Harvard Journal of Law & Technology
Volume 38, Digest Spring 2025

AI AUDITING: FIRST STEPS TOWARDS THE EFFECTIVE REGULATION OF ARTIFICIAL INTELLIGENCE SYSTEMS

Edwin A. Farley & Christian R. Lansang***

ABSTRACT

The unprecedented ways in which artificial intelligence (AI) can affect our lives motivate a need for greater regulation of AI. In the face of technological change and legal challenges, auditing of AI has the potential to deliver accountability for the impacts of AI systems. Drawing insights from financial auditing and the current AI auditing landscape, this article identifies reforms necessary to create effective forms of AI auditing. This Note argues that there is a need for government-mandated AI audits conducted by professional auditors following established standards subject to government oversight.

The emergence of an industry of AI auditors and government oversight could deliver accountability for AI without disincentivizing innovation. AI auditing standards should reflect a comprehensive approach targeting three components of AI development: (1) data, (2) model, and (3) deployment. Far from being an increased cost of business, AI auditing and oversight can become a means to accelerate advancement of the technology itself. Over time, as the public gains appreciation of the value of an AI audit alongside expanding audit mandates, a virtuous cycle can emerge, reining in the dangers of AI while advancing the technology in a way that is consistent with its positive potential.

---

* J.D., Harvard Law School, 2024; Sc.B., Brown University, 2019.

** J.D., Harvard Law School, 2022; A.B., Brown University, 2019. Christian is the Head of Growth and Product at Better Life Health.

The authors thank Professor Martha Minow for her help from the earliest stages of this project and for her feedback and encouragement throughout. The authors further thank Professor Daniel Rauch, Alan C. Raul, and Arthur Farley. The authors also thank the editors and staff of the *Harvard Journal of Law & Technology* for their commitment to this piece. This paper reflects the authors' personal views only and not those of any company, government entity, institution, or other person.

TABLE OF CONTENTS

I. INTRODUCTION

Applications of artificial intelligence (AI) systems have ballooned,[1] and AI can now be found in sensitive contexts such as hiring, credit scoring, and evaluations of loan applications.[2] The

1. An "AI system" or "automated system" is a "system, software, or process that uses computation as whole or part of a system to determine outcomes, make or aid decisions, inform policy implementation, collect data or observations, or otherwise interact with individuals and/or communities." THE WHITE HOUSE, BLUEPRINT FOR AN AI BILL OF RIGHTS 10 (2022), https://ai.org.tr/wp-content/uploads/2022/10/9-US-White-House-BLUEPRINT-FOR-AN-AI-BILL-OF-RIGHTS-.pdf [https://perma.cc/V4SP-44CL] [hereinafter BLUEPRINT]. These systems are to be distinguished from "passive computing infrastructure," which "[do] not influence or determine the outcome of decision, make or aid in decisions, inform policy implementation, or collect data or observations." *Id.* Thus, automated systems have the potential to "meaningfully impact individuals' or communities' rights, opportunities, or access." *Id.*

2. *See, e.g.*, Rebecca Heilweil, *Artificial Intelligence Will Help Determine if You Get Your Next Job*, VOX (Dec. 12, 2019, 8:00 AM EST), https://www.vox.com/recode/2019/12/12/20993665/artificial-intelligence-ai-job-screen [https://perma.cc/8ZB5-VM2W]; Emmanuel Martinez & Lauren Kirchner, *The Secret Bias Hidden in Mortgage-Approval Algorithms*, MARKUP (Aug. 25, 2021, 6:50 ET), https://themarkup.org/denied/2021/08/25/the-secret-bias-hidden-in-mortgage-approval-algorithms, [https://perma.cc

accompanying dangers of employing AI in sensitive domains have been documented as well, including the reification of existing biases and disproportionate negative effects on already vulnerable populations.[3] Although these harms predate AI technology, AI can "amplif[y] and exacerbat[e]" them.[4] While AI presents novel challenges, it is imperative not to understate its *positive* transformative potential.[5] AI systems have improved skin cancer diagnoses,[6] brought healthcare to underserved communities,[7] and made hiring fairer in certain cases.[8] Despite emerging public misgivings about AI in general,[9] determining if an AI application is *worse* than the existing, human alternative requires an inquiry into the AI system's input data and the structure and content of its model, including how the model is trained, and how the system is deployed.[10] These are the elements of the AI audits we envision.

Imagine that a private regional insurance agency begins using an AI system to determine premiums. In its marketing materials, the

---

/4Q6V-CG7Z]; *Comments Of the Electronic Privacy Information Center to the Consumer Financial Protection Bureau*, Dkt, No. CFPB-2020-0026 (Oct. 2, 2020).

3. *See, e.g.*, Aaron Sankin, Dhruv Mehrotra, Surya Mattu & Annie Gilbertson, *Crime Prediction Software Promised to Be Free of Biases. New Data Shows It Perpetuates Them*, MARKUP (Dec. 2, 2021), https://themarkup.org/prediction-bias/2021/12/02/crime-prediction-software-promised-to-be-free-of-biases-new-data-shows-it-perpetuates-them [https://perma.cc/2LCZ-ZKBF]. "Sensitive domains" are settings in which actions can cause "material harms, including significant adverse effects on human rights such as autonomy and dignity, as well as civil liberties and civil rights." BLUEPRINT, *supra* note 1, at 11. Examples include "health, family planning and care, employment, education, criminal justice, and personal finance." *Id.* A domain may be considered "sensitive" whether or not existing laws govern it. What is considered "sensitive" may change over time and in different contexts. *Id.*

4. Remarks of Chair Lina M. Khan Regarding Combatting Online Harms Through Innovation Report, FED. TRADE COMM'N, COMM'N FILE NO. P214501 (June 16, 2021).

5. *See, e.g.*, Kai-Fu Lee, *AI and the Human Future: Net Positive*, MEDIUM (May 20, 2022), https://kaifulee.medium.com/ai-and-the-human-future-net-positive-ae3a500c1846 [https://perma.cc/RCC7-QNLQ]; Kai-Fu Lee, *How AI Will Completely Change the Way We Live in the Next 20 Years*, MEDIUM (Mar. 16, 2022), https://kaifulee.medium.com/how-ai-will-completely-change-the-way-we-live-in-the-next-20-years-e27a855b1bd0 [https://perma.cc/BD4Y-ZUS7].

6. Peter Tschandl, *Human-Computer Collaboration for Skin Cancer Recognition*, 26 NATURE MED. 1229, 1229–34 (2020).

7. Emma Pierson, David M. Cutler, Jure Leskovec, Sendhil Mullainathan & Ziad Obermeyer, *An Algorithmic Approach to Reducing Unexplained Pain Disparities in Underserved Populations*, 27 NATURE MED. 136, 138 (2021).

8. Rebecca Greenfield & Riley Griffin, *Artificial Intelligence Is Coming for Hiring, and It Might Not Be That Bad*, BLOOMBERG (Aug. 8, 2018, 5:00 AM EDT), https://www.bloomberg.com/news/articles/2018-08-08/artificial-intelligence-is-coming-for-hiring-and-it-might-not-be-that-bad [https://perma.cc/HD2M-ANGB].

9. In a recent study by Pew Research, 37% of poll respondents said they were "more concerned than excited" about the increased use of AI in everyday life; 45% said they were "equally excited and concerned." Lee Rainie, Cary Funk, Monica Anderson & Alec Tyson, *AI and Human Enhancement: Americans' Openness Is Tempered by a Range of Concerns*, PEW RSCH. (Mar. 17, 2022), https://www.pewresearch.org/internet/2022/03/17/ai-and-human-enhancement-americans-openness-is-tempered-by-a-range-of-concerns/ [https://perma.cc/U2XY-Q94W].

10. Ash Carter, *The Moral Dimension of AI-Assisted Decision-Making: Some Practical Perspectives from the Front Lines*, 151 DAEDALUS 299, 301–02 (2022).

insurance agency claims to offer its customers "the fairest rates possible" by using "proprietary AI technology." Shortly after implementing its AI system, the insurance agency receives several complaints from customers about high rates in various local markets. During an AI audit, an auditor creates a test suite of applications for insurance policies, informed by customer complaints as well as its own study, and finds that, indeed, customers living in certain parts of town are consistently offered higher rates. Having been granted access to the agency's systems and development practices, the auditor can now identify the source of the issue: the AI system relies on older, "stale" data from a time when the disadvantaged areas had disproportionate amounts of young drivers and consequently more reported accidents.[11] The auditor then creates a plan to remedy the issues, introducing controls into the audited company's systems and development practices.[12] In subsequent audits, compliance is monitored, and the impacted areas begin to benefit from lower, *fairer* insurance premiums. In this example, a mandated audit would touch the "invisible" aspects of AI systems that are not accessible to consumers, as well as portions of the AI systems that currently escape the purview of existing regulators.[13]

AI audits can also operate on generative AI systems. While there is no clear demarcation between generative and non-generative AI models, generally speaking, systems that are trained to create new data — and not just predictions on existing datasets — are often characterized as generative AI.[14] AI models, such as the model in the example above, utilize machine learning to make predictions about a specific dataset and are non-generative (e.g., in the insurance example above, based on a population with a given set of characteristics, an insurance premium is set to optimize profitability for the company). OpenAI's GPT-4, for example, is a popular generative AI model: upon receiving input in the form of text from its user, the model provides

---

11. "Stale" data is data that has not been updated at the frequency interval required for its productive use. *See* Michael Segner, *Stale Data Explained: Why It Kills Data-Driven Organizations*, MONTE CARLO (Mar. 28, 2023), https://www.montecarlodata.com/blog-stale-data/ [https://perma.cc/MJ58-L4DS].

12. *See, e.g.*, Brett Frischmann & Paul Ohm, *Governance Seams*, 37 HARV. J.L. & TECH. 1117, 1128–32 (2023) (defining characteristics of "governance seams" for managing information).

13. *See* Jennifer Valentino-Devries, Natasha Singer, Michael H. Keller & Aaron Krolik, *Your Apps Know Where You Were Last Night, and They're Not Keeping It Secret*, N.Y. TIMES (Dec. 10, 2018), https://www.nytimes.com/interactive/2018/12/10/business/location-data-privacy-apps.html [https://perma.cc/C4R6-W4GZ]. Facebook even engaged in such practices itself, collecting information through the embedded "Like" button it made available to users. *See* Dina Srinivasan, *The Antitrust Case Against Facebook: A Monopolist's Journey Towards Pervasive Surveillance in Spite of Consumers' Preference for Privacy*, 16 BERKELEY BUS. L.J. 39, 65–67 (2019). For instance, the Consumer Protection Division of the FTC relies on tips from ordinary consumers at https://reportfraud.ftc.gov [https://perma.cc/SUQ9-RCML].

14. Adam Zewe, *Explained: Generative AI*, MIT NEWS (Nov. 9, 2023), https://news.mit.edu/2023/explained-generative-ai-1109 [https://perma.cc/EY2R-SFMG].

textual outputs in response.[15] If a user were to provide ChatGPT — a chatbot application built on GPT-4 — with ingredients in her fridge and instruct ChatGPT that she has an interest in eating only healthy meals, the application could provide various recipes with accompanying nutritional facts.

Despite generative AI's advancements from a technical and popularity standpoint, its nascency as a technology brings forth novel problems. For example, consider an online news editorial board that has deployed a new generative AI tool from a third-party provider that lauds its AI system's ability to "increase interaction and engagement for stories through the novel application of generative AI." A particular novelty of this generative AI tool is its ability to autonomously implement interactive features for its readers. For example, in articles covering the Met Gala, hovering over an attendee's outfit brings forth a pop-up for the reader: "Would you like to see other outfits attendees wore that were designed by Thom Browne?" In the first few months after implementing the tool, the website's readership increases drastically. Features like polls at the end of stories are evaluated less intensively by editors compared to when the AI tool was first implemented. Later, however, the editors notice that one of its stories covering a woman's tragic death received hundreds of thousands of views. Upon closer inspection, the editors learn the story's virality is not driven by its content but rather because the story includes a poll in the middle of the article asking, "How do you think Meredith died? (A) Suicide (B) Reckless driving (C) Complete accident (D) Who cares." The poll also includes accompanying photos of the woman and her children and footage of the woman's car accident. After public demands to understand why this poll was included, the editorial board learns from the AI provider that controversial polls — though none of this exact kind — were common features the AI system would implement to boost engagement. The AI-tool provider thought it had rectified this feature by banning certain words that polls could include, such as slurs and other pejoratives. However, the AI-tool provider never re-evaluated the system after implementing this ban, since polls were particularly powerful engagement-creating devices. Moreover, the AI provider never disclosed these issues with its polls in its dealings with the editorial board.

This hypothetical is not a warning for an AI tool multiple years away — we have *already* seen variants of this exact harm. Recently, in The Guardian, a story describing a woman's death that was syndicated on Microsoft Start,[16] a Microsoft-generated poll appeared alongside the

15. *See* CHATGPT, https://chatgpt.com/ [https://perma.cc/QC4Z-8A8R].

16. Tamsin Rose & Nino Bucci, *Woman Found Dead at St Andrew's School in Sydney Identified as Water Polo Coach Lilie James*, GUARDIAN (Oct. 26, 2023, 2:04 AM EDT), https://www.theguardian.com/australia-news/2023/oct/26/womans-body-found-at-central-sydney-school-as-police-investigate-suspicious-death [https://perma.cc/8HQG-NJ8M].

story asking readers what they thought the reason behind the woman's death was.[17] After learning of the poll, The Guardian alleged that Microsoft had damaged its journalistic reputation and stressed the importance of transparency and consumer-safeguards for AI tools in the journalism context.[18] Shortly afterwards, Microsoft stated that, "We have deactivated Microsoft-generated polls for all news articles and we are investigating the cause of the inappropriate content. A poll should not have appeared alongside an article of this nature, and we are taking steps to help prevent this kind of error from reoccurring in the future."[19]

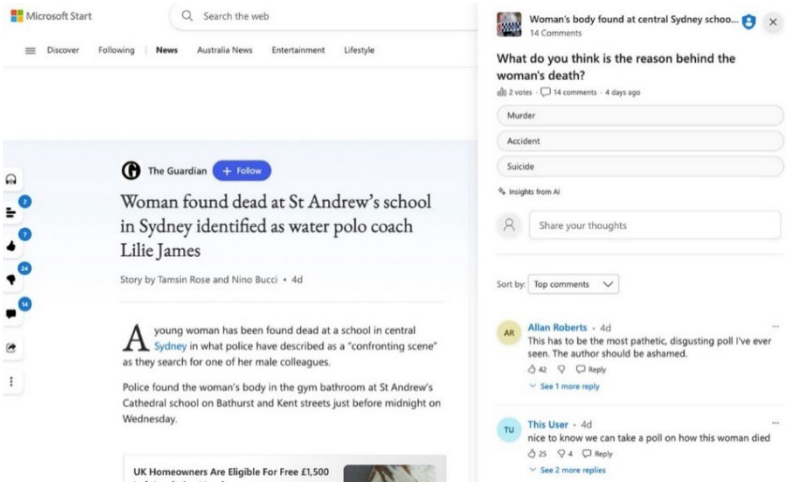

Figure 1: AI-generated poll accompanying news article

Any attempt to regulate AI will be technologically and legally challenging, but auditing may provide a flexible and adaptable approach with the potential to deliver accountability for AI systems and their impacts. Independent auditing activities have already revealed real AI harms, demonstrating that the auditing process is worthwhile.[20]

17. The poll has since been removed, and the screenshot is taken from The Verge reporting on the matter. Wes Davis, *Microsoft AI Inserted a Distasteful Poll into a News Report About a Woman's Death*, VERGE (Oct. 31, 2023, 12:24 PM EDT), https://www.theverge.com/2023/10/31/23940298/ai-generated-poll-guardian-microsoft-start-news-aggregation [https://perma.cc/T4W9-675D].

