Algorithmic Accountability in the Administrative State*

David Freeman Engstrom† & Daniel E. Ho††

How will artificial intelligence (AI) transform government? Stemming from a major study commissioned by the Administrative Conference of the United States (ACUS), we highlight the promise and trajectory of algorithmic tools used by federal agencies to perform the work of governance. Moving past the abstract mappings of transparency measures and regulatory mechanisms that pervade the current algorithmic accountability literature, our analysis centers around a detailed technical account of a pair of current applications that exemplify AI's move to the center of the redistributive and coercive power of the state: the Social Security Administration's use of AI tools to adjudicate disability benefits cases and the Securities and Exchange Commission's use of AI tools to target enforcement efforts under federal securities law. We argue that the next generation of work will need to push past a narrow focus on constitutional law and instead engage with the broader terrain of administrative law, which is far more likely to modulate use of algorithmic governance tools going forward. We demonstrate the shortcomings of conventional ex ante and ex post review under current administrative law doctrines and then consider how those doctrines might adapt in response.

* We thank the extraordinary law and computer science students in the policy practicum on Administering by Algorithm: Artificial Intelligence in the Regulatory State (Sandhini Agarwal, Matthew Agnew, Clint Akarmann, Nitisha Baronia, Cristina Ceballos, Shushman Choudhury, Alex Duran, Michael Fischer, Peter Henderson, David Hoyt, Caroline Jo, Sunny Kang, Urvashi Khandelwal, Minae Kwon, Joseph Levy, Larry Liu, Derin McLeod, Ben Morris, Ashley Pilipiszyn, James Rathmell, Patrick Reimherr, Geet Sethi, Stephen Tang, Nate Tisa, Florian Tramer, Emma Wang, Chase Weidnerm, and Alex Yu) for outstanding research, Kinbert Chou, Maddie Levin, and Coby Simler for research assistance, our practicum co-instructors Cathy Sharkey and Tino Cuéllar, and Scott Bauguess, Marco Enriquez, Kurt Glaze, Alex Jadin, Robyn Konkel, Dan Koster, Gerald Ray, David Saltiel, Liza Starr, Jonathan Vogan, Matt Wiener, and participants at the Roundtable Discussion on the Use of Artificial Intelligence in the Federal Administrative Process at N.Y.U. School of Law, the University of Texas Faculty Workshop, the C. Boyden Gray Center for the Study of the Administrative State's Research Roundtable on Technology, Innovation, and Regulation, the AALS meeting panel on Algorithmic Knowledge: Law, Science, and Democracy, the ABA Administrative Law Conference on Artificial Intelligence in Regulatory Enforcement, the AI in Government panel at Stanford's AI Ethics, Policy, and Governance conference, the AI4Law Workshop at Oxford University, and the Workshop on Artificial Intelligence and the Administrative State, International Conference on Artificial Intelligence and Law, for helpful conversations.

† Professor of Law, Associate Dean, Bernard D. Bergreen Faculty Scholar; Stanford Law School, 559 Nathan Abbott Way, Stanford CA 94305; Tel: 650-721-5859; Email: dfengstrom@law.stanford.edu.

†† William Benjamin Scott and Luna M. Scott Professor of Law; Professor of Political Science; Senior Fellow at Stanford Institute for Economic Policy Research; Stanford University, 559 Nathan Abbott Way, Stanford, CA 94305; Tel: 650-723-9560; Email: dho@law.stanford.edu.

Finally, we ask how else to build a sensible accountability structure around public sector use of algorithmic governance tools while maintaining incentives and opportunities for salutary innovation. Reviewing some commonly offered solutions, we propose a further and novel approach to oversight centered on prospective benchmarking. By requiring agencies to reserve a random set of cases for manual decision making, benchmarking offers a concrete and accessible test of the validity and legality of machine outputs, enabling agencies, courts, and the public to learn about, validate, and correct errors in algorithmic decision making.

Introduction

In 2018, IBM published a white paper touting artificial intelligence (AI) as a way to "reinvent[] the business of government."[1] With IBM's help, the paper proclaimed, governments will undergo a digital transformation, become more client-oriented, and "recognize each citizen as a whole individual, with a personalized set of needs, interests, capabilities, and vulnerabilities."[2] The new suite of AI-based tools, it continued, will "[i]mprove the decision making of civil servants for maximum impact,"[3] empowering agency administrators to "apply digital insights to predict and intervene for better citizen outcomes."[4] "[D]igital reinvention," came a final claim, will not just yield a government that is more responsive and effective in performing its duties and meeting citizen needs,[5] but also one that operates with "[g]reater transparency."[6]

The tenor of IBM's claims should have a familiar ring. Twenty-five years earlier, President Clinton made comparable promises to "reinvent . . . [g]overnment."[7] Speaking near a Sunnyvale community center in Silicon Valley, Clinton and Vice President Gore lauded the city as a model for reinvention.[8] Sunnyvale captured data on thousands of measures, developed targets for each governmental unit, and instituted performance-based pay and budgeting. As described by Osborne and Gaebler in their bestselling book *Reinventing Government*, Sunnyvale was "the performance leader,"[9] transforming government into a lean, responsive, customer-oriented business.[10] Per the New York Times, "If the Clinton Administration has its way, all of America will operate like this highly computerized, relentlessly self-evaluating city in the heart of Silicon Valley."[11] The new digital toolkit would also enable government to "empower citizens to shape the marketplace according to their

1. IBM, DIGITAL TRANSFORMATION: REINVENTING THE BUSINESS OF GOVERNMENT (2018).

2. *Id*. at 13.

3. *Id*. at 5.

4. *Id.* at 13.

5. *Id.* at 5.

6. *Id*. at 7.

7. President Bill Clinton, Remarks by President Clinton Announcing the Initiative to Streamline Government (Mar. 3, 1993), https://govinfo.library.unt.edu/npr/library/speeches/030393.html [https://perma.cc/C6SR-MTX3].

8. Paul Richter, *Clinton, Gore Hail Sunnyvale's City Efficiency*, L.A. TIMES (Sept. 11, 1993), https://www.latimes.com/archives/la-xpm-1993-09-11-mn-34070-story.html [https://perma.cc/K6Z2-96GP].

9. DAVID OSBORNE & TED GAEBLER, REINVENTING GOVERNMENT: HOW THE ENTREPRENEURIAL SPIRIT IS TRANSFORMING THE PUBLIC SECTOR 142 (1992).

10. *See id.* at 145.

11. Seth Mydans, *Where Trouble Is Rare and Governing Is Easy*, N.Y. TIMES (Sept. 10, 1993), https://www.nytimes.com/1993/09/10/us/where-trouble-is-rare-and-governing-is-easy.html [https://perma.cc/9GAV-4PXP].

own needs and values"[12] and, as Gore put it, "earn back the trust of Americans."[13]

Yet Sunnyvale floundered. When its performance index dropped, it changed the weights. When new weights did not fix matters, it abandoned the overall measure. By 1999, employees quit in droves and accused municipal leadership of mismanagement.[14] So went the beacon of public sector performance measurement and a newly responsive approach to government. When agency administrators can define and game performance measures and lack clear baselines for judging gains from technology adoptions, new systems can erode accountability and foil oversight rather than promote regulatory goals.[15]

What should we make of the new calls to reinvent government, this time using AI?[16] Can AI make good on a twenty-five-year-old promise to remake governance through technology? Will it, as IBM and many others suggest, yield a more nimble, responsive, and transparent public sector? Or will the new algorithmic governance tools fall prey to Sunnyvale's trap of promising a silver technology bullet? Worse, might AI tools erode, rather than promote, internal efficacy and external accountability, or even spark the same demoralized exodus from government as Sunnyvale's ill-fated experiment? And how might law manage these opportunities and risks?

In 2019, we led a unique, interdisciplinary team of three dozen lawyers and computer scientists to deliver a far-ranging report to the Chair of the Administrative Conference of the United States (ACUS) on the use of AI by

12. OSBORNE & GAEBLER, *supra* note 9, at 306.

13. OFFICE OF THE VICE PRESIDENT, CREATING A GOVERNMENT THAT WORKS BETTER & COSTS LESS: STATUS REPORT OF THE NATIONAL PERFORMANCE REVIEW 14 (1994).

14. Kelly Wilkinson, *Trouble in Paradise: Sunnyvale Is Nationally Recognized for Its Stable City Government. Now Employees Are Leaving En Masse*, SUNNYVALE SUN (Aug. 4, 1999) ("During the past five years, the city's employee turnover rates have nearly doubled, even though retirement rates have barely budged a percentage point.").

15. Daniel E. Ho, Cassandra Handan-Nader, David Ames & David Marcus, *Quality Review of Mass Adjudication: A Randomized Natural Experiment at the Board of Veterans Appeals, 2003-16*, 35 J.L. ECON. & ORG. 239 (2019); Daniel E. Ho & Sam Sherman, *Managing Street-Level Arbitrariness: The Evidence Base for Public Sector Quality Improvement*, 13 ANN. REV. L. & SOC. SCI. 251 (2017); *see also* John Buntin, *25 Years Later, What Happened to 'Reinventing Government'?*, GOVERNING (Sept. 2016), https://www.governing.com/topics/mgmt/gov-reinventing-government-book.html [https://perma.cc/XBG3-9RD5] (noting tendency of performance management systems to ossify and encourage agencies to "post good numbers" rather than develop innovative solutions to problems).

16. To be fair, IBM is hardly alone in its faith in a digitized revolution in the work of government. *See, e.g.*, Anusha Dhasarthy, Sahil Jain & Naufal Khan, *When Governments Turn to AI: Algorithms, Trade-Offs, and Trust*, MCKINSEY (2019); William D. Eggers, David Schatsky & Peter Viechnicki, *AI-Augmented Government: Using Cognitive Technologies to Redesign Public Sector Work*, DELOITTE (2017); Max Stier & Daniel Chenok, *The Future Has Begun: Using Artificial Intelligence to Transform Government*, IBM CENTER FOR THE BUS. OF GOV'T (2018); Franco Amalfi, *Building Government for the 21st Century*, ORACLE (2018); Hila Mehr, *Artificial Intelligence for Citizen Services and Government*, HARVARD ASH CTR. FOR DEMOCRATIC GOVERNANCE AND INNOVATION (2017); Miguel Carrasco et al., *The Citizen's Perspective on the Use of AI in Government*, BCG (2019).

federal regulatory agencies.[17] We canvassed the roughly 140 most important federal departments, agencies, and sub-agencies for evidence of adoption of AI and machine learning and conducted in-depth case studies, relying on extensive interviews and documentation, to unearth some of the most innovative uses of AI for core government functions.

Our research brings to light a wide catalog of algorithmic governance tools, thus confirming AI's extraordinary potential to reimagine core agency functions across the full range of agency processes and outputs, from enforcement and adjudication to citizen engagement and procurement. The project likewise confirms that the proliferation of new algorithmic governance tools throughout the administrative state will shift, perhaps substantially, the subtle balance among technical efficiency, democratic accountability, and regularity at the heart of sound administrative governance. Perhaps most significant of all, our project points up the poverty of existing thinking about how to build a sensible accountability structure around the new algorithmic governance. Most of the scholarly literature remains untethered from the actual state of technology, offering "thought experiments,"[18] focusing on potential rather than actual applications,[19] or abstracting away from any concrete applications at all.[20] Moreover, by fixating on a small set of criminal justice

17. *See* DAVID FREEMAN ENGSTROM, DANIEL E. HO, CATHERINE M. SHARKEY & MARIANO-FLORENTINO CUÉLLAR, GOVERNMENT BY ALGORITHM: ARTIFICIAL INTELLIGENCE IN FEDERAL ADMINISTRATIVE AGENCIES (2020), https://www-cdn.law.stanford.edu/wp-content/uploads/2020/02/ACUS-AI-Report.pdf [https://perma.cc/TSF5-TF2Q] [hereinafter GOVERNMENT BY ALGORITHM].

18. Eugene Volokh, *Chief Justice Robots*, 68 DUKE L.J. 1135, 1137 (2019).

19. *See*, *e.g.*, Niva Elkin-Koren & Michal S. Gal, *The Chilling Effect of Governance-by-Data on Data Markets*, 86 U. CHI. L. REV. 403 (2019) (considering use of data and AI to craft "personalized law"—for instance, a speed limit for each driver).

