}[X' \setminus (P_j \cup N)]\; \mathbf{E}_{x\sim X'}\big[h_j(x) \,\big|\, x \in X' \setminus \big(P_j \cup N\big)\big]$$

$$= p_j - q + p'z,$$

$$\mathbf{E}_{x\sim X'}[f(x)]$$

$$= \mathbf{Pr}[\hat{P}_j] \;-\; \mathbf{Pr}[\hat{N}] \;+\; \mathbf{Pr}[X' \setminus \big(P_j \cup N\big)]$$

$$= \mathrm{p_j} + \mathrm{p'} - \mathrm{q} = 1 - 2\mathrm{q}.$$

Hence, we have the covariance of h_j and f is

$$\mathrm{Cov}_{x\sim X'}\,\Big(h_j(x), f(x)\Big)$$

$$= \mathbf{E}_{x\sim X'}\big[h_j(x)f(x)\big] - \mathbf{E}_{x\sim X'}\big[h_j(x)\big]\,\mathbf{E}_{x\sim X'}[f(x)]$$

$$= 2q\big(p_j - q + p'z + 1\big).$$

Since $z \in [-1,1]$, we have

$$\mathrm{Cov}_{x\sim X'}\,\Big(h_j(x), f(x)\Big) \;= 2q\big(p_j - q + p'z + 1\big)$$

$$\geq 2q\big(p_j - q - p' + 1\big) = 4p_j q.$$

Therefore, there exists a linear classifier h such that,

$$\mathrm{Cov}_{x\sim X'}\big(h(x), f(x)\big) \geq 4q \max_{j\in[k]} p_j \geq \frac{4}{k} q(1-q),$$

since $\max_{j\in[k]} p_j \geq (1-q)\,/\,k$.

The impurity of X' equals $min\{q, 1-q\}$. If the impurity is larger than ε, then there exists a linear classifier h with covariance greater than $4\varepsilon(1-\varepsilon)\,/\,k = \alpha$. Hence, any cell with impurity larger than ε admits a linear classifier with covariance at least α, so the algorithm would keep splitting it. At the end of the algorithm, the total impurity of S is at most ε.