18. *Id.*; Anna Bateson, *Email to: Bradsmi@microsoft.com*, https://docs.google.com/viewerng/viewer?url=https://s3.documentcloud.org/documents/24101210/letter_to_brad_smith_from_anna_bateson_31_10_23.pdf [https://perma.cc/H6MW-V9H9] (message made available to The Verge).

19. Davis, *supra* note 17.

20. *See, e.g.*, Lisa Macpherson, *Observe and Report: Facebook Versus NYU Ad Observatory Proves the Need for Policy Interventions*, PUB. KNOWLEDGE (Aug. 11, 2021), https://publicknowledge.org/observe-and-report-facebook-versus-nyu-ad-observatory-proves-the-need-for-policy-interventions/ [https://perma.cc/F2QM-CWAC].

However, at present, auditing efforts are vulnerable to legal barriers and technical interference, and are further weakened by a lack of universally accepted auditing standards and mandates.[21] Furthermore, amid the current tension between proposed regulatory schemes versus pure reliance on market forces, a different approach could be a catalyzing first step.[22] This Note proposes a solution to regulate AI and improve AI accountability by mandating AI audits conducted by AI auditors following set standards and subject to government oversight.

This Note presents an analysis that combines concepts from the fields of law and computer science to build on scholarship at the intersection of law and AI. By identifying the different types of auditing relationships, this Note investigates the potential for the growth of an industry of external AI auditors.[23] It takes the further step of advocating for auditors to play a crucial role in a regulatory scheme for AI,[24] and includes government oversight of auditors.[25] Moreover, it takes a

---

21. Annie Lee, *Algorithmic Auditing and Competition Under the CFAA: The Revocation Paradigm of Interpreting Access and Authorization*, 33 BERKELEY TECH. L.J. 1307, 1309–11 (2018). A third-party auditor who reports scraping client data to conduct audits, explained: "[Our] biggest barrier is probably legal. The Computer Fraud and Abuse Act criminalizes terms of service violations, which often occur in automated data collection at scale of publicly available data." Sasha Costanza-Chock, Emma Harvey, Inioluwa Deborah Raji, Martha Czernuszenko & Joy Buolamwini, *Who Audits the Auditors? Recommendations From a Field Scan of the Algorithmic Auditing Ecosystem*, ACM FACCT CONF. (2022). This sentiment persisted after decisions that narrowed the CFAA. *See* Sandvig v. Barr, 451 F. Supp. 3d 73, 76 (D.D.C. 2020) (refusing to criminalize violations of terms of service under CFAA when data is collected for research purposes).

22. *See* Clark D. Asay, *Artificial Stupidity*, 61 WM. & MARY L. REV. 1187, 1190–93 (2020) (describing motivations for existing scholarship on AI in the law but noting that a crucial question is missing: "What, for instance, is the best innovation policy for spurring radical AI innovation?"). Regulation may also encourage innovation. *See* Alan C. Raul, *Who's Balancing Privacy Against Public Health and Everything Else?*, THE HILL (June 6, 2020, 11:00 AM ET), https://thehill.com/opinion/cybersecurity/502517-whos-balancing-privacy-against-public-health-and-everything-else/ [https://perma.cc/8JDG-26C2] (discussing the imperative both for data privacy considerations in innovation and innovation in methods for creating data privacy).

23. *Cf.* Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 661 (2017) (treats auditing as merely examining a black box); Andrew D. Selbst, *An Institutional View of Algorithmic Impact Assessments*, 35 HARV. J.L. & TECH. 117 (2021) (declining to take on auditing as an approach).

24. *See generally* Bodo, B., Helberger, N., Irion, K., Zuiderveen Borgesius, F., Moller, J., van de Velde, B. et al., *Tackling the Algorithmic Control Crisis-the Technical, Legal, and Ethical Challenges of Research into Algorithmic Agents*, 19 YALE J.L. & TECH. 143 (2017) (touching on challenges to the auditing of AI, especially for what this Note terms "third-party" auditors, but does not consider regulation of auditors or regulation of AI that includes audits).

25. *See* W. Nicholson Price II, *Black-Box Medicine*, 28 HARV. J.L. & TECH. 419, 461–62 (2015) (discussing potential regulatory solutions, including potential for third-party "validation" that could be nimbler than the FDA's existing regulatory approach); Danaë Metaxa, Joon Sung Park, Ronald E. Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock et al., *Auditing Algorithms*, 14 FOUNDS. & TRENDS HUM.-COMPUT. INTERACTION 272, 325–27 (2021) (identifying the importance of audits that are independent and impartial, but without investigating the process as a part of a regulatory program and the requirements that might accompany it); s*ee also* Alfred Ng, *Can Auditing Eliminate Bias from Algorithms?*, MARKUP

system-wide view that encompasses the application-specific insights of existing work that has focused on AI as applied in certain domains.[26]

This Note first evaluates current forms of AI auditing and proposes a structure for AI audits. It then articulates how the oversight of auditors can initiate a virtuous cycle of AI development, before examining legal solutions for the advancement and oversight of AI auditing. It concludes by illustrating an example of how auditing could operate with respect to AI exam proctoring software.

II. AI AUDITORS AND AUDITS

Analyzing AI auditing as a way to regulate AI requires careful consideration of *who* the auditors are and *what* they audit.

A. Categories of AI Auditors

Drawing from the industry of financial auditing and the current AI auditing landscape, three relevant categories of auditors emerge: internal auditors, third party auditors, and external auditors.[27] Although audits by completely independent researchers have yielded results, the fate of their operations often remains uncertain. Meanwhile, the growth of an industry of *external* AI auditors could be an effective approach when auditors' practices are regulated and subject to potential legal liability.

(Feb. 23, 2021, 8:00 AM ET), https://themarkup.org/the-breakdown/2021/02/23/can-auditing-eliminate-bias-from-algorithms [https://perma.cc/UA6R-J8RJ] (identifying shortcomings of current auditing practices, but stops short of describing qualities of an AI auditor and how government oversight of auditors could ameliorate the shortcomings it identifies). *See generally* David S. Rubenstein, *Acquiring Ethical AI*, 73 FLA. L. REV. 747 (2021) (introducing role for government in encouraging development of ethical AI, but through procurement processes).

26. *See* Price II, *supra* note 25, at 461 (focusing on the rise of "personalized medicine" in the pharmaceutical and medical device industries and the risks of employing opaque algorithms in their development); Ifeoma Ajunwa, *An Auditing Imperative for Automated Hiring*, 34 HARV. J.L. & TECH. 621, 670 (2021) (discussing need for audits of automated hiring systems); Cary Coglianese & Erik Lampmann, *Contracting for Algorithmic Accountability*, 6 A.L.R. ACCORD 175, 192–94 (2021) (touching on audits as a component of contractual requirements, thus recognizing their potential, but not beyond use by government contractors). *See generally* Susan S. Fortney, *Online Legal Document Providers and the Public Interest: Using a Certification Approach to Balance Access to Justice and Public Protection*, 72 OKLA. L. REV. 91 (2019) (examining how third-party certification of automated legal services offerings can broaden access to legal services and improve consumer protection).

27. *See* Costanza-Chock et al., *supra* note 21; Press Release, SEC, *SEC Implements Internal Control Provisions of Sarbanes-Oxley Act; Adopts Investment Company R&D Safe Harbor* (May 27, 2003), https://www.sec.gov/news/press/2003-66.htm [https://perma.cc/S6A3-UGS6]; *Basics of Inspections*, PCAOB, https://pcaobus.org/oversight/inspections/basics-of-inspections [https://perma.cc/N83T-MXMK]; *see also* Ellen P. Goodman & Julia Trehu, *Algorithmic Auditing: Chasing AI Accountability*, 39 SANTA CLARA HIGH TECH. L.J. 289, 294–96 (2023).

1. Internal Auditors

Internal auditors are employed by and are a part of the companies whose practices they assess. Several large technology companies presently have sizeable internal audit teams.[28] These auditing groups benefit from close proximity to the algorithms they are auditing.[29] If the internal auditors have questions, they may collaborate directly with peer departments. Additionally, because the auditors are part of the company itself, concerns over trade secrets disclosure are mitigated.[30]

Internal audits may fail to achieve effective regulation, though, because there are no standards for what these audits must entail, and insufficient independence exists between the auditor and auditee. If there are no outside standards for evaluating AI systems, a company may selectively investigate only certain aspects of their model or use whatever standards the system already meets in order to claim the AI system "passed" an audit.[31] Even when audits within a company do follow a consistent set of standards, there is no guarantee there is consistency in standards across companies that perform internal audits.[32] Results may not be published at all, and even in the event an internal auditor does publish audits, these auditors may have the incentive to only publish favorable results. Indeed, internal legal counsel may be careful to *avoid* finding out about an issue if their knowledge could lead to potential liability later on, and the same tendency can exist for hired auditors.[33]

---

28. *E.g.*, Isabel Kloumann & Jonathan Tannen, Building AI That Works Better for Everyone, META (Mar. 31, 2021), https://about.fb.com/news/2021/03/building-ai-that-works-better-for-everyone/ [https://perma.cc/TQ6V-F2HC] (describing Facebook's Responsible AI Team); *FATE: Fairness, Accountability, and Ethics Group*, MICROSOFT, https://www.microsoft.com/en-us/research/theme/fate/projects/ [https://perma.cc/9APS-V5M9].

29. *See* Goodman & Trehu, *supra* note 27, at 319–20.

30. *Id.*

31. For example, if AI model development team within a business is also the same team conducting the audit, that audit may not be properly validated since that staff has an incentive to find that model as valid. Patrick M. Parkinson, *SR 11-7: Guidance on Model Risk Management*, BD. OF GOVERNORS OF FED. RESERVE SYS. (2011), https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm [https://perma.cc/ZEC7-TXKD]; *see also* David Manheim, Sammy Martin, Mark Bailey, Mikhail Samin & Ross Greutzmacher, The Necessity of AI Audit Standard Boards (Apr. 11, 2024) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/2404.13060 [https://perma.cc/5ZYG-GBD2] (analogizing internal audits without external review tantamount to a company refusing to an external financial audit on the basis that their internal checking is sufficient).

32. The Institute of Internal Auditors sets standards for the internal accounting auditing profession but its understanding of the role of internal auditors is based on the expectations of external audits, which largely do not exist in the AI context. *Cf.* INSTITUTE OF INTERNAL AUDITORS, INTERNATIONAL PROFESSIONAL PRACTICES FRAMEWORK (2017 ed.) (authoritative guidance promulgated by the Institute of Internal Auditors for traditional auditing practices).

33. *See* Ng, *supra* note 25 ("Lawyers tell me, 'If we hire you and find out there's a problem that we can't fix, then we have lost plausible deniability and we don't want to be the next cigarette company,' ORCAA's founder, Cathy O'Neil, said. 'That's the most common reason I don't get a job.'"); United States v. OpenX Technologies, Inc., No. 2:21-cv-09693, 2021

Moreover, cherry picking and the selective publication of audits can lead to "audit washing," which diverts attention from or even excuses the very harm that the audit is supposed to mitigate.[34] When unfavorable findings are kept internal, discriminatory algorithms may remain in use, all while the public remains unaware of existing harm. As a result, remedial efforts — or even an acknowledgment of wrongdoing — may not be pursued.

Notwithstanding these shortcomings, internal audits can still serve a laudable function. In 2015, an internal review by Amazon halted deployment of a discriminatory resume screening system.[35] Amazon's efforts were still emblematic of the aforementioned concerns with internal audits: there was no indication of established discrimination standards, transparency as to existing standards, or accountability.[36] Nevertheless, these kinds of checks *should* still happen. As public sensitivity to algorithmic harms increases, intensive internal audits could become part of any AI software development process. But this episode inside Amazon shows that, even if these conditions are met, internal auditors cannot replace necessary *external* audits nor be a comprehensive regulatory solution on their own.[37] Amazon can afford entire teams of siloed internal auditors, but smaller enterprises surely cannot.[38]

WL 6751464 (C.D. Cal. Dec. 27, 2021) (imposing civil penalties where a human review of algorithmic decisions failed to ensure compliance with the Children's Online Privacy Protection Act (COPPA) and Federal Trade Commission (FTC) Act; it was the human review that provided "actual knowledge" necessary to obtain the penalties, whereas the algorithm alone would not).

34. Goodman & Trehu, *supra* note 27, at 302–03; Alex Engler, *Auditing Employment Algorithms for Discrimination*, BROOKINGS INST. (Mar. 12, 2021), https://www.brookings.edu/research/auditing-employment-algorithms-for-discrimination/ [https://perma.cc/K2GG-JX3M].

35. Jeffrey Dastin, *Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women*, REUTERS (Oct. 18, 2018), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G [https://perma.cc/V5TB-B2XG].

36. In this case, accountability could include improving future practices. *Id.* ("Amazon edited the programs to make them neutral to these particular terms. But that was no guarantee that the machines would not devise other ways of sorting candidates that could prove discriminatory"). It could also mean remedying harms caused by the system when it was in use in the absence of other avenues for relief. *Id.* (noting that it is likely difficult to sue for employment discrimination by algorithm).

37. Adequate internal controls will smooth an external audit. Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson et al., *Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing*, 2020 PROC. CONF. ON FAIRNESS, TRANSPARENCY & ACCOUNTABILITY 33, 35 (2020).

38. Engler, *supra* note 34 (noting that cost and necessary expertise may pose barriers to internal auditing at industry scale).

2. Third-Party Auditors

Third-party auditors conduct audits without the audited party's knowledge or approval. Auditors in this category are hardly auditors at all, often operating as researchers aiming to study potential societal issues created or reinforced by AI. While independence is crucial for auditing (and, as discussed above, is a notable shortcoming of internal audits), some cooperation is nevertheless necessary to achieve meaningful outcomes. This limitation is ever-present for third-party auditors, as they do not know what data is being used and can only speculate about the algorithms themselves.[39] Unless otherwise required, potential auditing targets will likely not make the most sensitive portions of their AI systems available for investigation and may even take deliberate actions to keep portions of AI systems opaque.[40] Additionally, third-party auditors' interests may be motivated by their own goals, which may be narrowly targeted. The results and methods of these auditors may consequently be limited in scope and *de minimis* in impact.[41]

Additionally, the work of third-party auditors is vulnerable to obstruction and even complete interdiction. Targets of investigative activities can take both technical and legal steps to severely stifle and even absolutely prohibit the work of researchers engaging in an audit. For example, researchers with the Ad Observatory Project at New York University (NYU) had been investigating how Facebook's ad delivery system may have amplified political misinformation.[42] But in August 2021, Facebook disabled the accounts of the NYU researchers because the company stated the researchers' actions, including data scraping, violated Facebook's Terms of Service (TOS).[43]

Additionally, an uncertain legal landscape can chill independent investigation. The Computer Fraud and Abuse Act (CFAA) was first enacted in 1984 to prevent accessing computer systems "without authorization."[44] In the last decade, it has been wielded to enforce TOS

39. Access to training data, training procedures, untrained models, and trained models are four different data points valuable to auditors that are frequently made inaccessible to auditors. *See generally* Sarah H. Cen et al., *Auditing AI: How Much Access Is Needed to Audit an AI System?*, THOUGHTS ON AI POL'Y (Sep 14. 2023), https://aipolicy.substack.com/p/ai-accountability-transparency-2 [https://perma.cc/3AEH-8JWD].