20. *See*, *e.g.*, Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1258 (2008) (offering a "new framework for administrative and constitutional law designed to address the challenges of the automated administrative state" but abstracting from use cases save occasional references to no-fly lists and state-level social welfare benefit eligibility determinations); Cary Coglianese & David Lehr, *Transparency and Algorithmic Governance*, 71 ADMIN. L. REV. 1, 29 (2019) [hereinafter Coglianese & Lehr, *Transparency*] (offering a "general analysis" of conceptions of transparency in the context of algorithmic governance, but rooting the analysis almost entirely in *potential* uses of algorithmic tools by, among others, the Occupational Safety and Health Administration, the Federal Aviation Administration, and the Pipeline and Hazardous Materials Safety Administration); Cary Coglianese & David Lehr, *Regulating by Robot: Administrative Decision Making in the Machine-Learning Era*, 105 GEO. L.J. 1147 (2017) [hereinafter Coglianese & Lehr, *Regulating by Robot*] (focusing mostly on potential uses of algorithmic governance tools by the U.S. Pipeline and Hazardous Materials Safety Administration and the Occupational Safety and Health Administration, among others, and making no effort to isolate and examine specific existing use cases). The one exception is a growing literature on use of algorithmic "risk assessment" tools to assist bail, sentencing, and parole decisions within the criminal justice system. *See* Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CAL. L. REV. 671 (2016); Andrew Guthrie Ferguson, *Predictive Prosecution*, 51 WAKE FOREST L. REV. 705 (2016); Aziz Z. Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 DUKE L.J. 1043 (2019); Sandra G. Mayson, *Bias In, Bias Out*, 128 YALE L.J. 2218 (2019); Andrew D. Selbst, *Disparate Impact in Big Data Policing*, 52 GA. L. REV. 109 (2017); Joel Tito, *Destination Unknown: Exploring the Impact of Artificial Intelligence on Government*, CTR. FOR PUB. IMPACT (2017), https://resources.centreforpublicimpact.org/production/2017/09/Destination-Unknown-AI-and-government.pdf [https://perma.cc/DYJ5-33NN] (exploring criminal justice use cases only).

applications and commingling public and private sector use of AI despite their very different characteristics and imperatives, the existing literature operates at a high level of abstraction and, perhaps of necessity, focuses narrowly on constitutional principles, particularly procedural due process and equal protection.[21] Only a trickle of research treats the more fine-grained statutory requirements of administrative law and, even then, offers mostly a surface-level tour of potentially applicable doctrines.[22]

This Article seeks to shift debate onto a more concrete footing by offering a more technically informed account of the new algorithmic governance tools and the practical and legal challenges they raise and then advancing a novel proposal for their regulation. In so doing, we make three distinct contributions.

First, we push past the abstractions of the existing literature and, drawing on extensive in-depth interviews and research into technical and operational details, offer rich descriptive insight into frontier AI applications at two federal agencies, highlighting their likely evolution and the key normative and distributive implications of their adoption. A trio of tools in use at the Social Security Administration (SSA) aids in the adjudication of disability benefits, by (i) clustering together similar cases for more efficient and equitable disposition by administrative judges, (ii) identifying cases likely to be full grants, enabling the SSA to conserve resources required for a full hearing, and (iii) flagging errors in draft decisions by administrative judges, thus potentially avoiding costly appeals and reversals and improving decision consistency. Turning to agency enforcement, the Securities and Exchange Commission (SEC) is, like several other key enforcement agencies (e.g., the Internal Revenue Service and the Environmental Protection Agency), developing and deploying machine learning applications that help focus scarce agency investigative and enforcement resources on high-risk individuals and entities.[23] We focus in

21. *See*, *e.g.*, FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (2016); Mike Ananny & Kate Crawford, *Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability*, 20 NEW MEDIA & SOC'Y 973 (2016); Citron, *supra* note 20; Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1 (2014); Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93 (2014); Andrew Guthrie Ferguson, *Big Data and Predictive Reasonable Suspicion*, 163 U. PA. L. REV. 327, 329–30 (2015). A more recent line of inquiry pushes to an even higher level of abstraction, exploring the implications of algorithmic governance tools for the legitimacy of the administrative state. *See* Ryan Calo & Danielle Keats Citron, *The Automated Administrative State: A Crisis of Legitimacy*, EMORY L.J. (forthcoming 2020).

22. *See, e.g.*, Citron, *supra* note 20; Mariano-Florentino Cuéllar, *Cyberdelegation and the Administrative State, in* ADMINISTRATIVE LAW FROM THE INSIDE OUT: ESSAYS ON THEMES IN THE WORK OF JERRY L. MASHAW 134 (Nicholas R. Parrillo ed., 2017); Coglianese & Lehr, *Regulating by Robot*, *supra* note 20; Deirdre K. Mulligan & Kenneth A. Bamberger, *Procurement as Policy: Administrative Process for Machine Learning*, 34 BERKELEY TECH. L.J. 773 (2019).

23. In what follows, we adopt a wide definition of enforcement that includes not just formal enforcement actions but also investigations, audits, and other forms of regulatory monitoring that may or may not lead to enforcement actions. For a contrary effort to distinguish monitoring and enforcement, see Rory Van Loo, *Regulatory Monitors: Policing Firms in the Compliance Era*, 119 COLUM. L. REV. 369 (2019).

particular on a set of tools that predict insider trading and also which investment advisors and other agency "registrants" are violating their obligations under the federal securities laws. Looking across adjudication and enforcement—both critically important governance tasks, but with very different logics and features—permits an analysis that is at once concrete and synthetic, yielding well-grounded but generalizable insights about whether, and how, to regulate public sector AI use.

Second, we move beyond the existing literature's focus on constitutional law and consider how administrative law will or should adapt to the shift to algorithmic governance. While the existing literature's focus on constitutional law has yielded welcome insights,[24] we argue that much, if not most, of the hard work of regulating the government's new algorithmic toolkit will come not in the clouds of constitutional doctrine but in the statutory streets of administrative law. Yet administrative law's virtual absence in academic and policy discussion is concerning not merely because of its likely centrality. Its neglect also matters because how current doctrine will resolve the most pressing cases is far from certain. Using adjudication and enforcement as examples and comparing and contrasting their legal treatment under administrative law, we show that current case law is unclear about whether algorithms used to make or support enforcement or adjudication decisions can be subjected to judicial review under the Administrative Procedure Act (APA) at all, or whether algorithms constitute legislative rules that must undergo notice and comment. To date, none have. This uncertainty is a problem because administrative law, at least in its current guise, is unlikely to translate into a consistent and comprehensive approach to regulating public sector AI use that consciously balances the imperatives of internal administration with the legal demands of external accountability.

Third, we offer a novel and generalizable solution for monitoring, oversight, and accountability as the administrative state adopts these tools. We begin by spelling out some of the limitations of a pair of prescriptions. A minimalist option would be to retrofit the APA to enable prudent ex ante review of algorithmic tools through the notice and comment process or judicious ex post review by courts. We offer some tentative suggestions in this regard but also raise questions about whether front-end rulemaking and back-end judicial review of the usual APA sort are well suited to wrestle with the systemic considerations relevant to AI adoption. Forcing algorithms into the current template of notice and comment is over-inclusive and risks retarding the regulatory state's adoption of modern technology, thus exacerbating the public-private technology gap. At the same time, ex post judicial review of algorithmic governance tools and their outputs under current doctrine, where it can be had at all, does not address key concerns and suffers from a substantial mismatch in judicial capacity and the technical demands of algorithmic

24. *See supra* notes 21-22 and accompanying text.

oversight. We also consider a commonly discussed but similarly limited solution that would look to an oversight board (e.g., an "FDA for algorithms"[25]) staffed with technologists, academics, lawyers, and agency representatives to monitor, investigate, and recommend adjustments to agency adoption and use of AI.

We then argue that a further and promising intervention would require agencies to engage in what we call "prospective benchmarking." In a nutshell, agency administrators would reserve and then analyze a random sample of decisions using the agency's conventional, non-algorithmic approach. This "human alongside the loop" approach provides critical information and a comparison set to help smoke out when an algorithm has gone astray, when encoding the past may miss new trends, when an algorithm may create disparate impact, or when "automation bias" causes excessive deference to machine outputs. In the end, modernizing the administrative state will entail both adapting AI and crafting administrative procedures to address the mix of technical, distributive, and bureaucratic challenges raised by AI. Benchmarking, we conclude, is a promising step in that direction that deserves consideration and further elaboration.

Before launching, some clarifications are in order. First, we use "artificial intelligence" to mean any instance where an agency deploys models to learn from data with the goal of prediction. AI is thus used interchangeably with machine learning but excludes simple process automation (e.g., a case management system to digitally process benefits applications) and conventional statistical analysis (e.g., regression with the aim of drawing a causal inference).[26] Second, our description of AI techniques aims for the mid-level between the technical and abstract. Government agencies rarely publish technical manuals that spell out all of a tool's machine learning methodology, whether because of reliance on third-party contractors to develop systems or understandable concern about gaming by the regulated community. By focusing our analysis on algorithmic governance tools developed in-house by agency technologists, we can provide richer insights into how the systems function. Finally, while our ACUS project encompassed nearly the entire federal administrative state,[27] we limit our analysis to core adjudicatory and enforcement functions for expositional clarity and analytic leverage over the

25. Andrew Tutt, *An FDA for Algorithms*, 69 ADMIN. L. REV. 83 (2017).

26. Note that our definition may exclude use cases that predominate in much of the emerging literature on algorithmic accountability in the administrative state, particularly the relatively simple, rule-based, "logical AI" systems adopted by many state agencies to automate public benefits determinations. *See*, *e.g.*, Calo & Citron, *supra* note 21. Perhaps unsurprisingly, litigation challenging those tools has not centered on the opacity concerns that animate much of the algorithmic accountability literature but rather has leveled relatively straightforward procedural due process claims against the use of grossly deficient means of snail-mail notice and the imposition of draconian fines after unrealistically short time periods for those targeted to respond. *See, e.g.*, Cahoo v. SAS Analytics, 912 F.3d 887 (6th Cir. 2019); Barry v. Lyon, 834 F.3d 706 (6th Cir. 2016).

27. As noted previously, we set aside only those domains in which little public information exists, such as national security.

challenges of public sector AI use. Similarly, while our ACUS project treats numerous legal, technical, and practical implications of the new algorithmic governance, we focus here on the twin challenges of internal administration and external accountability. Readers interested in AI uses supporting other governance tasks or who seek to understand other challenges of public sector AI use—from machine learning's technical limits to adversarial learning and capacity building—are directed to the ACUS report itself and related work.

Our article proceeds from here as follows. Part I provides an in-depth view of two leading use cases in agency adjudication and enforcement and spells out the trajectory and challenges of AI adoption in each. Part II considers administrative law's response under existing doctrine. Part III evaluates prescriptive proposals, including retrofitting the APA and an oversight board, and then fleshes out the novel solution built around prospective benchmarking. A concluding part returns to Sunnyvale and offers some brief reflections on the promise and peril of the new algorithmic governance.

I. The Algorithmic Trend in Adjudication and Enforcement

This Part describes the shift to algorithmic decision making in adjudication and enforcement. We focus on these areas because they represent core areas of administrative governance, where two agencies in particular have engaged in considerable experimentation with AI: formal adjudication at the SSA and enforcement at the SEC. Our aims are three-fold. The first is to paint a rigorous, ground-level portrait of the tools in use at both agencies. Facts matter in law, and ungrounded speculation about errant "robot judges," while vivid, do little to separate fact from fiction. Surfacing key technical and operational details of the tools in use at the SSA and SEC is a critical first step in understanding the substantial challenges algorithmic governance poses for administrative law—the subject of Part III. The second aim is to offer an informed prediction, based in a mix of legal and engineering judgment, about the likely trajectory of AI-based adjudication and enforcement tools. Third and finally, we aim to connect up tools in use at the SSA and SEC to the wider algorithmic accountability literature and show where that literature does and does not capture the realities of the new algorithmic technologies of governance.

In pursuing each of these ends, the rich descriptions that follow highlight the significant potential, and also the risks, of AI-based governance technologies. In adjudication, AI holds the promise of solving the accuracy challenge that has bedeviled the SSA for generations. As Jerry Mashaw put it, in adjudication "the process is the product."[28] On the one hand, more efficient and accurate processing of cases can reduce gross backlogs and might even

28. Jerry L. Mashaw, *Reinventing Government and Regulatory Reform: Studies in the Neglect and Abuse of Administrative Law*, 57 U. PITT. L. REV. 405, 412 (1996).

revive the lost constitutional value of dignity by freeing up judges to provide hearings independent of their accuracy benefits. At the same time, supplanting human decision making entirely or relegating human decisions to ratification of machine recommendations may gut legal process of its dignitary values even if the system proves accurate.[29] In the enforcement context, machine learning promises to aid the SEC in identifying likely violators of the securities laws, by enabling the agency to sift through mountains of data. And yet, being singled out and made to defend against a regulatory or legal action, even if ultimately vindicated, is costly. The process itself, as the saying goes, can be the punishment.[30]

As a final note, our look beneath the hood at the SSA and SEC concretely confirms the transparency concerns that feature heavily in, and indeed dominate, the emerging algorithmic accountability literature. But our descriptive portrait also introduces issues that have barely registered in that literature. First, there are steep technical challenges to automating government tasks that trade in large amounts of unstructured text. Second, internal capacity building will be central to the AI transition, given the iterative process of developing useable tools and the constant threat of gaming. Last, the demand for intelligible models may not solely be driven by regulated parties or courts, but rather from within the agency itself. Demands for intelligible and explainable AI outputs by the front-line staff attorneys at the SSA and line-level enforcers at the SEC hold the promise of a form of internal due process that can help mitigate accountability concerns. While these are important insights, much of our analysis in the Parts to come focuses largely on external accountability mechanisms. We once more refer the reader interested in these other issues to the full ACUS report.

A. Adjudication of Disability Benefits at the Social Security Administration

1. The Problem of ALJ Arbitrariness

Begin with a classic problem of formal adjudication: decisional independence risks arbitrariness.[31] Figure 1 displays disposition data for SSA ALJs in 2018. Each dot represents one ALJ, with number of decisions on the *x*-axis and the award rate on the *y*-axis. We observe extreme variation in award rates. In one region, one judge awarded 8% of all cases and another awarded 98% of all cases. Because cases are randomly assigned within an office, we can compare the extent of variation expected under chance alone, plotted in grey.

29. *Id*.

30. *See* MALCOLM M. FEELEY, THE PROCESS IS THE PUNISHMENT: HANDLING CASES IN A LOWER CRIMINAL COURT (1979).

31. *See* HAROLD J. KRENT & SCOTT MORRIS, ACHIEVING GREATER CONSISTENCY IN SOCIAL SECURITY DISABILITY ADJUDICATION: AN EMPIRICAL STUDY AND SUGGESTED REFORMS 15 (2013); JERRY L. MASHAW ET AL., SOCIAL SECURITY HEARINGS AND APPEALS: A STUDY OF THE SOCIAL SECURITY ADMINISTRATION HEARING SYSTEM 21 (1978).

We can resoundingly reject the notion that these disparities are the result of chance variability.

Much ink has been spilled on the topic, including the potential for appellate review, performance measurement, quality assurance, and peer review to cure these deficits.[32] Yet while Professor Jerry Mashaw famously highlighted the problem of inconsistency some 40 years ago, decisional arbitrariness persists to the present day.[33]

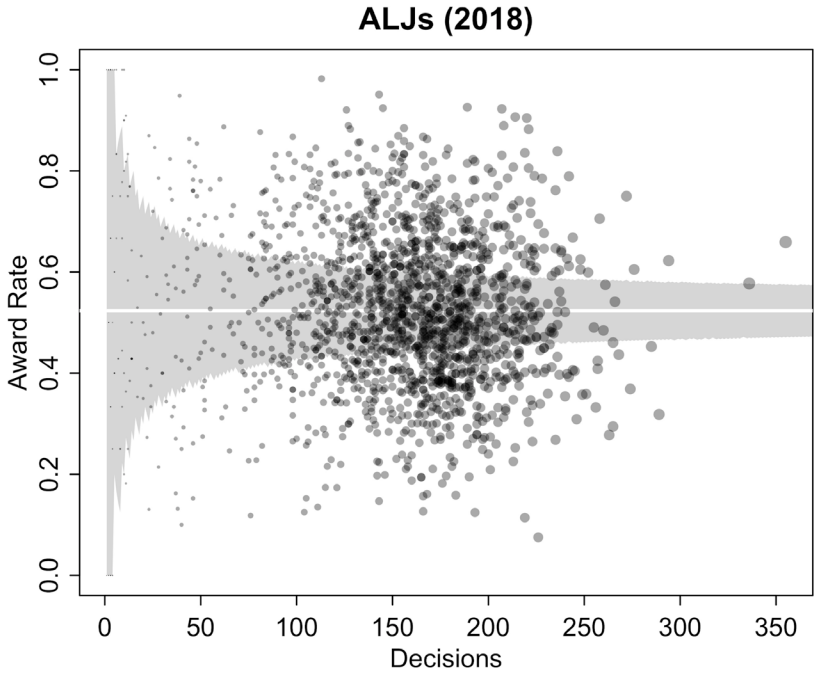

Figure 1. Number of decisions on *x*-axis against the award rate on the *y*-axis for all ALJs in 2018. The grey interval indicates the pointwise 95% interval under the null hypothesis that ALJs have the same underlying grant rate.

32. JONAH B. GELBACH & DAVID MARCUS, A STUDY OF SOCIAL SECURITY DISABILITY LITIGATION IN THE FEDERAL COURTS (2016); Daniel E. Ho, *Does Peer Review Work? An Experiment of Experimentalism*, 69 STAN. L. REV. 1 (2017); Daniel E. Ho & Sam Sherman, *Managing Street-Level Arbitrariness: The Evidence Base for Public Sector Quality Improvement*, 13 ANN. REV. L. & SOC. SCI. 251 (2017); Jerry L. Mashaw, *The Management Side of Due Process: Some Theoretical and Litigation Notes on the Assurance of Accuracy, Fairness, and Timeliness in the Adjudication of Social Welfare Claims*, 59 CORNELL L. REV. 772 (1974); Kathleen G. Noonan et al., *Legal Accountability in the Service-Based Welfare State: Lessons from Child Welfare Reform*, 34 L. & SOC'Y INQUIRY 523 (2009); William H. Simon, *Legality, Bureaucracy, and Class in the Welfare System*, 92 YALE L.J. 1198 (1984).

33. *See* Paul Verkuil, *Meeting the Mashaw Test for Consistency in Administrative Adjudication*, *in* ADMINISTRATIVE LAW FROM THE INSIDE OUT: ESSAYS ON THE THEMES IN THE WORK OF JERRY L. MASHAW (Nicholas Parrillo ed., 2017); David Ames, Cassandra Handan-Nader, Daniel E. Ho & David Marcus, *Due Process and Mass Adjudication: Crisis and Reform*, 72 STAN. L. REV. 1 (2020).

2. Pioneering Applications of AI

Can AI change this state of affairs? The SSA Appeals Council has developed three applications of AI in adjudication.[34] The first application aimed to address a particular challenge of the existing case assignment system to adjudicators: because cases were randomly drawn, adjudicators were necessarily forced to crisscross from one body of law to the next. Each different area involves a complex set of decisions, with (manual) decision trees mapping roughly 2,000 possible paths of disability cases.[35] The Appeals Council hence developed a clustering algorithm to enable individuals to process cases by substantive similarity, enabling adjudicators to develop familiarity with the same part of the decision tree. The latent class model used hearing level information (e.g., age of claimant, functional impairments, and state of origin) to create clusters of comparable cases. Due to labor-management concerns, clustering only reordered how cases were processed within an adjudicator's docket, and did not change the composition of cases across adjudicators. In that sense, clustering facilitated "micro-specialization," not macro-specialization across adjudicators. Through an early pilot, where branch chiefs could elect to use the clustering results, the Appeals Council reported a 7% gain in productivity and a 12.5% reduction in errors.

The second application was aimed to save resources on costly in-person hearings by developing a model to predict cases likely to result in full grants. In 2010, SSA finalized a rule to enable a "Quick Disability Determination" (QDD) at the initial decision level.[36] The model would use information about medical history, treatment protocols, medical symptoms, and findings to predict easy grants, to be reviewed by state QDD examiners. Similarly, SSA developed a pilot program for expediting claims at the ALJ hearing level. The model used Naive Bayes classification with state-level information to predict fully favorable dispositions (as opposed to dispositions that are favorable, unfavorable, or dismissals), again to be reviewed manually for a recommended grant.

The third, and most ambitious, application is the "Insight" system developed by Kurt Glaze, an attorney-turned-analyst at SSA. The system draws on the decision trees and policies developed beginning in the 1990s and uses structured input to test for adherence with policies. In addition, the system uses natural language processing (NLP) (regular expressions, semantic parsing, and supervised classification) to flag potential errors and inconsistencies in draft

34. Gerald K. Ray & Jeffrey S. Lubbers, *A Government Success Story: How Data Analysis by the Social Security Appeals Council (with a Push from the Administrative Conference of the United States) Is Transforming Social Security Disability Adjudication*, 83 GEO. WASH. L. REV. 1575 (2015).

35. How Data Analysis is Transforming Disability Adjudication at the Social Security Administration, Presentation at the Government Performance Summit, May 4-5, 2015.

36. *Administrative Review Process for Adjudicating Initial Disability Claims*, 71 Fed. Reg. 16,242 (Mar. 31, 2006); 20 C.F.R. §§ 404.1619, 416.1019 (2020).

decisions. For instance, Insight extracts functional impairments and compares whether the impairment is consistent with the job classification in the Dictionary of Occupational Titles from the Department of Labor.[37] Figure 2 presents an early prototype screenshot of the kind of flag meant to guide attorneys and ALJs in the adjudicatory process. The Insight system was adopted on a voluntary basis at the Appeals Council in 2016 and at hearing offices in 2017. Early results suggested a reduction in processing time and a reduction in "returns" to adjudicators for error.

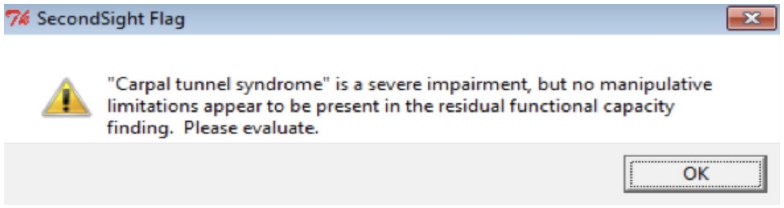

Figure 2. Prototype screenshot (modified for visibility) from Insight system flagging a potential inconsistency in a draft decision.

3. Trajectory

SSA's adoption of AI has been more advanced than at other adjudicatory agencies. While the effect on hearing-level decisions by ALJs remains unclear, these applications are suggestive of the future adoption of AI in formal adjudication, particularly taking into account rapid advances in natural language processing (NLP). Such techniques have wide applicability across adjudicatory agencies, from immigration adjudication at the Executive Office for Immigration Review to veterans adjudication at the Board of Veterans' Appeals to Medicare disputes at the Office of Medicare Hearings and Appeals.

In the near future, disposition forecasts may improve the accuracy and consistency of decisions by attorneys and ALJs. For instance, each adjudicator might be presented with a probabilistic forecast of a grant, against which the attorney can compare her own inclination, much in the way that "risk assessment scores" in criminal justice are used in pretrial detention decisions. In the medium run, feature extraction from claims records folders may help adjudicators identify important elements of the case. The claims file currently is displayed to attorneys and ALJs in digital (PDF or TIFF) format, and the process of manually identifying relevant entries (e.g., medical exam results) is time-consuming. Either by adapting NLP-based information extraction tools or converting to an electronic health data standard, systems may speed up this review of claims folders. The most ambitious version would be the deployment of language models to aid in drafting benefits decisions. By extracting

37. DEP'T OF LABOR, DICTIONARY OF OCCUPATIONAL TITLES (1991), https://www.oalj.dol.gov/LIBDOT.HTM [https://perma.cc/A2FF-96PU].

information from the claims folder and using meta-information about the case (e.g., knee injury of Gulf War veteran involving a claim for a ratings increase), an AI application may someday be capable of predicting the likely language, legal authority, and evidentiary basis of the decision: auto-complete for law.

4. Implications

On the one hand, the benefits to these tools appear clear: AI might finally help crack the code of mass adjudication, improving accuracy, reducing inconsistency, and cutting down on rampant backlogs that plague agencies like the SSA, the Office of Medicare Hearings and Appeals, the Board of Veterans Appeals, and the Executive Office of Immigration Review. Perhaps most tantalizing is that if AI can generate more "accurate" (or at least more consistent) decisions, it may help reclaim a lost part of constitutional due process. The post-*Goldberg* consensus has been that accuracy is the linchpin of due process. As Justice Brennan reasoned in *Goldberg*, the "hearing has one function only . . . to protect a recipient against an erroneous termination . . . ."[38] Yet QDD challenges us to think whether we would indeed want to skip hearings when the hearing may not contribute to accuracy. Indeed, eliminating hearings may cause the very "societal malaise" that *Goldberg* worried about.[39] Alternatively, by taking much of the rote and repetitive work out of judging, AI might free up judicial resources to focus on procedural fairness elements of the job: to hold hearings, provide tentative orders, and engage individuals with explanation. One need not look very far into litigant reviews of ALJs to find evidence of the dignity value of hearings. Wrote one litigant: "I know I had to have shown my complete nervousness but, after speaking and listening to him talk with kindness, I felt relief. He was truly a great Judge even though I was denied . . . ."[40]

On the other hand, the adoption of AI, particularly in light of the trajectory of use cases, raises serious questions. First, each of the SSA use cases may increasingly displace the exercise of judicial discretion, even when manual review remains nominally present. Concrete examples of the resulting "automation bias," or overreliance on machine outputs, are not hard to imagine.[41] For example, a machine-predicted disposition might allow an ALJ to compare her inclination to the wisdom (or foolishness?) of the crowd, potentially threatening notions of decisional independence. The search tool may

38. Goldberg v. Kelly, 397 U.S. 254, 267 (1970).

39. *Id*. at 265.

40. *Administrative Law Judge Case Statistics*, DISABILITY JUDGES (Aug. 5, 2012, 11:27 PM), https://www.disabilityjudges.com/state/virginia/norfolk/james-j-quigley [https://perma.cc/W92C-KGP5].

41. "Automation bias" refers to the tendency of humans to unreasonably defer to automated outputs over time. *See* Citron, *supra* note 20, at 1272; R. Parasuraman & D.H. Manzey, *Complacency and Bias in Human Use of Automation: An Attentional Integration*, 52 HUM. FACTORS 381 (2010); Linda J. Skitka, Kathleen L. Mosier & Mark Burdick, *Does Automation Bias Decision-Making?*, 51 INT'L J. HUM.-COMPUTER STUD. 991 (1999).

allow an ALJ to spend less time reviewing the full record of a case, thus losing nuance and eroding de novo review. A machine-generated decision template might, whether gradually or more abruptly, convert an ALJ's role from drafting to simply signing an automated body of text, much in the way that standard form contracts are signed. And because there will surely be disparities in how much effort ALJs will expend to review AI-assisted product, the present inter-ALJ disparities noted previously[42] may be exported into willingness to deviate from the automated default.

Second, if these tools enable centralization of policy control, they raise deep questions about separation of functions and powers within agencies. In immigration adjudication, for instance, the exemption from performance reviews was only secured by letter, not statute or regulation.[43] As a result, the exemption was later removed, enabling greater forms of presidential control of immigration adjudication.[44] To the extent that tools like Insight promote such control, they may facilitate converting an adjudicatory agency into an executive one.