40. There is a litany of reasons why companies may take measures to keep their AI systems secretive, such as protecting trade secrets, maintaining competitive advantages over competitors, and mitigating the risk of company practices. *Id.*

41. Christian Sandvig, Kevin Hamilton, Karrie Karahalios & Cedric Langbort, *Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms*, 64 INT'L COMM'N ASS'N 1, 12–14 (2014) (describing the "scraping" and "sock-puppet" audit methods, both of which are feasible for third party auditing of AI).

42. Barbara Ortutay, *Facebook Shuts Out NYU Academics' Research on Political Ads*, AP NEWS (Aug. 4, 2021, 6:46 PM EST), https://apnews.com/article/technology-business-5d3021ed9f193bf249c3af158b128d18 [https://perma.cc/K4BP-ZWK2].

43. *Id.*

44. *See* 18 U.S.C. § 1030(a)(1).

violations and halt outside auditing efforts.[45] Even apparent victories for independent researchers, such as *Sandvig v. Barr* in 2020[46] and *Van Buren v. United States* in 2021,[47] left several questions unresolved.[48] For example, although the TOS policies considered in these cases did not amount to cognizable access restrictions, neither decision determined under what circumstances a written policy *would* create a CFAA violation.[49] Additionally, even if researchers' methods are deemed to not violate the CFAA, companies can still develop technical barriers to restrict audits' efficacy.[50]

3. External Auditors

External auditors are third-party entities that are hired to conduct an audit of an AI system but are not part of the everyday operations of the audited company. There is a growing industry of AI auditors that fall into this category.[51] While no general regulation mandating AI audits currently exists, audited companies may hire external auditors to confirm compliance with human rights standards,[52] sector-specific regulations,[53] or particularized measures of fairness.[54] Audits may also verify the accuracy of an AI system, in turn creating reputational credibility for the company deploying the AI.[55] External auditors benefit

45. Metaxa et al., *supra* note 25, at 298–99.

46. 451 F. Supp. 3d 73 (D.D.C. 2020).

47. 141 S. Ct. 1648 (2021).

48. Lee, *supra* note 21, at 1321 ("The Court seems to suggest that certain forms of technological savvy are permitted by law while others are prohibited with the force of the CFAA. . . . Yet it does not provide specific guidance as to how parties should navigate the difficult question of what online information is technically 'in the public forum' and consequently freely accessible.").

49. *See id.*; Aaron Mackey & Kurt Opsahl, *Van Buren is a Victory Against Overbroad Interpretations of the CFAA, and Protects Security Researchers*, EFF (June 3, 2021), https://www.eff.org/deeplinks/2021/06/van-buren-victory-against-overbroad-interpretations-cfaa-protects-security [https://perma.cc/F3AF-Z59G]. *See generally* Facebook v. Power Ventures, 844 F.3d 1058 (2016) (holding cease-and-desist letter created a CFAA violation for unauthorized access, but that violation of website TOS, without more, did not create liability under the CFAA).

50. *See* Issie Lapowsky, *Platforms vs. PhDs: How Tech Giants Court and Crush the People who Study Them*, GW INST. FOR DATA, DEMOCRACY & POLS. (Mar. 19, 2021), https://www.protocol.com/nyu-facebook-researchers-scraping [https://perma.cc/2PU5-K5P6].

51. *See, e.g.*, O'NEIL RISK CONSULTING & ALGORITHMIC AUDITING, https://orcaarisk.com [https://perma.cc/ADD3-W3LX]; FIDDLER, https://www.fiddler.ai [https://perma.cc/8X4L-ZVZJ]; ARTHUR, https://www.arthur.ai [https://perma.cc/9FWN-GVDA].

52. *See, e.g.*, Lorna McGregor, Daragh Murray & Vivian Ng, *International Human Rights Law as a Framework for Algorithmic Accountability*, 68 INT'L & COMPAR. L.Q. 309, 319 (2019).

53. *See, e.g.*, HEALTH INFORMATION ASSOCIATES, https://hiacode.com/ai-validation-audit [https://perma.cc/6TL9-SYT9].

54. *See* Goodman & Trehu, *supra* note 27, at 324–25, 309.

55. *See* Shlomit Yanisky-Ravid & Sean K. Hallisey, *"Equality and Privacy by Design": A New Model of Artificial Intelligence Data Transparency via Auditing, Certification, and Safe*

from a combination of greater system access than independent auditors and a greater level of detachment as compared to internal auditors, allowing external auditors to design and perform tests that probe an audited entity's models for adherence to applicable laws, fairness standards, and the audited entity's own policies.[56]

Although external auditors can provide benefits that other forms of auditing cannot, some shortcomings of other auditing approaches can nevertheless appear with external audits. For example, while the relationship between external auditors and the audited party is cooperative, an external auditor may be inclined to care more about *its* reputation than that of its client. Furthermore, the financial component for the external auditor may drive its motivations, which could pressure the auditor to publish narrow and positive findings so that the auditor is retained to conduct audits in the future.[57] A lack of mandated auditing standards likewise means that external auditors could shift their standards of success, potentially also resulting in a form of audit washing.

4. Political Developments Related to the Role of Auditors

Recent political developments have highlighted the crucial role that auditors will play in the future of AI regulation. The Biden administration issued an executive order in October 2023 on the "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence."[58] The Order built on the administration's earlier "Blueprint for an AI Bill of Rights," which established principles of equity and civil rights for AI systems by directing agencies to study and implement protective AI measures.[59] Notably, the Order identified a purpose for AI auditors in multiple areas that coincides with what this Note defines as the role of the external auditor. For instance, the order

---

Harbor Regimes, 46 FORDHAM URB. L.J. 428, 475–79 (2019); *see also* Shea Brown, Jovana Davidovic & Ali Hasan, *The Algorithm Audit: Scoring the Algorithms that Score Us*, 8 BIG DATA & SOC'Y 1, 2, 7 (2021).

56. *See* Goodman & Trehu, *supra* note 27, at 318–19 (describing how an audit's potential rigor is contingent on the level of access that auditors are provided and that conducting "reasonably competent inquiries" necessarily requires access to information); *see also* Engler, *supra* note 34 (noting that when the auditee is also the person paying the auditor, the auditor may make subjective choices in the audit's methodology to paint the auditee in a favorable light).

57. Engler, *supra* note 34 (describing HireVue's selective release of findings from an already limited audit, a decision criticized for making the audit little more than a PR stunt); BLUEPRINT, *supra* note 1, at 20 (stressing the importance of independent evaluation as to ensure that such audits can be "trusted to provide genuine, unfiltered access to the full system.").

58. Exec. Order No. 14110, 88 Fed. Reg. 75,191, 75,193 (Oct. 30, 2023) (Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence) [hereinafter AI Order].

59. *See FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence*, WHITE HOUSE (Oct. 30, 2023), https://www.dhs.gov/archive/news/2023/10/30/fact-sheet-biden-harris-administration-executive-order-directs-dhs-lead-responsible [https://perma.cc/X7AN-CMDP]. *See generally* BLUEPRINT, *supra* note 1.

contemplates establishing "guidance and benchmarks for evaluating and auditing AI capabilities, with a focus on capabilities through which AI could cause harm, such as in the areas of cybersecurity and biosecurity"[60] and reporting requirements for "dual use" technologies based on computing power.[61] The Order also singles out "synthetic content," which is content created by generative AI, for auditing and maintenance connected to its ability to produce illegal content such as child sexual abuse imagery. Lastly, to ensure fair access to public benefits, the Order calls on all public benefits administrators to "enable auditing and, if necessary, remediation of the logic used to arrive at an individual decision or determination to facilitate the evaluation of appeals" of benefits.[62] While the Order does not provide a detailed or concrete framework for AI auditing methodologies, it acts as an indication confirming that the time is ripe to consider how AI audits should be structured.

B. Structure of AI Audits

The structure of AI audits should reflect the reality of AI development with a targeted focus on the areas of data, model, and deployment.[63] Through an integrated and comprehensive approach, this breakdown attempts to account for the importance of each component of AI development — and highlight the deleterious effects of overlooking their interactions. Each of these components presents a potential avenue wherein human intervention and decision-making involved with a product's development can cause downstream consequences to end users and communities more broadly. Data collection, provenance, and management are accompanied by risks, such as encoding and reifying bias, either deliberately or unintentionally. Model development introduces risks associated with opaque algorithms; decisions related to model parameterization and training error levels are of critical attention in this component as well.[64] The deployment component steps beyond technical development and includes the governance aspects of AI systems, as well as the applicability of laws and regulations. Moreover,

60. AI Order, *supra* note 58, § 4.1(a)(i)(C).

61. *See id.* § 4.2(a)–(b).

62. *Id.* § 7.2(b)(ii)(E).

63. *See* Carter, *supra* note 10.

64. "Parameterization" refers to the process of selecting and initializing the "parameters" and "hyperparameters" of a model. Parameters are internal to the model. They may be "learned" or estimated during training of the model. While the parameters are updated during training, parameters must be selected and given a starting value. Hyperparameters are "top level" parameters that control the learning process, but they are not a part of the resulting model after training and are set by engineers. Kizito Nyuytiymbiy, *Parameters and Hyperparameters in Machine Learning and Deep Learning*, TOWARDS DATA SCI. (Dec. 30, 2020), https://towardsdatascience.com/parameters-and-hyperparameters-aa609601a9ac [https://perma.cc/92UJ-2BUE].

this component acknowledges that while an entity that deploys AI may not be responsible for its technical development, it is responsible for a deployment that runs afoul of regulation or established norms.

The goal of AI audits should be two-fold: preserve fairness and protect rights. An AI system has failed to preserve "fairness" when it is misleads its users or subjects about its operation or when it allows for fraud to be perpetrated against the same. Thus, audits should curtail companies from overpromising users of its AI system's capabilities. In some cases, auditing should protect against pernicious deception that can severely harm users, customers, or those indirectly impacted by such systems. This conception of fairness attempts to protect users from deliberate *human* actions that facilitate the design of harmful AI systems.

From a rights perspective, as AI advances, existing rights must be protected even when, and perhaps especially when, an AI system is not purposefully designed to violate any right but nonetheless does. This mandate related to "rights" includes, but is not limited to, rights rooted in the Constitution,[65] non-discrimination, statutory rights that regulate individuals' relationships with employers and necessary services,[66] and interests in "new property" that often define individuals' interactions with government today.[67] Rights protected at law represent a minimum for the protection we envision. Data privacy rights, for example, are not legally protected to the same extent everywhere in the United States but are nevertheless an example of additional rights that AI audits can and should seek to preserve. This is an expansive vision of protected rights by design, guided by the premise that technological advancement should not erode rights.[68]

We note that while illuminating the *technical* portions of AI systems is of critical importance, a holistic and effective audit advances fairness and protects rights by also capturing and explaining how and

---

65. These include privacy and associational rights. *See, e.g.*, NAACP v. Alabama, 357 U.S. 449 (1958) (associational rights); Carpenter v. United States, 585 U.S. 296 (2018) (privacy of cell site location information). And liberty-preserving due process rights. *See* Griswold v. Connecticut, 381 U.S. 479 (1965); United States v. Carolene Prods. Co., 304 U.S. 144, 153 n.4 (1938); *see also* Katherine Freeman, *Algorithmic Injustice: How the Wisconsin Supreme Court Failed to Protect Due Process Rights in* State v. Loomis, 18 N.C. J.L. & TECH. ON. 75, 99 (2016).

66. *See, e.g.*, Fair Housing Act, 42 U.S.C. § 3601 *et seq.*; Fair Credit Reporting Act, 15 U.S.C. § 1681 *et seq.* The statutory rights we are focused on can be essentially distilled to the generally applicable protections that rebuke the "*Lochner* era's" view of economic rights that prevented such legislation.

67. *See generally* Goldberg v. Kelly, 397 U.S. 254 (1970) (establishing that entitlements such as welfare, pensions, professional licenses are forms of "new property" that trigger due process rights already recognized for "traditional property").

68. *Cf.* United States v. Jones, 565 U.S. 400, 421 (2012) (Alito, J., concurring) (quoting Kyllo v. United States, 533 U.S. 27, 34 (2001)) (expressing concern over Fourth Amendment doctrine not adapting to technology that trivializes intrusions that the Fourth Amendment would have protected at the time of the adoption of the Constitution).

where AI systems become enmeshed with *human* actions. No AI system exists without its team of human technicians, and these human decisions play a large, if not primary, role on an AI system's output and corresponding impact.[69]

The continued adoption of AI will assuredly usher in entirely new legal paradigms, but policymakers are not powerless today to shape the development of potential paradigms. AI audits can be deployed to protect the twin maxims that AI must be "fair" — it cannot overpromise or deceive — and that rights must be protected — AI cannot erode what we already have. AI audits can achieve a balance where these maxims do not compromise technological development. The audit components introduced below are aimed at this goal. Thus, if an audit component is not advancing this goal, the audit has failed, and the audit approach must be reevaluated with an application-specific approach; what serves an audit's purpose in one area may not be enough in another. This goal of AI audits is non-exhaustive, and there are other plausible aims for audits, including audits with technical objectives for security and efficiency.[70]

1. Data

As articulated above in the insurance rates example, an inquiry into the data that an audited entity uses to train its AI systems can be revealing. An auditor may be able to identify prohibited practices or potential harm from the data the audited party has used: data may encode bias,[71] be prone to abuse, invade privacy, or not be relevant to the purpose of the system.[72] Although data is often an input, it does not come out of

69. *See* Goodman and Trehu, *supra* note 27, at 320 (noting that the "complex human and sociotechnical choices" are important variables to analyze in any review as human decisions ultimately set the AI system's objectives and what results the AI system should optimize for); *see also* Jennifer Cobbe, Michelle Seng Ah Lee & Jatinder Singh, *Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems*, 2021 PROC. CONF. FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 598, 599 (2021) ("[Algorithmic decision-making] itself — even understood as a process — operates not on its own, but as part of a wider socio-technical system.").

70. For instance, identifying the use of "memory-safe" programming languages. *Statements of Support for Software Measurability and Memory Safety*, WHITE HOUSE (Feb. 26, 2024), https://bidenwhitehouse.archives.gov/oncd/briefing-room/2024/02/26/memory-safety-statements-of-support/ [https://perma.cc/TF3J-Q6GA].

71. *See* FED. TRADE COMM'N, DATA BROKERS: A CALL FOR TRANSPARENCY & ACCOUNTABILITY (2014), https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf [https://perma.cc/JY6G-2SBX] [hereinafter Data Brokers Report] (expressing that in the context of credit reporting, companies using AI must be wary of bias in data they use as input and the data they produce).

72. *Statement of Commissioner Alvaro M. Bedoya Regarding Report to Congress on Combating Online Harms Through Innovation*, FED. TRADE COMM'N, MATTER NO. P214501 (June 16, 2022) (discussing risks of errors from the general application of natural language models trained using only certain kinds of language).

nowhere. In fact, cleaning and preparing data is a significant portion of what AI professionals do.[73]

Moreover, the generation and use of "synthetic data" is becoming a crutch for companies using AI.[74] What might merely be the input for one company or one team within a company may turn into the output of another team or outside vendor.[75] These processes, like the implementation of a model, could themselves introduce bias to otherwise unbiased data and compound harmful trends by creating feedback loops when used in training.[76] Consequently, these synthetic data systems should receive the same scrutiny as the AI systems themselves.[77]

In the rapidly growing field of generative AI, recent research has demonstrated that the data used to train large language models, such as OpenAI's ChatGPT and Meta's LLaMA, can cause these models to have distinctive political dispositions in the outputs they produce. For example, Google's AI model has been found to be more politically conservative than that of GPT-3. Research has demonstrated that this difference may be due in part to the difference in each model's training data — Google's AI model training set consists of books which may be more conservative than GPT-3's training set of internet sources. Irrespective of this normative question, it is beneficial for customers to know that a model may produce outputs with a particular ideological bent and auditing the data component of AI systems can help uncover any such biases and foster holistic disclosure.

2. Model

AI auditors should seek to encourage the explainability of AI systems to fulfill the notions of fairness, equity, and safety that often underlie calls for algorithmic transparency.[78] Transparency is the

73. *See* Sean Michael Kerner, *What Are Data Scientists' Biggest Concerns? The 2022 State of Data Science Report Has the Answers*, MACH. (Sep. 14, 2022, 6:00 AM), https://venturebeat.com/ai/what-are-data-scientists-biggest-concerns-the-2022-state-of-data-science-report-has-the-answers/ [https://perma.cc/NX77-ATQL].

74. Sara Catellanos, *Fake It to Make It: Companies Beef Up AI Models with Synthetic Data*, WALL ST. J. (July 23, 2021, 5:30 AM ET), https://www.wsj.com/articles/fake-it-to-make-it-companies-beef-up-ai-models-with-synthetic-data-11627032601 [https://perma.cc/M3S2-ZVY2]. Synthetic data is computer-generated data "to augment or replace real data" in training of AI models. Kim Martineau, *What is Synthetic Data?*, IBM (Feb. 8, 2023), https://research.ibm.com/blog/what-is-synthetic-data [https://perma.cc/8Z5Q-SH3S].

75. Data Brokers Report, *supra* note 71 (describing the early phases of the life cycle of big data, including data procurement and processing, noting that datasets may be bought and sold).

76. *See generally* Kristina Lum & William Isaac, *To Predict and Serve?*, 13 SIGNIFICANCE 14 (2016).

77. In addition, any data that a company makes public or more transparent can increase the efficiency of audits. Yanisky-Ravid & Hallisey, *supra* note 55, at 477–78 (expounding the benefits and meaning of transparency in data).

78. *See* Margot E. Kaminski, 5 *Understanding Transparency in Algorithmic Accountability*, *in* THE CAMBRIDGE HANDBOOK OF THE LAW OF ALGORITHMS 121, 121 (Woodrow

disclosure, public or through an intermediary, of characteristics of a model, such as its inputs, its parameters, error levels, relevant data, or even source code.[79] Explainability refers to the capacity of an algorithm to generate, or lend itself to the development of, a human-comprehensible accounting of its decision-making process.[80]

Seeking explainability of algorithms can remedy shortcomings of transparency in evaluating AI systems.[81] Transparency in itself does not contribute towards AI auditing's goal of preserving fairness or protecting rights. First, even if a company were to make its model's code available, that may present little to no value due to the technical complexity of the code. Additionally, transparency of code alone without corresponding inputs does not recreate the context in which the model produces an output. Relying on transparency alone could allow for obfuscation of harmful algorithms.[82] Transparency as an approach to AI regulation does not adapt to the different outcomes of its decisions, either based on their seriousness or potential for discrimination, for example.[83] Context will necessarily dictate the remedy to algorithmic harms, so this is a significant shortcoming of a transparency-focused

---

Barfield ed., 2020). With a broad understanding of "transparency," FTC Commissioner Rebecca Slaughter argues for this kind of inquiry. *See* Rebecca Kelly Slaughter, *Algorithms and Economic Justice: A Taxonomy of Harms and a Path Forward for the Federal Trade Commission*, 23 YALE J.L. & TECH. 1, 47–48 (2021); *see also* BLUEPRINT, *supra* note 1, at 6–7.

79. Kaminski, *supra* note 78, at 127–28.

80. *See* Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter & Lalana Kagal, *Explaining Explanations: An Overview of Interpretability of Machine Learning*, 5 INT'L CONF. ON DATA SCI. & ADVANCED ANALYTICS 80, 81 (2018). A subset of "explainability" includes the concept of "interpretability," which emphasizes "human-understandable" representations of the decision process. *Id.*; *see also* Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085, 1109–10 (2018) (introducing explainability approaches embraced by machine learning practitioners); Kaminski, *supra* note 78, at 127 (offering disambiguation of multiple "orders" of transparency, where "first-order" transparency applies to the technology and is our focus here).

81. *See* Kaminski, *supra* note 78.

82. Even weak forms of explainability can be misleading. Erwan Le Merrer, Ronan Pons & Gilles Trédan, *Algorithmic Audits of Algorithms, and the Law*, HAL OPEN SCI. 1, 4 (2022). But simple automated auditing can detect this obfuscation. *Id.* The "bouncer problem" is illustrative: A bouncer with an intent to (impermissibly) discriminate will simply provide a fake accounting of his "inputs" if asked. The factors a transparent bouncer reveals will not change from visit to visit, thus yielding no information about the discrimination over multiple visits; however, once the bouncer begins to provide explanations, contradictory explanations for individuals who are identical according to the provided explanation become evidence of the discrimination. Erwan Le Merrer & Gilles Trédan, *The Bouncer Problem: Challenges to Remote Explainability*, 2 NATURE MACH. INTEL. 529, 529 (2020); *see also* Melissa Heikkilä, *A Bias Bounty For AI Will Help to Catch Unfair Algorithms*, MIT TECH. REV. (October 20, 2022), https://www.technologyreview.com/2022/10/20/1061977/ai-bias-bounty-help-catch-unfair-algorithms-faster/ [https://perma.cc/Y5ZU-BE5T].

83. This is still the case when the target audience of transparency is the general public. *See, e.g.*, Matthew Gooding, *Elon Musk's Plan for an Open-Source Algorithm Won't Solve Twitter's Problems*, TECH MONITOR (Apr. 26, 2022), https://www.techmonitor.ai/digital-economy/ai-and-automation/open-source-twitter-algorithm-elon-musk [https://perma.cc/G6NT-J9RZ].

approach.[84] Furthermore, the value of transparency is limited for more complex models. Linear regression and some statistical methods lend themselves to human understanding,[85] even based only on inputs and parameters of the model.[86] However, these methods cannot compete with the performance of more complex models, such as neural networks.[87] Transparency inputs and parameters that may provide a clear picture for simpler models will be meaningless on their own for more complex models.[88]

Due to transparency's shortcomings, AI audits should extend beyond the algorithm and encourage explainability as a means to reach accountability.[89] Auditors should determine if an algorithm is explainable but is not yet explained, purposefully obscuring its processes, or is just very complex.[90] Furthermore, an auditor should seek justification when an algorithm is "unexplainable" and reconcile that justification with the algorithm's performance.[91] Auditors may review an algorithm's description alongside code and results of technical analyses to determine if the audited entity has implemented its objective honestly and in the manner marketed towards its customers.[92] Model documentation is therefore both part of the audit and can itself be a product of

---

84. Mike Ananny & Kate Crawford, *Seeing Without Knowing: Limitations of the Transparency Ideal and its Application to Algorithmic Accountability*, 20 NEW MEDIA & SOC'Y 973, 978 (2018).

85. *Interpretability Versus Explainability*, subsection to *Model Explainability with AWS Artificial Intelligence and Machine Learning* Solutions, AWS (2022), https://docs.aws.amazon.com/whitepapers/latest/model-explainability-aws-ai-ml/interpretability-versus-explainability.html [https://perma.cc/L3MQ-W7UN].

86. Paul B. de Laat, *Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?*, 31 PHIL. & TECH. 525, 536 (2017).

87. *See* Madalina Busuioc, *Accountable Artificial Intelligence: Holding Algorithms to Account*, 81 PUB. ADMIN. REV. 825, 830 (2021) (remarking that the most powerful AI methods today do not lend themselves to easy comprehension). Indeed, "deep" and "generative" networks may even count as a strength the fact that they buck conventional human thought processes. *How to Use AI to Discover New Drugs and Materials with Limited Data*, IBM RSCH. BLOG (Apr. 13, 2022), https://research.ibm.com/blog/ai-discovery-with-limited-data [https://perma.cc/78SR-P6UG] (discussing the use of "generative" AI for drug discovery, including its surprising breakthroughs).

88. Ananny & Crawford, *supra* note 84, at 982–83 (explaining that a direct link between visibility and understanding cannot be assumed for all models and applications).

89. *Id.* at 983 (recognizing an algorithmic system as an "assemblage" of human and computer actors); *see also* BLUEPRINT, *supra* note 1, at 18–20 (establishing that explainability fits into a broader effort to stem the harm and promote the benefits of automated systems).

90. Amina Adadi & Mohammed Berrada, *Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)*, 6 IEEE ACCESS 52138, 52147 (2018).

91. *See* BLUEPRINT, *supra* note 1, at 18–20 (establishing "what should be expected of automated systems").

92. Because AI deployment relies on prepackaged libraries or models, it could be worth considering whether the target of auditors should be the business applying an AI system or, for example, the developers of so-called "foundation models." Stanford Engineering Staff, *The Future of AI Chat: Foundation Models and Responsible Innovation*, STAN. ENG'G (Dec. 1, 2023), https://engineering.stanford.edu/news/future-ai-chat-foundation-models-and-responsible-innovation [https://perma.cc/NWP7-T32B] (discussion with Professor Percy Liang).

the process.[93] In enforcing this component, auditors should seek to limit perverse incentives. Companies may use "unexplainable" methods from the start to avoid inconvenient explanations.[94] For example, a linear regression *could* be implemented with a neural network, obscuring regression coefficient values and the inherent interpretability of the regression model. Moreover, explainability may be of particular importance to generative AI where exact outputs may be unknown but particular *categories* of outputs could be inferred. For example, in the news editorial example above, knowledge of the exact interactive feature the AI-tool provider will implement on any given story might be impossible to predict or explain.[95] However, it could be technically feasible to evaluate features the generative AI tool has implemented previously as well as the deleterious consequences that accompanied these features. This knowledge alone could lead to more holistic disclosures on the AI-tool provider's end and more attenuated oversight and prophylactic issue spotting by outside observers to mitigate the risk of reputational damage and harm to individuals who are functionally bystanders.

In order for technical audits of the model component to be effective, access to data and the model is of utmost importance.[96] Indeed, the importance of access strengthens the need for government authority behind audits to neutralize any adverse incentives the auditee may provide to an auditor.[97] Auditors may examine a model's operative logic and in conjunction run their own analyses of its performance.[98] This

---

93. *See* STEPHEN GOLDSMITH & WILLIAM D. EGGERS, GOVERNING BY NETWORK: THE NEW SHAPE OF THE PUBLIC SECTOR 123–24 (2004) (detailing how lack of documentation complicated government auditing efforts and ultimately narrowed results and potential benefit of the entire process).

94. *Contra* Deven R. Desai & Joshua A. Kroll, *Trust But Verify: A Guide to Algorithms and the Law*, 31 HARV. J.L. & TECH. 1, 16–17 (2017) (discussing designing algorithmic systems to enable audits by regulators).

95. *See supra* note 16 and corresponding discussion.

96. Kroll et al., *supra* note 23, at 661 (establishing shortcomings of "black-box" audits compared to "white-box" audits, in which the auditor is given access to peer inside the system, not solely to query it). This access would likely need to be mandated because incentives of audited entities are likely slanted against it, but there is consensus among auditors that mandates for audits and disclosure should exist. *See* Figure 2 *in* Costanza-Chock et al., *supra* note 21.

97. *See* Miriam Seifter, *Rent-a-Regulator: Design and Innovation in Privatized Governmental Decisionmaking*, 33 ECOLOGY L.Q. 1091, 1126 (2006) (noting it is crucial that interests between the auditors and the government are aligned); Engler, *supra* note 34 (describing and contrasting audits of HireVue and Pymetrics based on access, disclosure, and funding); Ajunwa, *supra* note 26, at 671–72; *see also* Costanza-Chock et al., *supra* note 21, at 7 (noting that entities using AI might try to *avoid* understanding their systems to maintain plausible deniability as to potential effects).

98. Ananny & Crawford, *supra* note 84, at 981 ("Learning about complex systems means not simply being able to look inside systems or take them apart. Rather, it means dynamically interacting with them in order to understand how they behave in relation to their environments."); Walter A. Mostowy, *Explaining Opaque AI Decisions, Legally*, 35 BERKELEY TECH. L.J. 1291, 1304 (2020) (describing approaches with strengths in different applications,

would likely include constructing test inputs to the model targeted to reveal decisive factors and where the model performs in unintended or prohibited ways. In doing so, auditors assess compliance with applicable laws and best practices developed by auditors, experts, and government.[99] For example, differing error levels for predictions could be indicative of disparate impacts for protected classes, and a lack of accuracy could be grounds for enforcement actions.[100] New York City has required that auditors compute and report specific "scores" with defined formulas.[101] Accordingly, auditors should insist on mitigating steps during model training and validation processes informed by concrete tests, and with an eye towards advancing explainability for the future.[102]

The relationship between AI auditing and explainability is mutually reinforcing. A virtuous cycle within audits begins as auditors incorporate precepts of explainability into their practices, avoiding the pitfalls of transparency on its own. It continues as audited entities employ explainable algorithms, thereby approaching the root of the concept of AI accountability that both the developer and deployer of an AI system can answer for its decisions.[103] Furthermore, auditors' engagement with algorithm development can advance the promise of AI as a force for equity.[104]

---

including visualization, knowledge extraction, influence measurement, and example generation).

99. *See* Ajunwa, *supra* note 26, at 667 (discussing how government and non-governmental groups may cooperate to develop standards, as in the case of energy efficiency); Andrew Smith, *Using Artificial Intelligence and Algorithms*, FED. TRADE COMM'N (Apr. 8, 2020), https://privacysecurityacademy.com/wp-content/uploads/2021/01/Using-Artificial-Intelligence-and-Algorithms-_-Federal-Trade-Commission.pdf [https://perma.cc/Q34B-5RES] (explaining the requirement of accuracy and robustness of models that can be found in the FCRA).

100. *See, e.g.*, *Texas Company Will Pay $3 Million to Settle FTC Charges That it Failed to Meet Accuracy Requirements for its Tenant Screening Reports*, FED. TRADE COMM'N (2018), https://www.ftc.gov/news-events/news/press-releases/2018/10/texas-company-will-pay-3-million-settle-ftc-charges-it-failed-meet-accuracy-requirements-its-tenant [https://perma.cc/JT2F-KUTY].