Third, while automating adjudication may be cost-effective, it may undercut the perceived legitimacy of agency decision making. The contrary view is expressed by Professor Eugene Volokh, who argues that we should "focus on the quality of the proposed AI judge's product, not on the process that yields that product."[45] But for mass adjudicatory agencies, there exists no exogenous measure of quality—or, as Jerry Mashaw put it, "no objective, external referent for determining [an 'accurate' decision]." Hence, "to change the process of decision, to 'reengineer' it, is to change the product as well." From that perspective, each step of displacing human discretion changes the product of adjudication. Without an external referent for accuracy, we should be cautious about the implications. Do these use cases undercut the tailoring of law to fact? Does it matter if QDD can only be applied to initial decisions that are filed electronically, hence disbursing expedited benefits determinations to a demographically distinct (albeit large) set of applicants? Does the Insight system in fact create a new binding policy in a way that violates administrative law's demands for transparency and explanation? In the long term, these developments may erode the APA understanding of formal adjudication.

Finally, and despite these open questions, we lack even the most basic understanding of the impact of algorithmic tools on agency adjudication. To be sure, SSA conducted internal studies that indicated that employees who opted to use the Insight system identified more errors and processed cases more quickly than employees opting against using the Insight system. But usage was voluntary, therefore making it hard to attribute performance differences to the

42. *See supra* Section I.A.1.

43. *See Threat to Due Process and Judicial Independence Caused by Performance Quotas on Immigration Judges*, NAT'L ASS'N OF IMMIGRATION JUDGES 1 (Oct. 1, 2017).

44. *See id*.

45. Volokh, *supra* note 18, at 1191.

Insight system itself. If more motivated employees adopted the Insight system, the performance differences may stem simply from different levels of buy-in. In an audit of the Insight system, SSA's Office of the Inspector General echoed this sentiment and concluded that "management should define objectives in measurable terms so performance toward achieving those objectives can be assessed."[46] Given what is at stake, it is critical that administrative law take seriously the turn to algorithmic adjudication, which we consider beginning in Part III.

B. Regulatory Enforcement at the Securities and Exchange Commission

1. The Challenge of Enforcement

Agency enforcement poses a classic tradeoff between discretion and accountability.[47] Discretion is necessary because agency resources are finite but regulatory targets, and the monitoring and search costs that an agency accrues in identifying them, are nearly limitless.[48] Monitoring and search costs can quickly eat up agency budgets. Moreover, optimal deterrence does not support proceeding against every possible regulatory target. Even enforcement actions that are formally cost-justified—that is, actions in which the social benefit exceeds the social cost of bringing them—may not be a sound use of agency resources given other agency imperatives and priorities.[49] But prosecutorial discretion—and an agency's decision when to wield the coercive power of the state and when not to—also brings risks. Agency forbearance can mask an agency's infidelity to statutory design and purposes.[50] It can also conceal arbitrary selection of enforcement targets, which is itself socially costly.[51] Indeed, the mere fact of being targeted for audit or investigation by an agency can impose significant harms on regulated parties, even if they are ultimately

46. SOC. SEC. ADMIN., OFFICE OF THE INSPECTOR GEN., THE SOCIAL SECURITY ADMINISTRATION'S USE OF INSIGHT SOFTWARE TO IDENTIFY POTENTIAL ANOMALIES IN HEARING DECISIONS 5 (Apr. 2019).

47. *See* Margaret H. Lemos, *Democratic Enforcement? Accountability and Independence for the Litigation State*, 102 CORNELL L. REV. 929, 935 (2017) (noting "the challenge of designing enforcement institutions in a way that promotes accountability while preserving a role for independent, professional judgment").

48. *See* Robert A. Kagan, *Editor's Introduction: Understanding Regulatory Enforcement*, 11 L. & POL'Y 89, 110 (1989) ("Most regulatory agencies feel chronically understaffed and underbudgeted relative to their caseload.").

49. *See id*. at 93 (noting ideal agency pursues welfare-maximization by "focus[ing] its energies where it can do the most good, guided by a sense of what is legally, technologically, economically, and politically possible"); Gary Becker, *Crime and Punishment: An Economic Approach*, 76 J. POL. ECON. 169 (1968) (offering classic account of optimal deterrence).

50. *See* Rachel E. Barkow, *Overseeing Agency Enforcement*, 84 GEO. WASH. L. REV. 1129, 1150 (2016) (noting that agencies can "behave improperly if the targets it selects for enforcement are disproportionately singled out in ways that are unwarranted under the legal standards").

51. *See* Elizabeth Magill, *Foreword: Agency Self-Regulation*, 77 GEO. WASH. L. REV. 859, 901 (2009) ("If the agency chooses to pursue one class of violators instead of others, that places a burden on those who are pursued, and, if the two classes compete with one another, the agency's action provides a relative benefit to those who are not pursued.").

vindicated.[52] Process, as we have already noted, can be a costly and undue form of punishment.[53]

2. Pioneering Applications of AI

The Securities and Exchange Commission's (SEC) development of a suite of algorithmic enforcement tools provides a window into the possibilities and limits of predictive analytics in reducing agency search costs and making agency enforcement decisions more precise and less arbitrary. Three tools in particular illustrate the agency's approach.

The first two tools target trading-based market-based misconduct. One of these, known as the Advanced Relational Trading Enforcement Metrics Investigation System, or ARTEMIS, identifies and assesses suspicious trading. ARTEMIS "analyzes patterns and relationships among multiple traders using the Division's electronic database of over six billion electronic equities and options trading records."[54] This tool aims to catch all instances of insider trading in the market and powerfully enhances the SEC's monitoring and surveillance powers. ARTEMIS's focus is serial offenders and cheaters. This is generally thought to be an easier demographic of offenders to find as compared to first-time insider trading activities. The other tool, called the Abnormal Trading and Link Analysis System (ATLAS), is the newest of the SEC's algorithmic enforcement tools and complements the ARTEMIS tool by focusing on first-time, rather than serial, insider trading.

Neither ARTEMIS nor ATLAS is fully automated. Both first require agency enforcement staff to identify a suspected offender before more targeted data collection and a full-fledged investigation can proceed. The ARTEMIS process typically starts with automated analysis of the public filings of a company that has experienced significant stock movement. While companies announce important events in scheduled 10-K and 10-R filings, they are also required to make announcements regarding material events of particular relevance to shareholders in a separate 8-K form. In the first step of the process, SEC analysts systematically pool these 8-K forms and then use two separate algorithmic tools to parse them. The first tool is an NLP topic model to classify documents into categories of significant reported events[55]—for instance, M&A

52. *See* Marshall v. Jerrico, Inc., 446 U.S. 238, 249 (1980) (noting that enforcement decisions can "result in significant burdens on a defendant or a statutory beneficiary, even if he is ultimately vindicated"). As noted below, however, the Court has not found these costs to be legally cognizable.

53. *See* FEELEY, *supra* note 30.

54. Mary Jo White, Chair, Sec. & Exch. Comm'n, Remarks at the International Institute for Securities Market Growth and Development (April 8, 2016), https://www.sec.gov/news/statement/statement-mjw-040816.html [https://perma.cc/NP5Z-3FHS].

55. The topic model represents filings in a term-document matrix ("bag of words") and models term generation as a function of latent topics. *See* David M. Blei & John D. Lafferty, *Topic Models: Classification, Clustering, and Applications*, *in* TEXT MINING 101 (Ashok N. Srivastava & Mehran Sahami eds., 2009).

targeting, CEO termination, or FDA approval decisions.[56] At the second step, this labeled data is pushed through a supervised learning algorithm to flag current filings and trigger events that may warrant further investigation. Once the data has been sifted and analyzed, a human examiner reviews the results.

An agency examiner who concludes that trading around a specific company's stock warrants further investigation prepares a "bluesheet request" and thus begins the more targeted collection and analysis of trading data.[57] Staff must identify which broker/dealers traded the security at issue by obtaining the clearing reports submitted to FINRA.[58] Staff must also must decide how far back in time to request data.[59] In order to ensure that bluesheet-derived data is high-quality, the SEC and FINRA[60] regularly bring charges against brokerage firms for inaccurate or incomplete submissions.[61]

Once bluesheet data have been collected, SEC staff use the ARTEMIS and ATLAS tools to analyze those data alongside data from prior bluesheet requests to determine whether the trading activity in question constitutes a suspicious anomaly.[62] The SEC has not disclosed the precise features the

56. Event categories include: M&A transaction target, bankruptcy, major commercial announcement, scheduled earnings announcements, unscheduled earnings announcement, clinical trial, FDA decision announcements, and court judgment.

57. A bluesheet is a statutorily authorized investigatory tool the SEC uses to request detailed trading data on a particular company's stock from the broker/dealer community. This information includes standard trading information (name of the security, whether the transaction was a buy or a sell, long or short, price, and date), as well as personal information about the trading participants (name, address, social security number). 17 C.F.R. § 240.17a-25 (2020). An example electronic bluesheet (also referred to as "EBS") is publicly available through the FINRA website, and can be examined to understand the criteria of data requested by the SEC. FIN. INDUS. REG. AUTH., ELECTRONIC BLUESHEET SUBMISSIONS 6 (Jan. 29, 2018), http://www.finra.org/sites/default/files/notice_doc_file_ref/Regulatory-Notice-18-04.pdf [https://perma.cc/5DGJ-2X9S].

58. WILLIAM M. PRIFTI, SECURITIES: PUBLIC & PRIVATE OFFERINGS app. J7 (2d ed. 2013); SEC. & EXCH. COMM'N, DIV. OF ENF'T, ENFORCEMENT MANUAL § 3.2.2 (Nov. 28, 2017) [hereinafter SEC Enforcement Manual] (describing bluesheets).

59. SEC Enforcement Manual, *supra* note 58, § *3.2.2*; Telephone Interview with Scott Bauguess, Former Deputy Dir. & Deputy Chief Economist, Sec. & Exch. Comm'n (Feb. 15, 2019) [hereinafter Bauguess Interview].

60. FINRA is the acronym for the Financial Industry Regulatory Authority that, in its own words, is "a not-for-profit organization authorized by Congress to protect America's investors by making sure the broker-dealer industry operates fairly and honestly." *See* FINRA, https://www.finra.org/about [https://perma.cc/F7WV-GUTY].

61. For instance, in June 2016, FINRA fined Deutsche Bank Securities Inc. $6 million for failing to meet regulatory reporting requirements in bluesheets generated from 2008-2015. The firm had submitted thousands of bluesheets that misreported or omitted critical information on over a million trades. *See* Press Release, Sec. & Exch. Comm'n, *Citigroup Provided Incomplete Blue Sheet Data for 15 Years* (Jul. 12, 2016), https://www.sec.gov/news/pressrelease/2016-138.html [https://perma.cc/KED7-5Y3G]. And in July 2016, Citigroup Global Markets Inc. was fined $7 million by the SEC for submitting 2,382 erroneous bluesheets from 1999 to 2014. Citigroup contended that these errors were attributable to a coding failure in Citigroup's internal electronic bluesheet system. Press Release, Fin. Indus. Reg. Auth., FINRA Fines Deutsche Bank Securities Inc. $6 Million for Submitting Inaccurate and Late Blue Sheet Data (June 29, 2016), http://www.finra.org/newsroom/2016/finra-fines-deutsche-bank-securities-inc-6-million-submitting-inaccurate-and-late-blue [https://perma.cc/U6JW-TQGC].

62. Bauguess Interview, *supra* note 59.

agency uses in either tool to make these determinations. However, the features used in the ATLAS tool are said to be "intuitive" and presumably focus on whether the trade was explicable for the trader given the context and also the trader's historical behavior.[63] The ARTEMIS tool uses an unsupervised learning model for anomaly detection. The ATLAS tool, by contrast, uses a supervised model called a one-class support vector machine (SVM) to determine if a particular trade is suspicious.[64] The potential regulatory targets who are fed into the ATLAS model are split into two categories: those who lost money on a trade, and those who made money. The SVM is trained on the former, then fit to the latter. The assumption is that the behavior of those who made money should not differ significantly from those who lost money over time. Outliers identified by both tools are treated as suspicious.[65]

A third AI-based enforcement tool parses the narrative disclosures that investment advisors and other agency "registrants" make to the SEC to predict which among them may be violating the securities laws and so should be subject to more stringent treatment under the agency's examination program.[66] Under that program, the SEC is responsible for conducting examination of a wide range of entities registered with the SEC, including tens of thousands of investment advisors, broker dealers, and mutual funds and exchange traded funds.[67] The sheer scope of the program creates significant opportunities to economize on scarce agency resources by concentrating examination efforts on a subset of registrants.

Because the Form ADV disclosures that investment advisors make to the agency are comprised, at least in part, of free text, NLP algorithms are used to normalize the inputs for analysis.[68] That process consists of three steps: (i) text

63. Features might include how often a trader trades the company's stock, how often she trades other stocks, how many shares were traded in comparison to the trader's other trades, and the time between the announcement and the trade. Telephone Interview with Staff, Complex Financial Instruments Unit, Sec. & Exch. Comm'n (Feb. 15, 2019) [hereinafter SEC Staff Interview I].

64. An SVM is a classifier that uses training data to create an optimal hyperplane that categorizes new examples. Savan Patel, *Chapter 2: SVM (Support Vector Machine) — Theory*, *in* MACHINE LEARNING 101 (May 3, 2017), https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72 [https://perma.cc/EGB4-DJDA].

65. SEC Staff Interview I, *supra* note 63.

66. The full name for these filings is the "Uniform Application for Investment Adviser Registration and Report by Exempt Reporting Adviser."

67. The SEC puts it this way:

> OCIE is responsible for conducting examinations of entities registered with the SEC, including more than 13,200 investment advisers, approximately 10,000 mutual funds and exchange traded funds, roughly 3,800 broker-dealers, about 330 transfer agents, seven active clearing agencies, 21 national securities exchanges, nearly 600 municipal advisors, FINRA, the MSRB, the Securities Investor Protection Corporation, and the Public Company Accounting Oversight Board, among others.

Press Release, U.S. Sec. & Exch. Comm'n, SEC Office of Compliance Inspections & Examinations Announces 2019 Examination Priorities (Dec. 20, 2018), https://www.sec.gov/news/press-release/2018-299 [https://perma.cc/GT2H-GGJJ].

68. Telephone Interview with Staff, Office of Research and Data Services, Div. of Econ. & Risk Analysis and Office of Analytics and Research, Sec. & Exch. Comm'n (Feb. 20, 2019) [hereinafter SEC Staff Interview II]. That form contains two parts. The first concerns the investment

extraction from PDF forms and segmentation into sections that answer specific questions from the form;[69] (ii) unsupervised learning to cluster types of documents and detect anomalies;[70] (iii) supervised learning using prior Form ADVs associated with referrals to the agency's enforcement arm to classify each investment advisor as "high," "medium," or "low" risk.[71] Entities flagged as "high" risk are passed on to SEC enforcement staff, with an explanation detailing the weight each feature was given by the model in calculating the score.[72]

3. Trajectory

As a well-resourced agency with significant technical capacity, the SEC sits well ahead of most other federal agencies in developing AI-based enforcement tools. But the SEC is by no means alone in its efforts. The Center for Medicare and Medicaid Services (CMS), the Internal Revenue Service, and the Environmental Protection Agency are all at various stages of development and deployment of algorithmic tools designed to predict illegal conduct or more precisely allocate scarce agency resources toward audit or investigation.[73] More, and more advanced, enforcement tools are surely on the way. Continued advances in NLP are likely to improve the accuracy of enforcement targeting. While word embeddings are not easily adapted to the securities domain—one of the top 10 embeddings for the word "insider" is "bigwig"[74]—cutting-edge language models (e.g., Google's BERT model) may facilitate transfer learning to adapt large-scale models to more jargon-filled legal texts and more domain-specific tasks with less training data. In the long-term, the most audacious

advisor's "business, ownership, clients, employees, business practices, affiliations, and any disciplinary events of the adviser or its employees." SEC. & EXCH. COMM'N, FAST ANSWERS: FORM ADV (Mar. 11, 2011), https://www.sec.gov/fast-answers/answersformadvhtm.html [https://perma.cc/HVD9-P2JL]. The second involves the advisor's services offered, fee schedule, "disciplinary information, conflicts of interest, and the educational and business background of management and key advisory personnel of the adviser." *Id*.

69. SEC Staff Interview II, *supra* note 68.

70. This approach can be done via Latent Dirichlet allocation (LDA), which uses a "bag of words" representation of text. This approach finds all of the words that are in a document and finds how many times they are repeated. The document "John bought stocks. Mary bought stocks", would be converted to BoW = {"John":1,"bought":2,"stocks":2,"Mary":1}. *See* David M. Blei, Andrew Y. Ng & Michael I. Jordan, *Latent Dirichlet Allocation*, 3 J. MACHINE LEARNING RES. 993 (2003).

71. This is done using a random forest model, an ensemble learning technique that generates many decision trees to classify data given a set of predicative labels. At inference time, each decision tree votes on how the data should be classified. *See* TREVOR HASTIE, ROBERT TIBSHIRANI & JEROME FRIEDMAN, THE ELEMENTS OF STATISTICAL LEARNING: DATA MINING, INFERENCE, AND PREDICTION 587-604 (2d ed. 2009).

72. Feature importance is calculated by calculating Gini importance. Cechine Lee, *Feature Importance Measures for Tree Models – Part I: An Incomplete Review* (Oct. 28, 2017), https://medium.com/the-artificial-impostor/feature-importance-measures-for-tree-models-part-i-47f187c1a2c3 [https://perma.cc/F9QQ-RZM8].

73. *See* GOVERNMENT BY ALGORITHM, *supra* note 17.

74. This is based on word2vec trained on the English GoogleNews Negative300 corpus. TURKUMLP GROUP, http://bionlp-www.utu.fi/wv_demo [https://perma.cc/UDB4-CG5C].

applications of AI could automate each step of an investigation (e.g., sending inquiry letters, compiling answers) all the way to filing an enforcement action. Just as the SSA's tools could become auto-complete for adjudication, these tools could become auto-complete for enforcement.

That said, the SEC's new algorithmic enforcement toolkit highlights some challenges as well. Such tools can only be as good as their data inputs, and unlocking AI's full potential in the enforcement context requires abundant, well-labeled data that accurately reflect ground truth. An initial challenge is non-randomness, as exemplified by the bluesheet process that feeds the SEC's ARTEMIS and ATLAS tools. That process, as noted previously, is neither comprehensive nor random. Instead, it is hypothesis-driven and reflects the assumptions, heuristics, and biases of enforcement staff in each case. Moreover, the ARTEMIS and ATLAS tools are trained on a pool of trading data that includes only prior bluesheet requests, not total trading activity.[75] When searching for patterns that suggest insider trading, the SEC's algorithmic tools compare current behavior to previously flagged behavior, not traders in the market as a whole, potentially impairing system accuracy. The IRS has historically solved this problem with random audit data, but no such gold standard data exists for the SEC.

A second and related challenge arises at the nexus of automation and human discretion. When line-level enforcement staff retain ultimate authority to initiate enforcement, uncritical reliance on automation may displace investigatorial attention away from false negatives and crowd out the exercise of discretion with suspected positives.[76] If prior enforcement actions serve as training data or determine the labels applied to that data, the system may unduly confine enforcement to a subset of violations. This phenomenon is well established in the predictive policing context: when a predictive model is used to deploy police patrols, and the resulting arrest data is used to retrain the

75. Bauguess Interview, *supra* note 59. Interestingly, the extent of the agency's transparency across the entire market may soon change. On November 15, 2016, the SEC approved a joint plan with FINRA and SROs to develop a consolidated audit trail ("CAT"). Press Release, Sec. and Exch. Comm'n, SEC Approves Plan to Create Consolidated Audit Trail (Nov. 15, 2016), https://www.sec.gov/news/pressrelease/2016-240.html [https://perma.cc/EL63-6KMT]. Adopted under SEC Rule 613, CAT requires SROs and broker-dealers to significantly enhance their information technology capacities to maintain a comprehensive database of granular trading activity in the U.S. equity and options markets. *Rule 613 (Consolidated Audit Trail)*, SEC. AND EXCH. COMM'N (Apr. 22, 2020), https://www.sec.gov/divisions/marketreg/rule613-info.htm [https://perma.cc/LCW7-K49H]. Rule 613 establishes a timeline for implementing CAT in the national market system ("NMS"). The reporting requirement went into effect for SROs on November 15, 2017, and to large broker-dealers on November 15, 2018. Smaller broker-dealers will have to be compliant for CAT reporting by November 15, 2019. CAT is poised to become the biggest central repository of stock exchange data, and broadens the reporting requirement to every trade quote and order, origination, modification, execution, routing, and cancellation. *Id.; see also Perspectives: Consolidated Audit Trail: The Wait Is Over,* DELOITTE (Aug. 2019), https://www2.deloitte.com/us/en/pages/financial-services/articles/sec-rule-613-consolidated-audit-trail-national-market-system-nms-plan-banking-securities.html [https://perma.cc/J4CP-KGHZ].

76. For more on automation bias, see note 41, *supra*.

model, the potential for a "runaway feedback loop" exists.[77] Police may be sent to the same neighborhoods over and over again, even if the underlying crime rate was random. In short, without proper safeguards, algorithmic detection at the SEC and elsewhere could become dominated by superficial features from prior enforcement decision making, replicating the idiosyncrasies and biases of line-level enforcers rather than building richer and more precise models of noncompliance.

The dynamic nature of wrongdoing poses a final challenge. For many agencies, enforcement resembles a game of "whack-a-mole" as regulated parties develop new artifices designed to evade, or narrowly navigate between, announced rules. Taking an example from tax, an algorithmic tool might be able to flag the complicated and choreographed set of transactions needed to implement an illegal tax shelter. But once agency enforcement begins, taxpayers and the tax compliance industry shift away and develop new artifices that are identifiable to algorithmic enforcement tools only if they are sufficiently similar to the prior ones.[78] For agencies using algorithmic enforcement tools, the challenge is how to continually and iteratively update them to capture new modes of wrongdoing.[79]

4. Implications

Development and deployment of algorithmic tools at the SEC and other agencies hold significant implications for the future. Two broad implications dominate.

First, the new algorithmic enforcement tools hold important implications for the accountability of agency enforcement activities. It remains unclear whether the new tools will *degrade* or *enhance* legal and political accountability relative to the status quo. On the one hand, the technical opacity and "black box" nature of the more sophisticated AI-based tools may erode overall accountability by rendering agency enforcement decisions even more inscrutable than the human judgments of dispersed agency enforcement staff. But the opposite might also prove true: formalizing and making explicit agency priorities could render an agency's enforcement decision making relatively more tractable compared to pools of agency enforcement staff. We return to these possibilities in Part II's exploration of administrative law's response to the new algorithmic governance tools.

77. Danielle Ensign et al., *Runaway Feedback Loops in Predictive Policing* (Dec. 22, 2017), https://arxiv.org/abs/1706.09847 [https://perma.cc/4APU-N8QK]. For a more general survey of potential problems with predictive policing, see Andrew Guthrie Ferguson, *Policing Predictive Policing*, 94 WASH U.L. REV. 1109 (2017).

78. Erik Hemberg et al., *Tax Non-Compliance Detection Using Co-Evolution of Tax Evasion Risk and Audit Likelihood*, ICAIL '15 (2015), https://taxprof.typepad.com/files/taxpaper.pdf [https://perma.cc/Y3BR-M4AR].

79. *Cf.* Deven R. Desai & Joshua A. Kroll, *Trust but Verify: A Guide to Algorithms and the Law*, 31 HARV. J.L. & TECH. 1, 21 (2017) (noting that systems require "ongoing monitoring and evaluation to ensure the model remains accurate given that the real-world changes").

Second, AI-based enforcement tools may work a fundamental change in the structure and legitimacy of the regulatory state. Algorithmic enforcement tools can, by reducing agency search costs, facilitate more robust enforcement activity by permitting agencies to identify regulatory targets more efficiently and shift scarce resources away from regulatory search and toward prosecution. These tools can also serve as force multipliers, narrowing the public-private technology gap and helping to level the playing field between underfunded agencies and well-resourced regulated parties. They could hence halt or even reverse the decades-long shift away from public enforcement and toward private litigation as a regulatory mode.[80] Indeed, one explanation for the shift from public to private enforcement, achieved largely via legislative creation of private rights of action, was a fiscally focused legislative desire to move enforcement costs from on-budget to off-budget forms.[81] A significant reduction in regulatory search costs could alter that core legislative calculus.

That said, as algorithmic tools move closer to the core of the state's coercive power, they may systematically shift patterns of state action in ways that raise distributive and, ultimately, political anxieties about a newly digitized public sector. One broad concern centers on the potential distributive effects of algorithmic enforcement, particularly from gaming.[82] Well-heeled regulated parties may be better able than their less advantaged peers to reverse-engineer an agency's algorithmic tools and take actions to avoid or even foil detection. As just one example, major investment banks may be more likely to have a stable of sophisticated employees with computer science and quantitative training who can reverse-engineer the SEC's algorithmic tools, thus shielding their own registrants from agency enforcement efforts.[83] If citizens come to believe AI systems are rigged, political support for a more effective, tech-savvy government could quickly evaporate.

Still another concern is that the advent of algorithmic enforcement will supplant expertise within the bureaucracy, exacerbating a perceived trend toward politicized administration and the hollowing out of the expertise of the

80. SEAN FARHANG. THE LITIGATION STATE: PUBLIC REGULATION AND PRIVATE LAWSUITS IN THE UNITED STATES (2010).

81. Sean Farhang, *Public Regulation and Private Lawsuits in the American Separation of Powers System*, 52 AM. J. POL. SCI. 821, 823-28 (2008) (reviewing the debate); David Freeman Engstrom & David Hausman, *Rights, Redistribution, and the Rise of the "Litigation State": The Case of Disability Discrimination Laws*, 45 LAW & SOC. INQUIRY (forthcoming 2020) (testing the theory).

82. *See* David Freeman Engstrom et al., *Enforcement by Algorithm* (June 2020) (unpublished manuscript) (on file with author); *see also* Jane R. Bambauer & Tal Zarsky, *The Algorithm Game*, 94 NOTRE DAME L. REV. 1, 10 (2018) (exploring more general phenomenon of "gaming," in which a clever adversary identifies and then exploits weaknesses in an algorithmic system).

83. A further example we discuss in more detail below comes from the adjudication side of things: A firm that knows that the PTO is using "deep learning" to detect similar trademarks could, in theory, develop an adversarial model to fool trademark examiners into thinking that a trademark is distinctive.

administrative state.[84] This is especially worrying because, at least for the moment, line-level enforcers can play a key role in bolstering accountability. Currently, use of algorithmic enforcement tools by SEC enforcement staff is entirely voluntary. As a result, agency technologists must sell skeptical line-level staff on their tools and provide user-friendly interfaces. Perhaps more importantly, SEC technologists report that line-level enforcement staff are often unmoved by a model's sparse classification of an investment advisor, based on dozens of pages of disclosures, as "high risk."[85] They want to know which section of a disclosure triggered the classification and why. This has further pressed agency technologists to focus on explainability in their models. Staff skepticism thus raises the interesting possibility that governance of public sector AI tools will at times come from "internal" due process, not the judge-enforced, external variety.[86] That said, SEC officials could change the agency's current approach, making use of the tools mandatory or keying agency decisions primarily or entirely to machine outputs. To that extent, the SEC could quickly increase the centrality and significance of the tools and sharpen accountability concerns.