101. *See* 6 R. CITY N.Y. § 5-301 (effective July 5, 2023) (defining the "selection rate" and "impact ratio" measures and the categories over which to compute them).

102. *See* Kroll et al., *supra* note 23, at 688 (describing the method of "regularization"); Oren Bar-Gill, Cass R. Sunstein & Inbal Talgam-Cohen., Algorithmic Harm in Consumer Markets 37 (unpublished manuscript) (Jan. 2023). Required improvements could include taking steps in the model design phase to chronicle the purpose of different components of the decision, an accounting of inputs and the parameter space. *See* Andrew D. Selbst, *An Institutional View of Algorithmic Impact Assessments*, 35 HARV. J.L & TECH. 117, 146 (2021) (discussing the utility of transparency at various stages of the development process, even for complex algorithms).

103. *See* Mark Bovens, *Analysing and Assessing Accountability: A Conceptual Framework*, 13 EUR. L.J. 447, 450 (2007) ("The most concise description of accountability would be: 'the obligation to explain and justify conduct.'").

104. *See* Jon Kleinberg et al., *Human Decisions and Machine Predictions*, Q.J. ECON. 237, 241 (2018) ("[A] properly built algorithm can reduce crime and jail populations while

3. Deployment

Investigating the deployment of AI recognizes that the "soft" aspects of AI systems — the people and institutions around them — can contribute to harms that audits attempt to stem and may also pose barriers to remedial measures.[105] The practice of governance auditing is instructive, as this practice supplements technical investigations with industry-specific risk mitigation.[106] Moreover, there may be human contributions to output decisions, even if AI is involved in some capacity, that must be interrogated.[107]

An audit should examine how an audited entity manages an AI system.[108] A model is not static, whether a model is updated in an effort to become more profitable or modified in response to vulnerabilities. Thus, an audit should review the protocol for updating a model.[109] For example, an auditor should investigate how the audited party has prepared to respond to requests from regulators and courts,[110] as well as individuals in some jurisdictions.[111] As standards are set across both industry and academia, auditors should assess how an audited entity engages with the broader AI community.[112] Even if algorithmic pre-

simultaneously reducing racial disparities. In this case, the algorithm can be a force for racial equity.").

105. *See* Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian & Janet Vertesi, *Fairness and Abstraction in Sociotechnical Systems*, 2019 PROC. CONF. FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 59 (2019) (arguing that definitions of fairness that consider only "the machine learning model, the inputs, and the outputs," miss "the broader context, including information necessary to create fairer outcomes, or even to understand fairness as a concept"). These aspects of AI deployment are often neglected in technical audits. Costanza-Chock et al., *supra* note 21 (finding only 52% of audits examine the "[c]orporate environment/engineering process" and the "[e]xistence/quality of systems to report harm/appeal decisions").

106. *See* Avi Gesser, Bill Regner & Anna Gressel, *AI Oversight is Becoming a Board Issue*, HARV. L. SCH. F. ON CORP. GOV. (Apr. 26, 2022) (presenting governance aspects of AI deployment); Kitty Kay Chan & Tina Kim, *Auditing AI Governance*, INTERNAL AUDITOR (Feb. 21, 2022), https://internalauditor.theiia.org/en/articles/2022/february/auditing-ai-governance/ [https://perma.cc/B4B2-R3AW].

107. *See generally* Kevin Macnish, *Unblinking Eyes: The Ethics of Automating Surveillance*, 14 ETHICS INFO. TECH. 151 (2012).

108. *See* Gesser et al., *supra* note 106.

109. The Dutch government has adopted an approach that considers the management of the algorithm deployment. *See* NETH. CT. AUDIT, *An Audit of 9 Algorithms Used by the Dutch Government* (May 18, 2022), https://english.rekenkamer.nl/publications/reports/2022/05/18/an-audit-of-9-algorithms-used-by-the-dutch-government [https://perma.cc/52NE-JAA7].

110. Courts have sought insight into the inner workings of AI systems in situations as disparate as those in *State v. Loomis*, 371 Wis.2d 235 (2016) and *Houston Fed'n of Tchrs., Loc. 2415 v. Houston Indep. Sch. Dist.*, 251 F. Supp. 3d 1168 (S.D. Tex. 2017).

111. *See* Rogers v. BNSF Railway Co., No. 1:19-cv-03083 (N.D. Ill. Oct. 12, 2022) (granting relief to plaintiffs under the private right of action of the Illinois Biometric Information Privacy Act).

112. *See, e.g.*, *Cybersecurity Framework*, NAT'L INST. STANDARDS & TECH., https://www.nist.gov/cyberframework [https://perma.cc/4BFT-2BAC]. These developments also provide industry-specific rules that can hone audit focus. *See, e.g.*, *Artificial Intelligence*

clearance procedures emerge, there is still a need for continued scrutiny on deployment.[113]

In addition to specially developed standards, auditing requirements should incorporate standards established by past enforcement actions and regulator guidance.[114] For example, the Federal Trade Commission (FTC) has recognized accuracy and robustness requirements for AI systems through the Fair Credit Reporting Act,[115] including for vendor models.[116] The Equal Employment Opportunity Commission issued guidance during the Biden administration on "reasonable accommodations" in AI systems under the Americans with Disabilities Act.[117] In addition, as long as AI regulation remains fractured, auditors should assess the applicability of state and municipal regulations and the audited systems' compliance therewith.[118]

While certain uses of algorithms may eventually be banned, the auditing approach is meant to be an alternative to an outright ban.[119] Motivating harms should not be ignored, but outright bans may have

---

Ethics Framework for the Intelligence Community, INTEL. (June 2020), https://www.intelligence.gov/artificial-intelligence-ethics-framework-for-the-intelligence-community [https://perma.cc/5PRE-4XZW]; *Artificial Intelligence in the Securities Industry*, FINRA (June 2020), https://www.finra.org/sites/default/files/2020-06/ai-report-061020.pdf [https://perma.cc/G2QT-GV36]. Relevant developments could include circumscribing accepted AI applications or setting a consensus risk tolerance.

113. *Cf.* Andrew Tutt, *An FDA for Algorithms*, 69 ADMIN. L. Rev. 83, 111 (2017) (describing "pre-market approval" measures among broader approaches to regulation).

114. *See, e.g.*, In the Matter of Everalbum, Inc., No. 192-3172, 2021 WL 1922417, at *1 (MSNET May 6, 2021).

115. Andrew Smith, *Using Artificial Intelligence and Algorithms*, FEDERAL TRADE COMMISSION (Apr. 8, 2020), https://privacysecurityacademy.com/wp-content/uploads/2021/01/Using-Artificial-Intelligence-and-Algorithms-_-Federal-Trade-Commission.pdf [https://perma.cc/SW49-QYKP]. This would include seeking out any paradoxical outcomes. For example, if given a certain individual's credit profile, an additional late payment would result in an increase in credit score, this flaw should be identified by an auditor. Edge cases and assumptions of a model are another example of useful testing.

116. *Id.* (stating that the notice requirement for "adverse action[s]" applies to information obtained from a vendor that uses AI).

117. *See Trump Rolls Back Bidens AI Executive Order and Makes AI Infrastructure Push: Key Takeaways for Employers*, FISHER PHILLIPS (Jan. 23, 2025), https://www.fisherphillips.com/en/news-insights/trump-rolls-back-bidens-ai-executive-order.html [https://perma.cc/RVQ6-T6FT] (describing new Trump administration shifts from the Biden Administration's federal AI guidance).

118. *Algorithmic Accountability Act of 2022 Section-by-Section*, RON WYDEN, https://www.wyden.senate.gov/imo/media/doc/2022-02-03%20Algorithmic%20Accountability%20of%202022%20Section-by-section%20(SxS).pdf [https://perma.cc/R9SX-6ZL5].

119. *See* Kashmir Hill, *Microsoft Plans to Eliminate Face Analysis Tools in Push for 'Responsible A.I.,'* N.Y. TIMES (June 21, 2022), https://www.nytimes.com/2022/06/21/technology/microsoft-facial-recognition.html [https://perma.cc/V5JN-J5FF]. *But see* Albert Fox Cahn & Justin Sherman, *We Don't Need a Weak Laws Governing AI in Hiring — We Need a Ban*, FAST CO. (Apr. 15, 2021), https://www.fastcompany.com/90625587/new-york-city-ai-hiring-rules-ban [https://perma.cc/C7DZ-GRH5].

unintended consequences when enacted with imperfect information,[120] including the possible foreclosure of beneficial AI applications.[121] Moreover, banning AI forecloses the possibility of the *positive* transformative potential of AI and may not be feasible given the reliance several companies and customers have already developed with AI products.[122] Accordingly, auditors should satisfy themselves that the audited entity is taking steps to ensure that their systems are not being used for unauthorized purposes, an issue the FTC took interest in with respect to voice-cloning technologies.[123]

III. AI REGULATION THROUGH AI AUDITING

Implementing AI regulation through auditing will require complex interconnected laws phased in over time, but purposeful reforms can jumpstart a virtuous cycle of AI development. Government has a crucial role to play in establishing and reinforcing the necessary conditions for audits to achieve their regulatory potential. In addition to mandating audits for at least some AI systems, auditing standards should be consistently enforced, lest another unregulated industry emerge in an effort to regulate another. The government should also undertake efforts to manage and advance AI auditing technology.

A. Creating a Virtuous Cycle Through Oversight

Oversight is essential for realizing a potential AI auditing virtuous circle. In order to facilitate such a virtuous circle, auditors should be certified by an oversight body.[124] Certification should be based on an auditor's adherence to evaluation and reporting standards in their

120. Bar-Gill et al., *supra* note 102, at 43; *see also* FED. TRADE COMM'N, BIG DATA: A TOOL FOR INCLUSION OR EXCLUSION? UNDERSTANDING THE ISSUES 31–32 (2016) (asking of companies, "[d]oes your reliance on big data raise ethical or fairness concerns?").

121. *See generally* Erik Brynjolfsson, Danielle Li & Lindsey R. Raymond, *Generative AI at Work*, Nat'l Bureau Econ. Rsch. Working Paper 31161 (Nov. 2023) (detailing introduction of generative AI tools in workplaces, including productivity impacts on novice workers).

122. In less than one year after its launch, Chat GPT stated that it already had 100 million weekly users. Jon Porter, *ChatGPT Continues to be One of the Fastest-Growing Services Ever*, VERGE (Nov. 6, 2023, 1:03 PM EST), https://www.theverge.com/2023/11/6/23948386/chatgpt-active-user-count-openai-developer-conference [https://perma.cc/2DE9-6E4P].

123. *You Don't Say: An FTC Workshop on Voice Cloning Technologies*, FED. TRADE COMM'N (Jan. 28, 2020), https://www.ftc.gov/news-events/events/2020/01/you-dont-say-ftc-workshop-voice-cloning-technologies [https://perma.cc/T9KG-S235] (describing the Commission's efforts to study the use of voice cloning technology in phishing schemes and other social engineering scams, making it more difficult for consumers to identify fraud).

124. *See* Catherine J.K. Sandoval, *Technology Law as a Vehicle for Technology Justice: Stop ISP Throttling to Promote Digital Equity*, 36 BERKELEY TECH. L.J. 963, 982–83 (2021) (acknowledging role of regulatory action in reinforcing "virtuous cycle" of technological development).

audits, such as those introduced in the preceding sections. The financial auditing industry offers one instructive possibility as a starting point for an oversight scheme for AI auditing. Auditing in the financial sector is meant to "obviate the fear of loss from reliance on inaccurate information, thereby encouraging public investment in these industries."[125] AI auditing and its oversight can be pursued towards analogous ends — to obviate concerns about unfair systems and encourage responsible deployment, creating a virtuous cycle of AI development.[126]

The Securities and Exchange Commission (SEC) requires qualifying public companies to have their financial statements audited and performs oversight of auditors through the Public Company Accounting Oversight Board (PCAOB).[127] Established by the Sarbanes-Oxley Act of 2002, the PCAOB registers, reviews, and investigates accounting firms that perform audits of public companies and provides a useful comparison and potential model for the oversight of AI auditors.[128] In both the AI and financial contexts, enhancing public understanding of the roles of auditors is a crucial component of the virtuous cycle, and accordingly, reports of auditors' performance should be made publicly available, as the PCAOB does.[129] While regulation via auditing is not perfect, when seen in the context of the events that prompted Sarbanes-Oxley and the lasting trends it set off, auditing has not only proven to be better than nothing, but in some cases, game-changing.[130]

The PCAOB's audit standards are enforced by regular reviews of auditors' work for "deficiencies."[131] An auditor may subsequently

---

125. United States v. Arthur Young, 465 U.S. 805, 819 n.15 (1984); *see also* Richard S. Panttaja, *Accountants' Duty to Third Parties: A Search for A Fair Doctrine of Liability*, 23 STETSON L. REV. 927, 932 (1994) ("a fundamental objective of the audit is to enhance credibility of management's representations in financial statements").

126. Coglianese & Lampmann, *supra* note 26, at 183 (detailing the positive reinforcement effect of setting industry standards with the example of energy efficiency standards).

127. *See* 15 U.S.C. § 7211 (establishing the PCAOB and enumerating its duties).

128. *Id.*

129. PCAOB firm inspection reports can be found at: https://pcaobus.org/oversight/inspections/firm-inspection-reports. Firms with "quality control criticisms" are clearly indicated when the PCAOB has made public additional portions of the previously issued inspection report because of a failure to address certain quality control issues to the satisfaction of the Board within the 12 months following the date of the report. *See also* 6 R. CITY N.Y. §§ 5-304, 5-305 (effective July 5, 2023) (requiring that results of bias audits of automated employment decision tools be published publicly and notice be provided to applicants of the details of the automated system).

130. *See* Michael W. Peregrine, *The Lasting, Positive Impact of Sarbanes-Oxley*, HARV. L. SCH. F. ON CORP. GOV. (Dec. 20, 2021), https://corpgov.law.harvard.edu/2021/12/20/the-lasting-positive-impact-of-sarbanes-oxley/ [https://perma.cc/39DM-WC47] ("[The Act] sparked the corporate responsibility movement, which continues to impact corporate and leadership ethics and compliance with law. It remains one of the most consequential governance developments in history and serves as an important lesson for corporate officers, directors and their professional advisors.").

131. *Basics of Inspections*, PCAOB, https://pcaobus.org/oversight/inspections/basics-of-inspections [https://perma.cc/99AJ-WWLV]; *Guide to Reading the PCAOB's New Inspection Report*, PCAOB, https://assets.pcaobus.org/pcaob-dev/docs/default-source/inspections/docu

clarify and fix any identified deficiencies.[132] An analogous regime with enumerated standards for AI auditors could add some predictability to oversight, along with the flexibility of a "deficiency" standard that can adapt to changes in AI development practices and technology.[133] For example, a common deficiency in financial audits is the failure to gather sufficient evidence.[134] While this may be replicated with AI auditing, perhaps in part due to efforts by the audited entity to limit auditors' access,[135] what counts as "Sufficient Appropriate Audit Evidence" is liable to change with new AI systems.[136]

The legitimacy that an oversight scheme modeled on the PCAOB could confer to the practice of AI auditing could in turn drive greater demand for audits from companies deploying AI. Mandating audits for certain AI systems could also be a vital catalyst, as New York City has recently instituted for "automated employment decision tools."[137] After an audit mandate is instituted, the virtuous cycle continues as public awareness increases related to what successful audits mean and what algorithmic harms entail. In turn, this leads to greater demand for audits by trusted auditors, which then further bolsters the legitimacy of the auditors, clients, their AI systems, and the emerging institution of AI auditing.[138] Together, these reforms and the accompanying virtuous cycle can empower AI auditing to become an effective, nimble regulatory approach that can uphold standards and promote public accountability.[139]

That being said, the auditing approach is meant to be flexible, so while oversight is crucial, its implementation need not be rigid. Existing administrative agencies have addressed the impact of algorithms in their respective fields of expertise, and these efforts could be combined

---

ments/inspections-report-guide.pdf?sfvrsn=bc066f32_0 [https://perma.cc/6X2T-75PY] (describing contents of PCAOB's inspection report of auditors whose work it reviews).

132. *See Guide to Reading*, PCAOB, *supra* note 131.

133. *But see* Ng, *supra* note 25 (noting that the New York City law requiring that algorithmic hiring systems be audited did not originally spell out how an audit should be conducted).

134. Mark S. Beasley, Joseph V. Carcello & Dana R. Hermanson, *Top 10 Audit Deficiencies*, J. ACCOUNTANCY (Apr. 1, 2001), https://www.journalofaccountancy.com/issues/2001/apr/top10auditdeficiencies.html [https://perma.cc/KB3P-XHUH].

135. Engler, *supra* note 34.

136. *AS 1105: Audit Evidence*, PCAOB, https://pcaobus.org/oversight/standards/auditing-standards/details/AS1105 [https://perma.cc/G4DW-M6FT].

137. *See* 6 R. CITY N.Y. § 5-301 (effective July 5, 2023) (requiring "bias audits" for AI systems that make hiring and promotion decisions). AI systems could be selected for mandated audits based on complexity or sensitivity. *See* Tutt, *supra* note 113, at 107–08 (Table 1); Ajunwa, *supra* note 26, at 646–47 (discussing inadequacy of existing employment law in the context of automated hiring).

138. *See, e.g.*, Ajunwa, *supra* note 26, at 667–68 (describing the "Fair Automated Hiring Mark" to signal to applicants that an AI system in use operates fairly, which could encourage more diverse candidates to apply).

139. Note that the recent New York City law does not include significant oversight of the auditors performing the required audits. *See* 6 R. CITY N.Y. §§ 5-304–5-305.

for the oversight of auditors.[140] Alternatively, given this Note's proposal's general, cross-industry aims, an entirely new agency could also be desirable.[141] In either case, and so long as other federal and state regulation of AI remains sparse, the oversight agency or agencies must engage in significant standard-setting.[142]

Government's interests and auditors' interests must be aligned.[143] Slippage can occur when conflicts of interest between auditors and their clients are not policed, or when standards are not communicated or enforced with disciplinary action in response to malpractice — either upon review or when alleged by the client.[144] Oversight can rebalance the dynamic between auditor and client, returning leverage to the auditor and ensuring clients cannot sway audit results with the prospect of future business.[145] Fundamentally, it is the role of the government to bolster the auditor's position of independence and keep them honest at the same time.[146]

To encourage improvements in harmful AI systems and simultaneously encourage complete disclosure to auditors, a safe harbor policy for audited systems could be instituted by the oversight agency.[147] In the event of a failed audit, liability could be suspended for claims arising from the operation of the AI system for the duration of a remediation period during which the system must be brought into compliance.

---

140. *See* Rachel E. Barkow, *Insulating Agencies: Avoiding Capture Through Institutional Design*, 89 TEX. L. REV. 15, 53 (2010) ("shared responsibility may create a healthy competition between the two agencies, and it will be harder to capture two agencies instead of one"); Jody Freeman & Jim Rossi, *Agency Coordination in Shared Regulatory Space*, 125 HARV. L. REV. 1131, 1134–35 (2012) (suggesting fragmentation and overlapping regulatory responsibility amongst agencies can lead to collaboration).

141. *Cf.* Tutt, *supra* note 113, at 115–16 (advancing the argument for a centralized agency to review algorithms before their release.); *see also* Freeman & Rossi, *supra* note 140, at 1134. It is doubtful, however, that the Supreme Court would uphold the constitutionality of an oversight agency for auditing that is structured like the FTC after *Seila Law LLC v. CFPB*. 140 S. Ct. 2183 (2020). To have independence in auditing oversight, this function would likely have to be tacked onto an existing agency, like the FTC, with an agency structure compliant with *Humphrey's Executor v. United States*, 295 U.S. 602 (1935).

142. Todd Feathers, *Why It's So Hard to Regulate Algorithms*, MARKUP (Jan. 4, 2022, 8:00 AM ET), https://themarkup.org/news/2022/01/04/why-its-so-hard-to-regulate-algorithms [https://perma.cc/6LCV-GA3N] (arguing that *legislative* efforts to study AI regulation have almost universally failed). Input from industry and users alike is needed to create a standard that will be feasible and still protect individuals. *See Transparency and Explainability (Principle 1.3)*, OECD.AI, https://oecd.ai/en/dashboards/ai-principles/P7 [https://perma.cc/WAS5-74KJ].

143. Seifter, *supra* note 97, at 1126.

144. *Id.*

145. *Id.* at 1129.

146. *Id.* at 1140 (arguing it is up to the oversight agency to "express its expectations clearly, provide incentives for adherence, and penalize deviations"). Client influence on auditors is likely greater when the industry oversight agency cannot act against the auditors themselves. *Id.*

147. *See* Yanisky-Ravid & Hallisey, *supra* note 55, at 476–77 (describing the purposes of a safe harbor policy to preserve judicial economy, encourage self-regulation, and protect users with better use of AI as a result).

The role of liability may also play an evolving role as AI auditing becomes more commonplace. For example, under securities law, financial auditors can be liable for certain practices of their clients.[148] Likewise, for AI auditors, this responsibility will evolve as new AI laws are enacted, and as existing laws are found to apply to AI.[149]

The rapid and seldom documented development of private AI systems make it crucial that external auditors can act with access and independence. There are currently no guidelines for AI auditing relationships, but oversight of auditors could introduce such standards. The often-cited O'Neil Risk Consulting & Algorithmic Auditing's (ORCAA) audit of HireVue's AI recruiting software illustrates the dangers of unregulated external auditing.[150] HireVue, a firm that uses AI to offer better hiring results, engaged ORCAA to perform an audit of one of its candidate assessment products. The ORCAA audit determined that HireVue met a legal bar for nondiscrimination, but the audit results also noted that the audit was not actually representative of HireVue's models: it turned out ORCAA did not evaluate models more likely to exhibit biased outcomes nor analyze any data involved in making recommendations.[151] Moreover, the full extent of the limited audit's findings was placed behind a nondisclosure agreement, leaving even ORCAA unable to rebut allegations the narrow scope of its work made it misleading.[152]

B. Spurring Development in Auditing and Oversight

AI auditing should not disincentivize innovation or technological development. Decisions such as *Nuvio v. FCC*[153] and legislation such as the Clean Air Act[154] demonstrate that regulatory regimes are not merely added costs for the regulated industry: regulation can promote technological advances in an industry. The approach that AI auditors employ will depend in large part on the technology they deploy. If the oversight agency were to require that auditors adopt a certain practice,

148. *See, e.g.*, Colleen Honigsberg, Shivaram Rajgopal & Suraj Srinivasan, *The Changing Landscape of Auditor Liability*, HBS Working Paper 19-113, 12–14 (Oct. 2018); *see also* McGann v. Ernst Young, 95 F.3d 821 (9th Cir. 1996) (holding that imposing liability on an auditor whose false assertions are reasonably calculated to influence the investing public is consistent with the Securities Exchange Act of 1934).

149. For example, the enactment of individual rights of action in a jurisdiction where a company operates could represent a salient incentive for companies using AI to seek comprehensive audits if it means reducing their liability.

150. *See, e.g.*, Ng, *supra* note 25.

151. *Id.* ("The ORCAA audit examined only HireVue's documentation of one of its job candidate assessments.").

152. Alex Engler, *Independent Auditors are Struggling to Hold AI Companies Accountable*, FAST COMPANY (Jan. 26, 2021), https://www.fastcompany.com/90597594/ai-algorithm-auditing-hirevue [https://perma.cc/LX8J-EXVQ].

153. Nuvio Co. v. FCC, 473 F.3d 302 (D.C. Cir. 2006).

154. Clean Air Act, 42 U.S.C. § 7401-7431 (1970).

it is likely that the technical difficulty of its implementation will not absolve an auditor of its duty to promptly comply. The D.C. Circuit's decision in *Nuvio* has not received due attention for its suggestion that when an agency acts in accordance with statutory authority in issuing an order, the technical feasibility of the request or the economic costs it implicates likely do not excuse the subjects of the order from compliance.[155]

In *Nuvio*, the United States Court of Appeals for the District of Columbia Circuit considered the validity of a Federal Communications Commission ("FCC") order requiring that Voice over Internet Protocol ("VoIP") providers implement a way for users to call 911 and connect to local emergency authorities.[156] The order gave providers 120 days to implement the capability.[157] Due to the technical difficulty of implementing this capability to the satisfaction of the FCC's order within its time limit, Nuvio, an interconnected VoIP provider ("IVP"), asserted both that the order was not feasible and that it would face significant economic cost as a result of the short 120-day window.[158] Nuvio argued the order was thus "arbitrary, capricious, an abuse of discretion, or otherwise not in accordance with law."[159]

Rejecting Nuvio's argument, the D.C. Circuit found the required technology was feasible, and the economic cost did not outweigh the FCC's interests in accordance with its statutory authority and purpose,[160] including its "duty to protect the public" established in its enabling act.[161] As required, the FCC relied on public safety needs in setting the terms of its order, including its self-described "aggressively short" time limit.[162] In addition, the FCC had considered tests that demonstrated the feasibility of the order's requirements.[163] The D.C. Circuit found that different technologies "may reasonably bear different regulatory burdens," so other instances in which the FCC had given

155. Articles citing to *Nuvio* do not touch on this aspect of the decision, focusing instead on its implications for net neutrality.

156. *Nuvio*, 473 F.3d at 303.

157. *Id.*

158. *Id.* at 305–06 (describing concerns with feasibility of meeting the FCC specifications for Nuvio's technology).

159. *Id.* at 305 (quoting 5 U.S.C. § 706(2)(A)).

160. *See id.* at 306–09; Wireless Communication and Public Safety Act of 1999 § 3, 47 U.S.C. § 615 (imposing on the FCC the responsibility to support the States in establishing "end-to-end emergency communications infrastructure"). Nor was the FCC required to hew to the terms of prior related orders. *Nuvio*, 473 F.3d at 306–09.

161. *Nuvio*, 473 F.3d at 307; Communications Act of 1934 § 1, 47 U.S.C. § 151 (establishing that the FCC would regulate with the "purpose of promoting safety of life and property through the use of wire and radio communications").

162. *Nuvio*, 473 F.3d at 308 ("[T]he threat to public safety if we delay further is too great and demands near immediate action." (quoting E911 Requirements for IP-Enabled Service Providers, First Report and Order and Notice of Proposed Rulemaking, 20 F.C.C.R. 10245, 10246 n.1 (2005))).

163. *Id.* at 306–07 (rejecting petitioner's claim that there was "no demonstrated way to overcome the technical and practical obstacles").

more generous time requirements did not necessarily bear on the current order.[164] Thus, it was within the FCC's authority to make a predictive judgment in setting the order's requirements, including the time limit, based on its expertise[165] and the necessity of the change.[166] The order was therefore not arbitrary and capricious.[167]

Writing in concurrence, then-Circuit Judge Kavanaugh went further, finding the order could be justified even if IVPs could *not* feasibly meet the 120-day deadline.[168] Judge Kavanaugh held that the FCC could reasonably protect public safety by banning the sale of a product until adequate 911 functionality could be ensured.[169] From this broader authority, the power to ban sales after 120 days naturally followed.[170] Notably, the majority did not rule out this argument.[171]

Nuvio is instructive for the organization of the agency that will oversee AI auditors. Like the growth of the Internet and introduction of VoIP technology at the time that prompted the FCC's action, the proliferation of AI and its heretofore scant regulation motivates action.[172] "[I]mmediate solution[s]" in the realm of AI might be as desirable as they were for VoIP.[173] The statute establishing an oversight agency to regulate the activities of auditors should therefore explicitly define that the agency has a duty to promote and apply novel technology in AI auditing as a means to stymie and mitigate algorithmic harms.[174] Judge Kavanaugh's principle in *Nuvio* is consistent with the goal of AI audits set out above: AI audits ensure technological development does not undermine existing rights, and Kavanaugh read this authority into the FCC's mandate with respect to the new, transformative technology of VoIP.

---

164. *Id.* (noting that the FCC had granted a greater time horizon for the development of 911 capability for satellite phones).

165. *Id.* (citing PPL Montana, LLC v. Surface Transp. Bd., 437 F.3d 1240, 1247 (D.C. Cir. 2006)) (finding latitude to make such distinctions between regulated technologies is left to agency expertise, accepting FCC's determination that satellite phone technology is sufficiently different from VoIP technology to justify different time requirements).

166. *See id.* at 309; AT & T Corp. v. FCC, 220 F.3d 607, 627 (D.C. Cir. 2000).

167. "Substantial deference" to an agency's predictive judgment within its area of expertise also applies to its determination of the technical scope of an order. *See Nuvio*, 473 F.3d at 309 (finding FCC's failure to "distinguish between the technological obstacles" of different types of VoIP technology did not invalidate its order).

168. *Nuvio*, 473 F.3d at 310–11 (Kavanaugh, J., concurring).

169. *Id.* at 311 (Kavanaugh, J., concurring).

170. *Id.* (Kavanaugh, J., concurring) ("Congress established the FCC in part for the purpose of promoting safety of life and property through the use of wire and radio communications." (internal quotations omitted)).

171. *Id.* at 305 n.5.

172. *See id.* at 303–04.

173. *Id.* at 304.

174. BLUEPRINT, *supra* note 1, at 24–25 (describing instances of algorithmic harms through discrimination).

Due to the nature of AI development, which takes place across industry and academia alike,[175] the oversight agency will likely be able to point to tests and proofs of concept to inform its decisions within the predictive discretion courts have afforded expert agencies.[176] If the oversight agency issues an order consistent with its purpose as defined by its statutory authority, the process of the implementation of the order's requirements likely does not bear on its validity. The implications of the *Nuvio* concurrence are likewise important. The *Nuvio* majority suggests that the oversight agency would be able to push adoption of existing best practices, and the concurrence suggests the oversight agency may accelerate progress by pushing adoption of cutting-edge technology. These developments will be crucial to the efficacy of auditing as a regulatory approach to AI.[177]

The "technology forcing" policy once pursued by the Environmental Protection Agency (EPA) is also instructive. The Clean Air Act of 1970 set aggressive emission limits that would kick in within five years of its passage even though they were technologically unachievable at the time of enactment.[178] In the AI setting it is conceivable that certain requirements may need to be set in advance of the advent of the implicated technology, such as anticipating the impact of quantum computing.[179] In the meantime, the *Nuvio* decision informs how an oversight agency for AI can be established to advance auditing technology in its oversight capacity.