II. Current Administrative Law and the Puzzle of Algorithmic Accountability

The new algorithmic governance tools like those on display at the SSA and SEC trigger a sharp collision. On the one hand, the body of law that governs how agencies do their work is premised on transparency, accountability, and reason giving.[87] When government takes action that affects

84. On the risks of hollowing out, see PETER H. SCHUCK, WHY GOVERNMENT FAILS SO OFTEN: AND HOW IT CAN DO BETTER (2014); and PAUL VERKUIL, VALUING BUREAUCRACY: THE CASE FOR PROFESSIONAL GOVERNMENT (2017).

85. GOVERNMENT BY ALGORITHM, *supra* note 17, at 28.

86. To that extent, bureaucratic implementation of algorithmic enforcement tools may roughly resemble a dynamic noted by others in which the interactions of internal and sometimes "rivalrous" bureaucratic actors shape agency behavior. *See* Neal Kumar Katyal, *Internal Separation of Powers: Checking Today's Most Dangerous Branch from Within*, 115 YALE. L.J. 2314 (2006); Amanda Leiter, *Soft Whistleblowing*, 48 GA. L. REV. 425, 429 (2014); Gillian E. Metzger, *The Interdependent Relationship Between Internal and External Separation of Powers*, 59 EMORY L.J. 423 (2009); Jon. D. Michaels, *Of Constitutional Custodians and Regulatory Rivals: An Account of the Old and New Separation of Powers*, 91 N.Y.U. L. REV. 227 (2016).

87. In the American context, this norm pervades administrative law, both in the Administrative Procedure Act, *see* 5 U.S.C. § 557(c)(3)(A) (2018) ("All [agency] decisions [with respect to procedures requiring a hearing] . . . shall include a statement of . . . findings and conclusions, and the reasons or basis therefor . . . ."), and in judicial decisions, *see Judulang v. Holder*, 565 U.S. 42, 45 (2011) ("When an administrative agency sets policy, it must provide a reasoned explanation for its action."); *FCC v. Fox Television Stations, Inc.*, 556 U.S. 502, 515 (2009) (noting "the requirement that an agency provide reasoned explanation for its action"). Similar versions can be found in many Western legal systems. For a review, see Henrik Palmer Olsen et al., *What's in the Box? The Legal Requirement of Explainability in Computationally Aided Decision-Making in Public Administration* 14-22 (iCourts Working Paper Series No. 162, 2019).

rights, it must explain why. On the other hand, the most advanced forms of AI are not yet, by their structure, fully explainable.[88]

This Part addresses this core collision. It begins by reviewing an emerging algorithmic accountability literature and highlights that literature's shortcomings, particularly its failure to reckon in any sustained manner with the possibilities and limits of conventional administrative law in achieving meaningful accountability. It then makes a start toward more sustained attention to administrative law as the front-line regulator of AI-based governance tools, following administrative law's foundational distinction between ex post and ex ante review of agency action and detailing some of the common and also distinctive legal puzzles that arise when applied to adjudication and enforcement. It concludes by mining Part I's technical and operational descriptions of the SSA's and SEC's new algorithmic tools to build an account of the non-legal, informational challenges of APA-based review of algorithmic agency action.

Throughout we make the case that judicial review of agency action, at least in its current guise, is unlikely to yield systematic, as opposed to pocketed and even idiosyncratic, oversight of agency use of algorithmic decision making tools. Under existing interpretations and doctrines, administrative law will provide few systematic incentives for agency administrators to improve internal administration and, at best, yield a checkerboard system of external accountability.

A. The Existing Algorithmic Accountability Literature

A fast-growing academic literature explores the clash between algorithmic opacity and legal demands of accountability and reason giving, much of it through the lens of constitutional due process. That high-level framing, with its focus on balancing the private interest, the government interest, and the marginal value of additional process, has spawned inquiry along two distinct tracks.

The first track asks what level of transparency into an algorithmic system's workings is necessary to gauge the system's fidelity to law. It starts from a well-established pair of ideas. One is that advanced machine learning outputs are *inscrutable* in the sense that even their own engineers cannot necessarily understand how the most advanced models arrived at a given result.[89] Machine learning outputs are also often *non-intuitive* in that the rules they derive to make predictions are so complex and multi-faceted that they defy practical inspection or do not comport with any practical human belief about

88. *See*, *e.g.*, Jenna Burrell, *How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms*, 3 BIG DATA & SOC'Y 1 (2016).

89. *See* Andrew Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085, 1094-96 (2018). For a highly accessible version, see JUDEA PEARL & DANA MCKENZIE, THE BOOK OF WHY: THE NEW SCIENCE OF CAUSE AND EFFECT 359 (2018).

how the world works.[90] As a result, even perfect transparency into an algorithmic system—that is, unfettered access to its source code and data and the chance to observe its operation "in the wild"[91]—may not yield accountability in the sense of rendering decisions fully legible to data subjects or surfacing all of a system's flaws.[92] Indeed, we know surprisingly little about why the most advanced neural networks work, although much active work exists to provide explainable AI.[93] Instead, desired transparency may only be approximated by mixing and matching multiple, partial modes of explanation. That is, one might be able to gain insight into a system's operation by combining a "decision-level" accounting of a given decision's "provenance" via the machine's inputs and outputs with a "system-level" accounting of the

90. *See* Selbst & Barocas, *supra* note 89, at 1096-99.

91. *See* Aaron Rieke, Miranda Bogen & David G. Robinson, *Public Scrutiny of Automated Decisions: Early Lessons and Emerging Methods*, UPTURN & OMIDYAR NETWORK 19 (2018), https://www.omidyar.com/sites/default/files/file_archive/Public%20Scrutiny%20of%20Automated%20Decisions.pdf [https://perma.cc/2755-G9B8].

92. Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 661 (2017) (noting that input-output testing—that is, basic "black box testing"—is "least powerful" among testing methods because of the inability to attribute a cause to a change in output or gauge its significance). For a more general version of the point, see Ananny & Crawford, *supra* note 21, at 980 ("Seeing inside a system does not necessarily mean understanding its behavior or origins."); and *id*. at 981 (noting that the "ephemeral nature of computational representations" may be incompatible with transparency). On the insufficiency of code alone, see Christian Sandvig et al., *Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms, Data and Discrimination: Converting Critical Concerns in Productive Inquiry*, 64TH ANN. MEETING OF THE INT'L COMM. ASS'N (May 22, 2014), http://www-personal.umich.edu/~csandvig/research/Auditing%20Algorithms%20—%20Sandvig%20—%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf [https://perma.cc/CTE7-L4UK]. Most agree that transparency requires, at a minimum, a description of a decision's "provenance," including an accounting of its inputs and outputs and the main factors that drove it. A more robust accounting of a decision's provenance would also convey the minimum change necessary to yield a different outcome and provide explanations for similar cases with different outcomes and different cases with similar outcomes. *See* Finale Doshi-Velez & Mason Kortz, Accountability of AI Under the Law: The Role of Explanation (Berkman Klein Ctr. Working Paper, 2019), https://dash.harvard.edu/bitstream/handle/1/34372584/2017-11_aiexplainability-1.pdf [https://perma.cc/D63K-8ELW]. However, while emerging techniques are rendering machine learning models more interpretable by ranking, sorting, and scoring data features according to their impact in the model, or using visualization techniques or textual justifications to lay bare a model's decision "pathway," challenges remain, especially with larger, multi-dimensional models. For a recent review of this highly active research area, see Ashraf Abdul, Jo Vermeulen, Dandling Wang, Brian Y. Lim & Mohan Kankanhalli, *Trends and Trajectories for Explainable, Accountable, and Intelligible Systems, An HCI Research Agenda*, CHI '18 (2018), https://dl.acm.org/doi/10.1145/3173574.3174156 [https://perma.cc/C2GS-PH25]. Another approach to interpretability uses visualization techniques or machine-based textual justifications to lay bare a model's decision "pathway." *See* L.A. Hendricks et al., *Generating Visual Explanations*, EUR. CONF. ON COMPUTER VISION 3-19 (2016); Chris Olah et al., *The Building Blocks of Interpretability*, DISTILL (Mar. 6, 2018), https://distill.pub/2018/building-blocks [https://perma.cc/HA2L-8MCP]. That said, input-output analysis need not be technical. Some advocate interactive "tinker" interfaces that allow data subjects to manually enter and change data and observe results, yielding a "partial functional feel for the logic of the system." Selbst & Barocas, *supra* note 89, at 1116.

93. Wojciech Samek, Thomas Wiegand & Klaus-Robert Müller, *Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models* (2017), https://arxiv.org/abs/1708.08296 [https://perma.cc/4BRM-V3XZ].

tool's "purpose, design, and core functioning,"[94] such as data descriptions, modeling choices, and the like.[95]

The second track in the literature tours the mechanisms that regulatory architects might choose—in an ideal world, and without political constraints—in order to translate a given level of transparency into desired accountability. The typical result is a menu of regulatory possibilities that tracks the options available in any regulatory context. These include individual, rights-based measures (e.g., private lawsuits; whistleblower schemes that incentivize those with insider knowledge to surface misconduct; vesting data subjects with rights to notice, consent, correction, and erasure), more systemic modes of oversight (e.g., public regulation by a separate oversight agency, or an FDA-like licensing or certification scheme, before an algorithmic system deploys), and assorted other accountability-boosting measures, including "soft" rules (e.g., impact assessments).[96]

This literature has generated an initial set of insights about the accountability challenges of algorithmic governance. An example is Danielle Citron's observation that the test for procedural due process, which requires courts to focus on the case at hand and weigh the private interest, government interest, and likely value of additional process, may miss the fact that algorithmic tools are designed to operate at scale. Potentially lost in case-level balancing is the possibility that a one-time but costly increase in procedural

94. Selbst & Barocas, *supra* note 89, at 1099-1110 (offering an accessible explanation of the debate over "outcome-based" and "logic-based" explanations). For similar efforts to categorize explanation types, 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, 55-59 (2017) (distinguishing between "model-centric" and "subject-centric" explanations); and Sandra Wachter, Brent Mittlestadt & Luciano Floridi, *Why a Right to Explanation of Automated Decision-Making Does not Exist in General Data Protection Regulation*, 7 INT'L DATA PRIVACY L. 76 (2017) (distinguishing between explanations of "system functionality" and "specific decisions").

95. *See* RIEKE ET AL., *supra* note 91, at 18; Coglianese & Lehr, *Transparency*, *supra* note 20, at 20-22; Selbst & Barocas, *supra* note 89, at 1129; *see also* Katherine J. Strandburg, *Rulemaking and Inscrutable Automated Decision Tools*, 119 COLUM. L. REV. 1851 (2019) (noting the potential for rulemaking processes to memorialize modeling and other design choices). Beyond the transparency issue, a second foundational point made along the first track is that algorithms are not self-executing technical creations, but rather human-machine "assemblages." Ananny & Crawford, *supra* note 21, at 983; *see also* Citron, *supra* note 20, at 1264-66 (providing a taxonomy of "mixed systems"). Programmers must make myriad decisions, from how to partition the data, what model types to specify, what dataset, target variables (or class labels), and data features to use, and how much to tune the model. David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653, 683-700 (2017); *see also* Coglianese & Lehr, *Transparency*, *supra* note 20, at 12 (noting that algorithms are "repeatedly guided and nudged, but not dictated, by humans in the establishment and refinement of the algorithm"). For an accessible account of target variables, class labels, and data features, see L. Jason Anastasapoulos & Andrew B. Whitford, *Machine Learning for Public Administration Research, with Application to Organizational Reputation*, 29 J. PUB. ADMIN. RES. & THEORY 491 (2019). As a result, arbitrary or biased outputs can result from tainted code and data, but also from numerous other human-made design choices. *See* Barocas & Selbst, *supra* note 20, at 678; Kroll et al., *supra* note 92, at 679-82.

96. *See* Margot Kaminski, *Binary Governance*, 92 S. CAL. L. REV. 1529, 1564 (2019); *see also* Desai & Kroll, *supra* note 79, at 46; Tutt, *supra* note 25.

scrutiny of an algorithmic tool can yield massive social benefits across the thousands or millions of cases to which the tool is applied.[97]

However, the existing literature also falls short on several counts. For instance, most treatments abstract away from the technical and operational details of actual algorithmic tools. These analyses often commingle public and private sector AI use despite the very different logics, stakes, and legal imperatives governing each.[98] Both shortcomings have pushed much of the inquiry to a level of abstraction that lends itself to broad mappings of normative concerns rather than concrete regulatory solutions.

A third issue is a more crippling blind spot: a near-total lack of any sustained or close consideration of administrative law. This is concerning because administrative law, far more than constitutional law, will modulate agency use of algorithmic governance tools as they are incorporated into the work of government. The canon of constitutional avoidance—which holds that courts should avoid ruling on constitutional issues in favor of other, often statutory, grounds—means that administrative law, not constitutional law, will often be the legal constraint of first resort.[99] The virtual absence of administrative law from the emerging literature on algorithmic governance tools is also narrowing and even self-defeating. Administrative law's approach to the issues of transparency and reason giving that are fueling concerns about the new algorithmic governance is multi-faceted and tailored to particular governance tasks, providing a richer and as yet unexplored set of frames for assessing and resolving the accountability dilemmas in an increasingly digitized government.

The rest of this Part makes a start toward more sustained attention to administrative law as the front-line regulator of AI-based governance tools. It does so by following administrative law's foundational distinction between ex post and ex ante review of agency action, detailing some of the legal puzzles raised by each. It then mines Part I's effort to surface the technical and operational details of the SSA's and SEC's new algorithmic tools to build an account of the non-legal challenges of APA-based review of algorithmic agency action. Throughout we make the case that judicial review of agency decisions incorporating algorithmic analyses will not, without significant adaptation of existing administrative law doctrine, yield systematic external accountability.

97. Citron, *supra* note 20, at 1249.

98. *See supra* notes 21-22 and accompanying text.

99. Escambia County v. McMillan, 466 U.S. 48, 51 (1984) ("[N]ormally the Court will not decide a constitutional question if there is some other ground upon which to dispose of the case.").

B. Current Administrative Law: Ex Post Review of Algorithmic Decisions

Under current administrative law, ex post judicial review of agency action using AI is unlikely to yield systematic scrutiny. In the adjudication context, administrative law's various reviewability doctrines do not shield agency action, but the chance of ex post review of something like the QDD algorithm remains slim. In the enforcement context, by contrast, a thicket of reviewability and related doctrines largely insulate algorithmic decision making from judicial oversight. In what follows, we explore these difficulties to securing judicial review, but it is worth noting at the outside that they might also account for *why and where* AI innovation has transpired in federal agencies. Indeed, our interviews corroborate that strategic agency officials have piloted use cases precisely with insulation from judicial scrutiny in mind.[100]

1. Ex Post Review of Adjudicatory Decisions

While we will see that reviewability poses less of a concern in formal adjudication than in enforcement, the chances of judicial review of existing algorithmic decision tools like the SSA's QDD tool for detecting "easy grants" remain slim.

For QDD beneficiaries, the early grant consummates the agency's decision process. Such QDD beneficiaries are unlikely to challenge the QDD methodology and likely lack standing to do so. On the other hand, individuals who were not selected for the QDD process may be able to challenge SSA's decision once final, but harmless error may insulate scrutiny of the algorithm. In *Webb ex rel. Z.D. v. Colvin*, the appellant challenged an ALJ's refusal to consider reclassifying the case as a "critical case" for expedited processing because the ALJ misunderstood the Hearing Appeals and Litigation Law Manual.[101] The court found that the judge's failure to reclassify the case did not prejudice the ultimate benefits determination, rejecting the claim. A similar logic would likely govern review of the QDD model. As in the enforcement context, for litigants who lost their ultimate determination, their challenge to QDD would simply merge with the merits. In that posture, litigants are unlikely to focus much effort on the QDD algorithm itself. For litigants who won their final claim, but did not receive the benefit of QDD, the question is closer. They are the group most likely to have been misclassified by the algorithm and hence harmed by the time delay in receiving benefits. (If resources were fixed, the net effect on these claimants may indeed have been to *prolong* the benefits adjudication process.) While back pay would ultimately be awarded, the hardship to either (a) borrow money or (b) restrict consumption while claims

100. GOVERNMENT BY ALGORITHM, *supra* note 17, at 39.

101. No. 12-1059, 2013 WL 5020495, at *20 (N.D. Tex. Sept. 13, 2013) (quoting Newton v. Apfel, 209 F.3d 448, 459 (5th Cir. 2000)) ("[B]ecause 'HALLEX does not carry the authority of law,' the ALJ's error warrants remand only if Plaintiff's claim was prejudiced by the error.").

are pending could be serious. The litigants might hence have standing to challenge the implementation of QDD.

On the merits, *Mathews v. Eldridge*'s due process framework offers limited hope.[102] The private interest—the earlier receipt of benefits in the presence of backpay—may not be deemed large. Second, the probable value of additional process—e.g., the ability to probe the validity of the algorithm—may not be high, at least relative to the additional cost in governmental procedures. Providing all SSA applicants notice and the ability to probe the validity of the QDD algorithm, when experts would need to participate in hearings, would be costly. To be sure, a hearing that allows parties to scrutinize the algorithm could lead to system-wide improvements in accuracy, but the piecemeal appeals process for SSA decisions arguably (a) provides little incentive for litigants to bear that cost when challenging a claimant-specific error, and (b) requires aggregating government cost of additional procedures to scrutinize algorithms. Indeed, subjecting all AI tools to opportunity for full-blown interrogation could undercut the incentive to adopt algorithmic decision tools in the first instance.

While procedural due process may not be well suited, litigants may challenge QDD—or the clustering tool or the Insight system—under standard APA review for adherence to the enabling act and for arbitrariness and capriciousness. Yet under such merits review, courts run into significant information challenges that we document further below.

2. Ex Post Review of Enforcement Decisions

In contrast to formal adjudication, modern administrative law erects substantial barriers to ex post judicial review of enforcement decisions.[103] Selective prosecution—i.e., the assertion that another entity is just as bad or worse, or "why me and not them"—is a non-starter absent constitutionally recognized racial or other bias.[104] Moreover, under the APA, courts generally

102. 424 U.S. 319 (1976).

103. Lisa S. Bressman, *Judicial Review of Agency Inaction: An Arbitrariness Approach*, 79 N.Y.U. L. REV. 1657 (2004); *see also* Van Loo, *supra* note 23, at 378 (noting that "regulatory monitors operate in the 'soft' administrative law space largely exempted from the APA's accountability mechanisms"). For a more general argument that administrative law focuses primarily on rulemaking and adjudicative hearings and thus misses large tranches of administrative action, see Edward Rubin, *It's Time to Make the Administrative Procedure Act Administrative*, 89 CORNELL L. REV. 95, 106-09 (2003); William H. Simon, *The Organizational Premises of Administrative Law*, 78 LAW & CONTEMP. PROBS. 61, 70-71 (2015).

104. Reno v. Am.-Arab Anti-Discrimination Comm., 525 U.S. 471, 489 (1999) (noting that selective prosecution claims are a "rara avis," and finding that the concerns that underscored its holding in Armstrong were "magnified" in the deportation context); United States v. Armstrong, 517 U.S. 456, 464-65 (1996) (articulating a strong presumption of regularity in prosecutorial decisions and requiring a defendant claiming selective prosecution to show discriminatory purpose and that the state's action was "'directed so exclusively against a particular class of persons . . . with a mind so unequal and oppressive' that the system of prosecution amounts to 'a practical denial' of equal protection of the law") (quoting Yick Wo v. Hopkins, 118 U.S. 356, 373 (1886)); Marshall v. Jerrico, Inc., 446 U.S. 238, 249 (1980) (noting that "traditions of prosecutorial discretion do not immunize from judicial scrutiny

lack jurisdiction to review an agency's decision whether or when to enforce. In the doctrine's standard formulation, a federal agency's decision to initiate a civil enforcement action is, like a criminal prosecutor's charging decision, insulated from judicial review as a core executive responsibility committed to agency discretion by law.[105] We consider first the case of a challenge to a non-enforcement decision (typically by a beneficiary of enforcement) and then a challenge to an enforcement decision (typically by the enforcement target itself).

a) Regulatory Beneficiaries—Challenging Agency Non-Enforcement

The principle that agency enforcement decisions should be insulated from judicial review extends to both agency decisions to enforce and *not* enforce, but it has particular force in the latter context, as when regulatory beneficiaries seek to compel rather than block agency action.[106] The well-known doctrinal fountainhead is *Heckler v. Chaney*, in which the Court created a strong presumption against review that can be rebutted only under narrow circumstances.[107]

The first exception, articulated a decade before *Chaney* in *Dunlop v. Bachowski*,[108] triggers when Congress has articulated guidelines for the agency's exercise of its enforcement authority by making enforcement mandatory ("shall enforce") coupled with a standard against which to judge agency refusals to do so.[109] Federal statutes meeting *Dunlop*'s requirements are rare, but, where they exist, an agency's use of an algorithmic tool can plainly be reviewed for its fidelity to congressional command. The resulting review can

cases in which the enforcement decisions of an administrator were motivated by improper factors or were otherwise contrary to law," but then making clear that judicial concern will be limited to the context of the "financial or personal interest on one who performs a prosecutorial function"). Short of this, only a class-of-one, rational-basis challenge is possible. *See* Vill. of Willowbrook v. Olech, 528 U.S. 562, 564 (2000); *see also* Engquist v. Or. Dep't of Agric., 553 U.S. 591, 601 (2008) (citing *Olech* and noting that the standard for a "class of one" equal protection challenge is that the plaintiff was "intentionally treated differently from others similarly situated and that there is no rational basis for the difference in treatment"). For analysis of the interplay between selective prosecution claims and claims under the APA, see *United States v. American Electric Power Service Corp.*, 258 F. Supp. 2d 804, 806 (S.D. Ohio 2003).

105. 5 U.S.C. § 701(a)(2) (2018) (prohibiting review of questions that are "committed to agency discretion").

106. For an overview of administrative law doctrine that "forestall[s] challenges to systemic nonenforcement and agency inaction," see Gillian E. Metzger, *The Constitutional Duty to Supervise*, 124 YALE L.J. 1836, 1872 (2015).

107. Heckler v. Chaney, 470 U.S. 821 (1985). Lower courts have extended the *Chaney* principle to pre-enforcement monitoring activities. *See, e.g.*, Gillis v. U.S. Dep't of Health & Human Servs., 759 F.2d 565, 576 (6th Cir. 1985) (finding HHS decision not to monitor hospitals' provision of reduced cost services was not reviewable); Madison-Hughes v. Shalala, 80 F.3d 1121, 1129-31 (6th Cir. 1996) (finding HHS decision not to collect data on race disparities was not reviewable because committed to agency discretion).

108. 421 U.S. 560 (1975).

109. *Chaney*, 470 U.S. at 832-33 (noting lack of review unless Congress "has provided guidelines for the agency to follow in exercising enforcement powers."); *see also Bachowski*, 421 U.S. at 566-68.

be thorough: courts regularly require agencies to provide a full explanation,[110] including how the agency assessed the importance of specific considerations[111] or why it departed from prior practice.[112] At least in the relatively few pockets of the federal code where an agency's organic statute presents the requisite mandate-plus-standards, a court can require an agency to explain the structure or even the precise specification of an algorithmic enforcement tool.[113]

The remaining exceptions, however, are narrower and offer only weak and irregular prospects of rebutting *Chaney*'s presumption of unreviewability. One exception triggers when an agency has adopted a policy of nonenforcement that rises to the level of an "abdication" of its statutory responsibilities.[114] Note, however, that this is not typically a free-standing exception. Rather, instances of abdication are reviewable only because the statute, in commanding that an agency safeguard the public health or safety, might thereby indicate that the agency lacks discretion to adopt a wholesale policy of nonenforcement.[115] The focus, as in *Dunlop*, remains on whether there are sufficient indicia of legislative intent to rebut the presumption of non-reviewability. Moreover, courts have been reluctant to find abdication in cases where the agency is

110. *See, e.g.*, Sullivan Indus. v. NLRB, 957 F.2d 890, 905 n.12 (D.C. Cir. 1992) ("Until the Board explains itself, we have no way of reviewing the Board's actions for consistency or rationality and no way of keeping our own precedents in harmony."); United States Dep't of Def. v. FLRA, 982 F.2d 577, 580 (D.C. Cir. 1993) ("Because the record and the FLRA's explanation for its decision are insufficient to support judicial review, the case is remanded to FLRA.").

111. *See, e.g.*, Sw. Pub. Serv. Co. v. FERC, 952 F.2d 555, 563 (D.C. Cir. 1992); Charlottesville v. FERC, 661 F.2d 945, 954 (D.C. Cir. 1981); City Fed. Sav. & Loan Ass'n v. FHLBB, 600 F.2d 681, 689 (7th Cir. 1979); City Nat'l Bank v. Smith, 513 F.2d 479 (D.C. Cir. 1975).

112. *See, e.g.*, Atchison, T. & S.F. Ry. Co. v. Wichita Bd. of Trade, 412 U.S. 800, 809 (1973) ("[I]t is enough to satisfy the requirements of judicial oversight of administrative action if the agency asserts distinctions that, when fairly and sympathetically read in the context of the entire opinion of the agency, reveal the policies it is pursuing."); Phila. Gas Works v. FERC, 989 F.2d 1246, 1247 (D.C. Cir. 1993) ("Of course, FERC can consider new facts and circumstances to limit *North Penn* and is entitled to weigh 'equitable' considerations as it thinks appropriate. But it must identify the facts, circumstances, and equitable factors on which it relies."); Pittsburgh Press Co. v. NLRB, 977 F.2d 652, 662 (D.C. Cir. 1992); Hatch v. FERC, 654 F.2d 825, 834 (D.C. Cir. 1981) (finding that FERC must provide "a reasoned explanation for any . . . failure to adhere to its own precedents").

113. That said, some courts have questioned how much proprietary information agencies might be required to disclose. A good example is *Corner v. Harris*, 519 F. App'x 942, 943 (7th Cir. 2013), a case challenging agency non-enforcement under the Labor–Management Reporting and Disclosure Act, which requires the secretary to "file suit if there is probable cause to believe that a violation of federal law probably affected the outcome of the election." *Id.* at 943. In declining to compel enforcement, the court there noted that "even federal statutes that, unlike § 402, create enforceable rights of access to information, have exceptions for material gathered in the course of pre-litigation investigations." *Id.* And it reasoned that "*Bachowski* was clear that the Secretary must give *reasons,* not open the agency's files to disclose whatever evidence the complainant desires to see." *Id.* After all, a "prosecutor (the Secretary occupies a prosecutorial role) needs to be able to promise confidentiality in order to gather information—especially when there is a deadline that may prevent resort to compulsory process." *Id.*

114. *Chaney*, 470 U.S. at 833 n.4.