The lawfulness of this approach depends now more than ever on how Congress grants authority to the oversight agency. The Supreme Court has eliminated the deference formerly granted to agencies to interpret the scope of their statutory authority, charging courts to "exercise independent judgment in determining the meaning of statutory provisions."[180] Thus, not only must the oversight agency be granted targeted authority to explore and recommend industry standards, the passage of accompanying statutory AI regulation may also be necessary to bolster agency actions against a skeptical court.

---

175. Kinjal Basu, *Our Approach to Building Transparent and Explainable AI Systems*, LinkedIn Engineering (Oct. 7, 2021), https://engineering.linkedin.com/blog/2021/transparent-and-explainable-AI-systems [https://perma.cc/8PTY-NYAE] (presenting implementation of explainable algorithms in LinkedIn engineering group).

176. *E.g.*, Gilpin et al., *supra* note 80, at 3–5 (describing a variety of methods for producing explanations for deep neural networks and building explanation-producing systems).

177. *Cf.* Desai & Kroll, *supra* note 94, at 49 (listing "technical infeasibility" and "complexity" as barriers to evaluating algorithms).

178. *Cf.* David Gerard & Lester B. Lave, *Implementing Technology-Forcing Policies: The 1970 Clean Air Act Amendments and the Introduction of Advanced Automotive Emissions Controls in the United States*, 72 TECH. FORECASTING & SOC. CHANGE 761, 775–76 (2005).

179. Vivek Wadhwa & Mauritz Kop, *Why Quantum Computing Is Even More Dangerous Than Artificial Intelligence*, FOREIGN POL'Y (Aug. 21, 2022, 6:00 AM), https://foreignpolicy.com/2022/08/21/quantum-computing-artificial-intelligence-ai-technology-regulation/ [https://perma.cc/52CR-ZKGU].

180. Loper Bright Enters. v. Raimondo, 603 U.S. 369, 394, 406 (2024).

IV. EMERGING ISSUES

AI auditing regulation prompts additional legal questions, many of which are still in nascent stages in legal scholarship. In particular, we have identified the relationship between auditing, trade secrets, and speech as burgeoning areas of inquiry that will only continue to garner interest and debate.

A. Trade Secrets

Oversight of auditors must be established in a manner such that trade secret claims do not limit the oversight agency's purview to access, review, and release information pertaining to AI audits. Trade secret claims from both audited entities and auditors will foreseeably accompany the adoption of AI auditing.[181] An audited entity may be concerned that an audit and oversight will reveal its technology, and an auditor may be concerned that its audit processes, especially components that themselves use AI,[182] could be revealed to other auditors.[183] These concerns could be minimized by timing the release of audits to minimize trade secret-related impact of disclosure[184] or creating a licensing scheme for auditors.[185] Technological solutions, such as tools that utilize "zero-knowledge proofs" may also mitigate the release of

---

181. *See* Lilian Edwards & Michael Veale, Slave to the Algorithm? *Why A "Right to an Explanation" Is Probably Not the Remedy You Are Looking For*, 16 DUKE L. & TECH. REV. 18, 70 (2017); Camilla A. Hrdy, *The Value in Secrecy*, 91 FORDHAM L. REV. 557, 564 (2022) ("Trade secrets stand in the way of the disclosure of information of high public interest.").

182. For instance, natural language processing techniques for digesting privacy policies. *See* Rahmadi Trimananda, Hieu Le, Hao Cui, Janice Tran Ho, Anastasia Shuba & Athina Markopoulou, OVRseen: Auditing Network Traffic and Privacy Policies in Oculus VR (Nov. 2021) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/2106.05407.pdf [https://perma.cc/75RM-KJVF] (using natural language processing models to parse Oculus VR's privacy policy, ultimately finding that 70% of the company's data flows were not properly disclosed).

183. *See* Coglianese & Lampmann, *supra* note 26, at 184–85. Making things worse, AI technologies are largely un-patented. *See, e.g.*, Eric Rosenbaum, *A Tesla Self-Driving Blind Spot that Few Are Focusing On*, CNBC (Feb. 8, 2018, 9:12 AM EST), https://www.cnbc.com/2018/02/08/a-tesla-self-driving-blind-spot-that-few-are-focusing-on.html [https://perma.cc/6PL8-J866] (reporting that Tesla has not secured any patents in recent years related to self-driving, with speculation that it might be relying on trade secret protections instead); Daisuke Wakabayashi, *Secrets of Knowledge? Uber-Waymo Trial Tests Silicon Valley Culture*, N.Y. TIMES (Jan. 30, 2018), https://www.nytimes.com/2018/01/30/technology/waymo-uber-lawsuit.html [https://perma.cc/423S-62VZ] (detailing lawsuit in which Waymo accused Uber of misappropriating its trade secrets).

184. Onora O'Neill, *Transparency and the Ethics of Communication*, in TRANSPARENCY: THE KEY TO BETTER GOVERNANCE? 75–90 (C. Hood & D. A. Heald, eds., Oxford Univ. Press 2006).

185. Some level of disclosure could be required in exchange for a license to operate as an auditor, qualified to perform required audits. Jeanne C. Fromer, *Machines as the New Oompa-Loompas: Trade Secrecy, the Cloud, Machine Learning, and Automation*, 94 N.Y.U. L. REV. 706, 733–34 (2019).

auditor and auditee trade secrets.[186] Auditors may also agree to maintain secrets via non-disclosure agreements.[187]

Professor Amy Kapczynski advances an argument rooted in the history of trade secrets to reconceptualize trade secret claims.[188] Maintaining trade secrets is not a sacrosanct right untouchable by legislatures and administrative agencies. Instead, Kapczynski's view recognizes that an enterprise's right to protect its trade secrets is a right extended by the state and subject to the "right of the state, in the exercise of its police power and in promotion of fair dealing."[189] Even disclosure of "secret formulas" would not necessarily unconstitutionally deprive a business of a property right in its secrets.[190] A mandate to ensure public awareness of the "nature of the product" may encompass core secrets such as "formulas" for products.[191] Kapczynski traces these principles from origins in the early twentieth century cases of *National Fertilizer v. Bradley* and *Corn Products Refining Co. v. Eddy* through recent decisions of state supreme courts, in which administrative processes facilitated disclosures to inform the public of information that otherwise would have been protected as a trade secret.[192]

While Kapczynski's view is not one that can or will be adopted immediately, it nevertheless carries relevant implications for Congress's design of an oversight agency's mission and for its success.[193]

---

186. Kroll et al., *supra* note 23, at 668.

187. Ajunwa, *supra* note 26, at 652; *see* Peter S. Menell, *Tailoring a Public Policy Exception to Trade Secret Protection*, 105 CALIF. L. REV. 1, 16–17 (2017).

188. Amy Kapczynski, *The Public History of Trade Secrets*, 55 U.C. DAVIS L. REV. 1367, 1436–37 (2022).

189. *Id.* at 1432 (quoting Corn Prods. Refin. Co. v. Eddy, 249 U.S. 427, 431–32 (1919)).

190. In *National Fertilizer Association v. Bradley*, a statute requiring fertilizer producers to reproduce on each package of fertilizer a chemical analysis of its contents was upheld. 301 U.S. 178, 181–82 (1937); *id.* at 179 n. 1 (reproducing statutory requirements, including disclosures of the form, "[a]vailable phosphoric acid . . . per cent. Ammonia equivalent of nitrogen . . . per cent. Potash soluble in water … per cent.").

191. *See* Corn Prods. Refin. Co. v. Eddy, 249 U.S. 427, 429–32 (1919) (finding no "constitutional right to sell goods without giving to the purchaser fair information of what it is that is being sold"). Moreover, in neither *Corn Products* nor *National Fertilizer* did the Court rely on "risk to public health, nor any likelihood of fraud," nor a balancing of public and private harms. Kapczynski, *supra* note 188, at 1379 (suggesting that disclosure requirements comparable to those in *National Fertilizer* are effectively subject to rational basis review); *see Nat'l Fertilizer*, 301 U.S. at 178.

192. *Id.* at 1437; *see* Powder River Basin Res. Council v. Wyo. Oil & Gas Conservation Comm'n, 320 P.3d 222, 235 (Wyo. 2014) (granting public request for release of fracking chemicals over trade secret objections); Lyft, Inc. v. City of Seattle, 418 P.3d 102, 105 (Wash. 2018) (finding that while likely a trade secret, Lyft's zip code information could be released publicly unless doing so would be " clearly not in the public interest" and would "pose[] substantial and irreparable harm"); *see also* Georgia v. Public.Resource.org, Inc., 140 S. Ct. 1498 (2020) (suggesting the applicability of "government edicts doctrine" to AI systems). *But see* Philip Morris Inc. v. Reilly, 312 F.3d 24, 28 (1st Cir. 2002) (*en banc*) (plurality opinion) (enjoining a proposed law that employed too low of a public interest standard for the disclosure of cigarette ingredients).

193. Congress could begin by not granting protection in advance. Kapczynski, *supra* note 188, at 1424. Congress should also clarify expectations surrounding the treatment of audits'

For instance, the animating concepts in *National Fertilizer* and *Corn Products* under which the corresponding laws were upheld can be extended to AI systems and the practice of AI auditing. AI systems are reducible to "formulas," whose components determine their performance — more complex, to be sure, but not unlike how the formulas for corn syrup and fertilizer define the products' properties and uses.[194] An auditor would need access to the components of the formula to perform an audit and the auditor's methods might need to be revealed — if not to the public, to the government — to enhance accountability for both parties to an audit.[195] This sharing of procedures and results of AI audits is meant to inform the public of the "nature" of AI systems, including elements and analyses of their "formulas." Such requirements would therefore be an exercise of the government's police power over markets, both for AI and for AI auditing.[196] Moreover, if advancing the capabilities of the auditing industry is considered an objective of oversight, the disclosure and spread of details related to any one auditor's technology could be justified as a means of advancing that objective.[197]

More advanced AI systems raise special concerns when trade secrets begin to intertwine with national security: the internal workings of certain AI applications that would be considered trade secrets in a commercial sense might also be national security secrets that bear on security and national competitiveness.[198] For these applications, requiring certain types of audit-related reporting to government regulators could create opportunities for national adversaries, and not just industry competitors, to access sensitive information.[199] This leakage would be

---

contents, thereby reducing the extent to which a regulated entity could claim reliance on promises and define the countervailing public interest the agency represents. *See id.* at 1438–39 (citing Ruckelshaus v. Monsanto Co., 467 U.S. 986, 1007–08 (1984)) (arguing *Monsanto* casts doubt on the enforceability of Congressional promises and clears the way for competitive uses of disclosed information). On the countervailing side, it is doubtful a company would openly argue that a harmful effect of its AI systems has some competitive value to be protected. *Id.* at 1437.

194. *Cf.* Viacom Int'l Inc. v. YouTube Inc., 253 F.R.D. 256, 260 (S.D.N.Y. 2008) (accepting that a new algorithm's source code was a trade secret, but this was not contested by the parties and the disclosure was sought by another corporation for its own use, not an administrative agency or for the purpose of public access).

195. *Id.* at 1437.

196. Kapczynski explicitly suggests as much, *supra* note 188, at 1436–37. In light of this, an argument that disclosing certain information is technically infeasible might be the stronger argument by a reluctant regulated entity, but the extension of the holding of *Nuvio* limits this argument as well. *See* 473 F.3d at 303–05.

197. As Kapczynski reminds us, the government is "*always* behind the structure of competition and therefore [may] rearrange it." *Id.* at 1441 (emphasis in original). The growth of the auditing industry in our proposal will be driven by government action to encourage and eventually require auditing.

198. *See generally* Greg Allen & Taniel Chan, *Artificial Intelligence and National Security*, BELFER CTR., July 2017, at 58–67.

199. *Cf.* Mathew Bultman, *Hedge Funds Warn SEC Cyber Lapses Risk Exposing Trading Secrets*, BLOOMBERG L. (Feb. 16, 2024, 11:43 AM EST), https://news.bloomberglaw.com/securities-law/hedge-funds-fear-sec-cyber-lapses-risk-exposing-trading-secrets

contrary to the government's own overarching objectives orthogonal to AI regulation. For such advanced and sensitive AI systems that could carry national security implications, a robust infrastructure for securely reporting and retaining trade secrets will be an important component to the auditing approach to AI regulation.

B. Auditing and Speech

The First Amendment has multiple implications for the oversight of auditing. With the Roberts Court's expanding view of what constitutes speech,[200] the invasiveness of audits themselves and the changes they may dictate could be constitutionally problematic without proper design.[201] On one extreme, consider a hypothetical regulatory regime targeted at political writing that requires publishers to engage auditors to assess articles' adherence to the ideologies the authors claim to represent. This sort of content-based regulation of quintessential political speech clearly violates the First Amendment.[202] On an opposite extreme, consider quantitative models for pricing financial derivatives, which are often the subject of academic study but nevertheless subject to intensive regulatory regimes.[203] An AI auditing regime falls somewhere along the continuum between these two extremes. An audit of an AI system implicates some notion of academic freedom,[204] but perhaps such auditing deals only with "commercial speech," which receives less protection than other forms of protected speech, like political speech.[205]

[https://perma.cc/L4CR-4SE8] (describing hedge funds' concerns over leaks of trading strategies once they are reported to the SEC in accordance with a new program).

200. *See, e.g.*, Citizens United v. FEC, 558 U.S. 310 (2010).

201. For example, *data* could be speech. Jane Bambauer, *Is Data Speech?*, 66 STAN. L. REV. 57, 114–116 (2014) (discussing implications for regulators if data is speech). While it may have intensified, the trend is not new. *See* Turner Broadcasting System, Inc. v. FCC, 512 U.S. 622, 643 (1994) ("As a general rule, laws that by their terms distinguish favored speech from disfavored speech on the basis of the ideas or views expressed are content based."); Buckley v. Valeo, 424 U.S. 1, 19 (1976); United States v. Playboy Ent. Grp., Inc., 529 U.S. 803 (2000); N.Y. Times v. Sullivan, 376 U.S. 254 (1964).

202. *Cf.* W. Va. State Bd. of Educ. v. Barnette, 319 U.S. 624, 642 (1943) (rejecting any exception to protection of speech concerning "politics, nationalism, religion, or other matters of opinion"); *Sullivan*, 376 U.S. at 276 (establishing broad prohibition on regulations that discriminate on the basis of speech content).

203. *See, e.g.*, 17 C.F.R. Part 43 (establishing real-time public reporting requirements of transaction and pricing data for "swaps," a class of financial derivatives).

204. *Cf.* Sweezy v. N.H., 354 U.S. 234 (1957) (outlining "four essential freedoms" of a university: "to determine for itself on academic grounds *who* may teach, *what* may be taught, *how* it shall be taught, and *who* may be admitted to study" (emphasis added)). There are layers of academic freedom, reflecting both outside influence of an academic institution, as in *Sweezy*, and influence from within. *See* Edwards v. Cal. Univ. of Pa., 156 F.3d 488 (3d Cir. 1998), *cert. denied*, 525 U.S. 1143 (1999) (reaffirming a principle that faculty as a whole, even if not necessarily individual professors, have the right to determine curricular foci, not the administration).

205. *See* Va. St. Bd. of Pharmacy v. Va. Citizens Consumer Council, Inc., 425 U.S. 748, 771 (1976).

Even if auditing could be limited to commercial speech, the line of business the entity is engaged in could still carry different implications for assertions of protected speech. In *Sorrell v. IMS Health*,[206] a Vermont law prohibiting pharmacies from selling prescription data and banning pharmaceutical manufacturers from using such data for marketing was found to violate the First Amendment as it "burden[ed] disfavored speech by disfavored speakers."[207] Auditing of certain uses of AI will implicate less "speech" content than others. For instance, audits of data intermediaries, who may use AI to aggregate, process, and eventually sell data, arguably implicate less "speech" than the actual use of purchased data for marketing purposes.[208] Avoiding forbidden content- and speaker-based prohibitions could leave more generalized approaches on the table.[209] However, the adaptability of auditing — i.e., how it can be effectively tailored application by application — is meant to be a benefit of the approach. A broad reading of *IMS Health* could conceivably jeopardize this strength, forcing a choice between adopting blanket approaches that could unintentionally hamper productive uses of AI, or skipping regulation altogether, which could result in harmful practices going unchecked.[210]

Disclosure requirements accompanying oversight of auditors must also withstand challenges on the basis of compelled speech. Compelled speech claims come part and parcel with disclosures of trade secrets.[211] Nevertheless, without transparency of auditing methods or results of third-party or internal audits, any assurance of fairness or legal compliance in AI applications will come from the company itself (that is, if internal audits are even conducted at all).[212] A structured disclosure

206. 564 U.S. 552 (2011).

207. *Id.* at 564.

208. "A data intermediary serves as a mediator between those who wish to make their data available, and those who seek to leverage that data. The intermediary works to govern the data in specific ways, and provides some degree of confidence regarding how the data will be used." Heleen Janssen & Jatinder Singh, *Data Intermediary*, 11 INTERNET POL'Y REV. 1, 2 (2022). Indeed, auditing of data intermediaries may be more readily likened to the kind of time, place, and manner restrictions the Court has upheld. *See, e.g.*, Police Dept. of Chicago v. Mosley, 408 U.S. 92 (1972). Data intermediaries were among the plaintiffs in *IMS Health*.

209. For example, requiring sufficient anonymization of data in general could be a way to regulate the specific use of location data. *Cf.* Erie v. Pap's A. M., 529 U.S. 277 (2000) (upholding against First Amendment challenge city ordinance banning public nudity in general meant to target establishments with nude erotic dancing).

210. The Court did not go so far as to state that data is "speech," however. Bambauer, *supra* note 201, at 71 (citing Trans Union LLC v. FTC, 536 U.S. 915, 916 (2002) (Kennedy, J., dissenting from denial of certiorari)); *see IMS Health*, 564 U.S. at 564.

211. *See* Kapczynski, *supra* note 188, at 1436–37.

212. Eric Rosenbaum, *Silicon Valley Is Stumped: Even A.I. Cannot Always Remove Bias from Hiring*, CNBC (May 30, 2018, 9:43 AM EDT), https://www.cnbc.com/2018/05/30/silicon-valley-is-stumped-even-a-i-cannot-remove-bias-from-hiring.html [https://perma.cc/5LXL-MNU8] ("The public will have little knowledge as to whether or not the firm really is making biased decisions if it's only the firm itself that has access to its decision-making algorithms to test them for discriminatory outcomes.").

format could be fashioned to resist compelled speech claims. Mitchell et al. describe "model cards" to accompany AI systems, complete with enumerated requirements with standardized metrics.[213] AI labeling of this sort could be designed in line with the disclosures upheld in *National Fertilizer*,[214] and likely would not implicate compelled speech of the kind the Supreme Court feared in *Pacific Gas & Electric Co. v. Public Utilities Commission*.[215] As the result of a comprehensive and rigorous investigation, an adverse auditing determination likely would not be "biased" or "controversial" in the context of compelled speech.[216] Audit results may be the subject of factual disagreement, but are likely distinguishable from "controversial" content, such as a poster "favoring unionization" based on editorial content choices.[217] Audit results disclosed on a digestible "model card" could be analogized to country-of-origin disclosure requirements for products sold in the United States: such results are simply facts relating to the manufacturing history of the product, and do not amount to compelled speech.[218]

V. CONCLUSION

Enacting effective regulation of AI through auditing will face significant challenges. Some emerge from the nature of AI, which include its rapid and seldom-documented development as well as the complexity of AI systems.[219] Other challenges are more familiar, such as the risk that auditors will fail to gather sufficient evidence or confront

---

213. *See* Margaret Mitchell et al., *Model Cards for Model Reporting*, 2019 CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 220, 223 (2019) (presenting example "model card" with suggested sections and prompts). Congress would have "substantial" leeway in determining what information must be disclosed, but it may not go so far as compelling speech that is contrary to the views of the affected entity. Pac. Gas & Elec. Co. v. Pub. Utils. Comm'n, 475 U.S. 1, 15–16 n. 12 (1986) (plurality opinion) (citing *Zauderer v. Off. Disciplinary Couns.*, 471 U.S. 626, 651 (1985)).

214. Kapczynski, *supra* note 188, at 1436–37 (citing *Nat'l Fertilizer*, 301 U.S. at 182); *see also* Zauderer v. Off. Disciplinary Couns., 471 U.S. 626, 628 (1985) (adapting notion from First Amendment treatment of commercial speech to allow the disclosure of "purely factual and uncontroversial information" about the product or service).

215. *Cf.* 475 U.S. at 15 ("Should TURN choose, for example, to urge appellant's customers to vote for a particular slate of legislative candidates, or to argue in favor of legislation that could seriously affect the utility business, appellant may be forced either to appear to agree with TURN's views or to respond."). Labeling based on an auditor's findings does not compel entities deploying AI systems to spread the messages of third parties. *Id.* at 15–16 n.12.

216. Am. Meat Inst. v. U.S. Dep't of Agric., 760 F.3d 18, 27 (D.C. Cir. 2014).

217. *Cf. id.* ("We also do not understand country-of-origin labeling to be controversial in the sense that it communicates a message that is controversial for some reason other than dispute about *simple factual accuracy*." (emphasis added)); *Pac. Gas & Elec. Co.*, 475 U.S. at 15.

218. *Cf. id.*

219. Siva Vaidhyanathan, *There's No Such Thing as a Tech Expert Anymore*, WIRED (Aug. 4, 2020, 8:00 AM), https://www.wired.com/story/theres-no-such-thing-as-a-tech-expert-anymore/ [https://perma.cc/F3PP-3UG7].

efforts by audited entities to limit access.[220] The risk of auditor capture likewise exists. The government must be able to intervene to realign incentives and offer protection to auditors and their clients alike. Through its capacity to compel the development of technology, government oversight makes AI accountability through auditing possible.

The potential for a virtuous cycle of AI development sets auditing apart as a regulatory approach to AI. The most sensitive applications or those in areas where existing regulations are insufficient can be targeted first with mandated audits. As standards solidify, audit mandates for additional applications of AI can be phased in. Then, as public recognition of the value of auditing increases alongside expanding audit mandates, a virtuous cycle can emerge, reining in the dangers of AI while advancing the technology in a way that is consistent with its positive potential.

220. Engler, *supra* note 34.

APPENDIX: AI EXAM PROCTORING CASE STUDY

There are several prominent providers of AI-driven exam proctoring software today.[221] A number of sensitive issues are implicated by proctoring software, leaving ample work for auditors.[222] The AI component of the software is commonly meant to identify behavior consistent with cheating, for which data is collected and analyzed, including audio/video, keystroke data, application activity, and personal information.[223] Aside from surveillance concerns, there is potential for discrimination in algorithmic cheating accusations:[224] facial recognition bias aside, systems may flag disability accommodations or underlying conditions as signs of cheating.[225] AI proctoring systems impact millions, and their users — students — are forced into these systems; unlike voluntary users of social media, for example, students often have no choice to opt out of exams and automated proctoring systems. As a result, lack of transparency and the risk of bias are all the more pertinent.

The active data collection methods of the proctoring systems are a natural place for the data component of the audit to focus. In particular, auditors may seek to determine which data is actually needed, based on how the audited entity defines its service.[226] AI proctoring has a readily comparable analog alternative: human proctoring or take-home exams that universities have long administered. Therefore, the necessity of certain data can be directly tested against these base cases.[227] In addition, race and disability status are two categories that have been the

---

221. Including Respondus (https://web.respondus.com, [https://perma.cc/S24B-6JDA]), ProctorU (https://www.proctoru.com, [https://perma.cc/D3VE-Y9AL]), Proctorio (https://proctorio.com, [https://perma.cc/8BB6-32W3]), Examity (https://www.examity.com, [https://perma.cc/P6GU-9M5Z]), and Honorlock (https://honorlock.com, [https://perma.cc/9YPE-W5GV]).

222. Complaint and Request for Investigation, Injunction, and Other Relief Submitted by The Electronic Privacy Information Center (EPIC), In the Matter of Online Test Proctoring Companies Respondus, Inc.; ProctorU, Inc.; Proctorio, Inc.; Examity, Inc., and Honorlock, Inc. (Dec. 9, 2020), available at https://epic.org/documents/in-re-online-test-proctoring-companies/ [https://perma.cc/MAK9-E6PA] [hereinafter EPIC Complaint].

223. *Id.*

224. Anushka Patil & Jonah Engel Bromwich, *How It Feels When Software Watches You Take Tests*, N.Y. TIMES (Sep. 29, 2020), https://www.nytimes.com/2020/09/29/style/testing-schools-proctorio.html [https://perma.cc/K7UB-6JFH].

225. Report on Concerns Regarding Online Administration of Bar Exams, NAT'L DISABLED L. STUDENTS ASS'N (July 29, 2020), available at https://ndlsa.org/wp-content/uploads/2020/08/NDLSA_Online-Exam-Concerns-Report1.pdf [https://perma.cc/XW5P-XLQK].

226. EPIC Complaint, *supra* note 222, at 6.

227. *See* George Watson & James Sottile, *Cheating in the Digital Age: Do Students Cheat More in Online Courses?*, ONLINE J. DISTANCE LEARNING ADMIN. (2010), https://www.westga.edu/~distance/ojdla/spring131/watson131.html [https://perma.cc/8B5N-4AXR] (presenting results that tend to show that students cheat more during in-person exams).

target of concern in this context.[228] Auditors should also examine privacy issues implicated by data collection practices, which is especially relevant here given documented privacy lapses and misleading practices by proctoring software companies.[229]

For the model component of audits, auditors could seek to demonstrate (or disprove) the reliability, accuracy, and the validity of the audited entities' AI systems. For example, keystroke analysis, which is used in proctoring software, has been shown to be inaccurate for predicting cheating.[230] Nevertheless, companies represent their product as "elegant, functional, powerful" while disclaiming any liability for "accuracy, content, completeness, legality, reliability, operability or availability of information or data."[231] An auditor would be able to assess accuracy and certify claims made by the audited entities.

Though this exam proctoring software lacks transparency, this is an instance where transparency alone could be counterproductive as it might enable gaming of the AI systems by those trying to cheat by designing illegitimate conduct to conform to or to spoof disclosed standards.[232] Transparency here should therefore not be confused with access: auditors must have *access* to systems so that they can perform meaningful analyses. Meanwhile, explainability could allow for more meaningful audits, and, ultimately accountability, including directly to concerned students through plain language explanations for why test-time activity was flagged as suspicious.[233]

The deployment of the proctoring systems may implicate multiple statutes and standards. For example, the OECD Principles on Artificial Intelligence may be "established public policies" within the meaning of the FTC Act.[234] Preventing users from knowing the basis of a risk assessment,[235] or failing to incorporate accountability mechanisms for

---

228. *See, e.g.*, EPIC Complaint, *supra* note 222, at 8, 11.

229. *Id.* at 4 ("The rapid growth of online test proctoring has all but forced many students to trade away their privacy rights in order to meet their academic obligations."); *id.* at 8.

230. EPIC Complaint, *supra* note 222, at 7 (describing ProctorU's use of keystroke analysis, for example); *see* Shea Swauger, *Remote Testing Monitored by AI Is Failing the Students Forced to Undergo It*, NBC News (Nov. 7, 2020, 4:30 AM EST), https://www.nbcnews.com/think/opinion/remote-testing-monitored-ai-failing-students-forced-undergo-it-ncna1246769 [https://perma.cc/2MLH-C89S] (presenting overview of keystroke analysis).

231. EPIC Complaint, *supra* note 222, at 12–13.

232. *See* de Laat, *supra* note 86, at 535–36 (introducing "gaming" as a "perverse effect" of transparency).

233. *Id.* at 20–21.

234. EPIC Complaint, *supra* note 222, at 16–17 (citing 15 U.S.C. § 45(n)) ("In determining whether an act or practice is unfair, the Commission may consider established public policies as evidence to be considered with all other evidence."); *see also* Fiona Alexander, *U.S. Joins with OECD in Adopting Global AI Principles*, NTIA BLOG (May 22, 2019), https://www.ntia.gov/blog/2019/us-joins-oecd-adopting-global-ai-principles [https://perma.cc/RHG5-WQ4S].

235. *Recommendation of the Council on Artificial Intelligence*, OECD (May 21, 2019), http://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 [https://perma.cc/

such a lapse,[236] could therefore constitute "unfair or deceptive" practices.[237] Auditors should thus investigate whether creators of proctoring software are prepared to make required accommodations for students with disabilities. The audited entity's marketing of its product may also be relevant, if an audit tends to prove it misrepresented its product.[238] The exam proctoring setting emphasizes the importance of the deployment prong of AI audits; even perfectly "fair" systems might be inappropriate to the setting.[239]

Oversight of auditors in this context would involve an inquiry into the tests the auditors ran to determine the accuracy and reliability of the algorithms. Explainability is important here, and oversight of auditors should verify if auditors have been able to reconcile (or not) the claims of the audited entities with the actual performance of audited systems.[240] Indeed, there have already been claims of deceptive practices against these exam proctoring companies.[241] Because these AI systems impact students who do not necessarily seek them out and yet could have a significant effect on a student's life, accountability is doubly important.[242]

BDB7-2HFR] (OECD AI Principle on Transparency and Explainability); *see also Universal Guidelines for Artificial Intelligence*, PUB. VOICE (Oct. 23, 2018), https://archive.epic.org/international/AIGuidleinesDRAFT20180910.pdf [https://perma.cc/7R3B-8BBY].

236. *Id.* (OECD AI Principle on Accountability). The OECD principles also cover "foreseeable" use cases of a system, so in this instance, auditors should assess the appropriateness of the system as deployed and in foreseeable circumstances. *Id.*

237. 15 U.S.C. § 45(a)(1).

238. Thompson Med. Co., Inc. v. FTC, 791 F.2d 189, 193 (D.C. Cir. 1986) (a company engages in a prohibited deceptive trade practice when it makes a representation to consumers yet for which it lacks a "reasonable basis" to support the claims made); *see, e.g.*, EPIC Complaint, *supra* note 222, at 21–23.

239. For example, surveillance. *Commercial Surveillance and Data Security Rulemaking*, FED. TRADE COMM'N (Aug. 11, 2022), https://www.ftc.gov/legal-library/browse/federal-register-notices/commercial-surveillance-data-security-rulemaking [https://perma.cc/3THD-6VZW].

240. *See* EPIC Complaint, *supra* note 222, at 10–11 (establishing opaque aspects of AI systems employed by proctoring companies).

241. *Id.* at 20.

242. *See* Patil & Bromwich, *supra* note 224.