115. *See, e.g.*, Riverkeeper, Inc. v. Collins, 359 F.3d 156, 168-69 (2d Cir. 2004) (noting that the court could find abdication "only if [the NRC] has established a policy not to protect . . . public health and safety," and foreclosing such a finding by listing the various measures the NRC took to that end).

engaged in at least some enforcement activity.[116] So long as an algorithmic tool does not foreclose enforcement entirely, and merely pares down the universe of targets, the exception is not typically triggered.

A creative route around *Chaney* would exploit a possible ambiguity in the case's normative foundation. Beyond *Chaney*'s separation-of-powers framing of enforcement as a core executive responsibility,[117] the doctrine is generally understood to be driven by the complexity and technical nature of enforcement decisions.[118] A key question is whether this focus on complexity sounds in deference or indeterminacy. At the core of a deference-based reading is comparative expertise: generalist judges should not second-guess expert administrators on how best to achieve regulatory goals, particularly where that determination turns on the optimal allocation of scarce agency resources.[119] That reading enjoys substantial support in *Chaney* itself,[120] and it is hard to see

116. *See, e.g.*, Citizens for Responsibility & Ethics in Washington v. Fed. Election Comm'n, 892 F.3d 434, 440 n.9 (D.C. Cir. 2018) ("CREW cites this footnote but its own submissions show that the Commission routinely enforces the election law violations alleged in CREW's administrative complaint."); Am. Disabled for Attendant Programs Today v. U.S. Dep't of Hous. & Urban Dev., No. 96-5881, 1998 WL 113802, at *3 (E.D. Pa. Mar. 12, 1998), *aff'd*, 170 F.3d 381 (3d Cir. 1999) (declining to entertain "broad-gauged review of HUD's entire agency-initiated enforcement program (or lack thereof), sought prior to any apparent recourse by plaintiffs to the privately-initiated administrative enforcement schemes").

117. A further normative foundation is the idea, however sound, that agency inaction is not as coercive an exercise of state power as agency action. *See* Balt. Gas & Elec. Co. v. F.E.R.C., 252 F.3d 456, 459 (D.C. Cir. 2001); *see also* Crowley Caribbean Transp., Inc. v. Pena, 37 F.3d 671, 675 (D.C. Cir. 1994) (declining to carve "reviewable legal rulings out from the middle of non-reviewable actions"). Courts have found normative foundations in other places as well, looking to pragmatic factors such as "the need for judicial supervision to safeguard the interests of the plaintiffs; the impact of review on the effectiveness of the agency in carrying out its congressionally assigned role; and the appropriateness of the issues raised for judicial review." Nat. Res. Def. Council, Inc. v. Sec. & Exch. Comm'n, 606 F.2d 1031, 1044 (D.C. Cir. 1979) (citing Hahn v. Gottlieb, 430 F.2d 1243 (1st Cir. 1970)); *see also* Nat'l Treasury Emp. Union v. Horner, 854 F.2d 490, 497 (D.C. Cir. 1988).

118. *See Baltimore Gas & Electric*, 252 F.3d at 459. In addition to the above discussion, complexity and technicality can mean many things. One commonly cited discussion is Judge Oakes concurrence in *Dina v. Attorney General of United States*, 793 F.2d 473, 477 (2d Cir. 1986), which notes that nonreviewability is determined primarily by the fact that it is "hard to review" cases without appropriate guidance. *See* Chong v. Dir., U.S. Info. Agency, 821 F.2d 171, 177 (3d Cir. 1987) (citing *Dina* as authority for this proposition). For this reason, Judge Oakes rejects the notion that it applies to a relatively narrow category of cases. Of course, this justification also sounds in comparative expertise. *See, e.g.*, NAACP v. Trump, 298 F. Supp. 3d 209, 227 (D.D.C. 2018), *adhered to on denial of reconsideration*, 315 F. Supp. 3d 457 (D.D.C. 2018). Note that comparative expertise will not always carry the day when the below-mentioned considerations are present. *See, e.g.*, Salazar v. King, 822 F.3d 61, 76 (2d Cir. 2016) (declining to apply *Chaney* where an agency's decision involves a "complicated balancing of factors," but where the agency was exercising its "coercive power").

119. It is also the case that an agency may pursue multiple goals simultaneously, such as maximizing its win rate, maximizing the total amount of fines or other sanctions, or achieving what the agency sees as an optimal, or congressionally specified, level of enforcement effort or deterrence. *See* David Freeman Engstrom, *Public Regulation of Private Enforcement: Empirical Analysis of DOJ Oversight of Qui Tam Litigation Under the False Claims Act*, 107 NW. U. L. REV. 1689, 1703 (2013) (noting different agency objective functions and "maximands" when engaged in enforcement-related decision making).

120. As the *Chaney* Court itself noted, an agency's enforcement decision "involves a complicated balancing of a number of factors which are peculiarly within [agency] expertise," *Heckler v. Chaney*, 470 U.S. 821, 831 (1985), making the agency "far better equipped than the courts to deal with the many variables involved in the proper order of its priorities." *Id*. at 831-32.

how an agency's use of an algorithmic tool to optimize its allocation of scarce resources could disturb this principle. But the picture is different if *Chaney* instead sounds in indeterminacy—that is, whether the precise grounds for agency non-enforcement decisions are knowable at all, and thus whether anyone, expert or otherwise, can reliably reconstruct the agency's decision in a particular case. This alternative reading also draws support in *Chaney* in a key passage where the Court contrasts affirmative agency action, which provides "a focus for judicial review," with agency failures to act, which often do not.[121]

Interestingly, recentering *Chaney* around indeterminacy could lead courts in either direction on reviewability. On one hand, algorithmic tools must typically specify an objective function formalizing the agency's priorities. This can potentially convert a line-level enforcer's opaque, all-things-considered weighing of factors into a rule-bound and tractable calculus, providing the focal point and foothold for judicial review found missing in *Chaney*. On the other hand, a complex NLP-based machine learning tool of the sort the SEC utilizes may, because of technical opacity, be *more* indeterminate than even the efforts of line-level enforcement staff to work up cases in the analog way using a multi-factor protocol or other guidance document. The result seems paradoxical: algorithmic tools that are *more* intelligible are subject to review, those that are *less* so, or even fully opaque, are insulated from it. We offer a fuller discussion of questions arising from the dynamic and adaptive nature of certain machine learning tools below.

Finally, lower courts have entertained still other innovative paths around *Chaney*. First, some courts have held that an agency can, by adopting a policy statement or formal or informal guidelines imposing binding limitations on the exercise of its own enforcement discretion, provide the necessary law against which to measure its failure to initiate enforcement.[122] Other courts, however, have expressed skepticism as to whether a mere policy statement, as opposed to a properly promulgated legislative rule, can provide the necessary law to apply.[123] This latter position, it should be noted, would permit review of agency use of algorithmic enforcement tools only in situations in which the tool has already been subject to ventilation via notice and comment and so might not appreciably increase accountability. Second, lower courts have worked around *Chaney* by casting an agency's enforcement decision as a general policy or rule rather than a particularized action, especially when it involves the application of a "permanent standard" or a rule that is "mechanical" in form.[124] There may,

121. *Id.* at 832; *see also* Texas v. United States, 809 F.3d 134 (5th Cir. 2015) (reasoning that deferred agency action is "affirmative agency action" because it confers lawful presence and employment authorization on a large class of people who would otherwise be removable).

122. GoJet Airlines v. FAA, 743 F.3d 1168 (8th Cir. 2014); Socop-Gonzalez v. INS, 208 F.3d 838 (9th Cir. 2000); Chong v. Dir., U.S. Info. Agency, 821 F.2d 171, 176 (3d Cir. 1987).

123. *See* Mass. Pub. Interest Research Grp. v. U.S. Nuclear Regulatory Comm'n, 852 F.2d 9 (1st Cir. 1988).

124. Edison Elec. Inst. v. EPA, 996 F.2d 326, 333 (D.C. Cir. 1993) (finding a policy statement reviewable on this basis); *see also* Kenney v. Glickman, 96 F.3d 1118, 1123 (8th Cir. 1996);

however, be limits to this logic: the Supreme Court has pointedly rejected judicial review proceedings that level a "broad programmatic attack" at an agency's administration of its statute or otherwise seek "wholesale improvement" of an agency's programmatic activities rather than focusing on particular agency actions that cause particularized harm.[125] Note as well that treating an algorithm as a rule in order to take the case outside *Chaney*'s ambit raises further and distinct questions of reviewability, particularly the availability of pre-enforcement review. The possibility that an algorithmic enforcement tool constitutes a rule, with all that such a designation would entail under the APA, is taken up more fully below.

Future research can and should parse this dense web of case law more deeply.[126] For now, however, it seems safe to conclude that, save situations in which the mandatory framing of an agency's organic statute brings it within *Dunlop*'s domain, or a judicial willingness to recenter *Chaney* around agency self-cabining or indeterminacy, agency use of algorithmic enforcement tools will remain substantially insulated from judicial challenges by regulatory beneficiaries or other non-targets seeking to compel agency enforcement.

b) Regulatory Targets—Challenging Agency Enforcement

Chaney's presumption against reviewability is flipped when judicial review of an algorithmic enforcement tool is sought by an enforcement target itself.[127] But even here, current administrative law erects substantial reviewability barriers that block the most likely avenues for judicial challenge, foiling anything resembling systematic review.

Arent v. Shalala, 70 F.3d 610, 614 (D.C. Cir. 1995) ("*Chaney* is of no assistance to the [agency] in this case because the [agency's] promulgation of a standard for 'substantial compliance' under the [Act] does not represent an enforcement action."); Nat'l Treasury Emp. Union v. Horner, 854 F.2d 490, 496 (D.C. Cir. 1988) ("OPM's decision to develop some but not other competitive examinations, in contrast, is a major policy decision, quite different from day-to-day agency nonenforcement decisions, or in its own context, from day-to-day personnel management decisions."); Capital Area Immigrant's Rights Coal. v. U.S. Dep't of Justice, 264 F. Supp. 2d 14, 24 (D.D.C. 2003) (per curiam) (declining to apply *Chaney* where "plaintiffs do not challenge any individual decision or agency enforcement action," but rather "general procedures for adjudicating immigration appeals"); *see also* Regents of the Univ. of California v. U.S. Dep't of Homeland Sec., 908 F.3d 476, 499 n.13 (9th Cir. 2018); OSG Bulk Ships v. United States, 132 F.3d 808, 812 (D.C. Cir. 1998); Crowley Caribbean Transp. v. Pena, 37 F.3d 671, 674-75 (D.C. Cir. 1994). To that extent, courts may be picking up on a point legal academics have made that enforcement occupies a kind of nether-space between rulemaking, which is typically general and prospective in form, and adjudication, which is individualized and retroactive in form. *See* Lemos, *supra* note 47, at 933 (noting that enforcement shares features of both – and involves both wholesale and retail decisions).

125. *See* Norton v. S. Utah Wilderness Alliance, 542 U.S. 55, 64 (2004); Lujan v. Nat'l Wildlife Fed'n, 497 U.S. 871, 891 (1990). For an argument that administrative unduly forestalls challenges to agency failures of "systemic administration," see Metzger, *supra* note 86.

126. *See* Engstrom et al., *supra* note 82.

127. *See* Bowen v. Mich. Acad. of Family Physicians, 476 U.S. 667, 670 (1986) (noting strong presumption in favor of reviewability of "final agency action by an aggrieved person . . . unless there is persuasive reason to believe that such was the purpose of Congress"); *see also* Citizens to Preserve Overton Park v. Volpe, 401 U.S. 402 (1971) (articulating strong presumption of reviewability of final agency action under the APA).

The main barrier extends from the Supreme Court's holding in *Standard Oil of California v. FTC* that an agency's decision to proceed with an enforcement action—that is, its decision to initiate an investigation, audit, or enforcement action—is non-final agency action and so not immediately challengeable.[128] In a key passage, the Court distinguished an agency's issuance of a complaint from the final rule at issue in *Abbott Laboratories* on the grounds that, in the latter, the FDA's rule had a substantial legal and practical effect on publicity-vulnerable pharmaceutical companies, who would otherwise, as the Court noted, be put to the Hobson's choice of costly compliance or a potentially ruinous public enforcement action. By contrast, the FTC's initiation of a complaint against Standard Oil had no similar impacts "other than the disruptions that accompany any major litigation."[129] Litigation costs, as the Court had put it several decades earlier, are "part of the social burden of living under government."[130]

For regulatory targets who seek to challenge an agency's use of an algorithmic enforcement tool, several implications follow. To begin, an enforcement target that believes it has been wrongly or arbitrarily identified by an algorithmic tool for investigation, audit, or enforcement cannot seek review of that decision on an interlocutory basis and instead must wait until the agency has brought its enforcement action to a conclusion.[131] At that point, a regulatory target that has mounted an unsuccessful defense, and thus found liable, could attempt to argue that even an agency enforcement action that is unassailable as a substantive matter is nonetheless voidable where the agency's process—including an upstream algorithmic process used to identify it as a regulatory target at the outset—was inconsistent with the agency's organic statute or implementing regulations. The APA specifically contemplates such actions via § 704's decree that a "preliminary, procedural, or intermediate agency action or ruling not directly reviewable is subject to review on the review of the final agency action."[132]

However, practical barriers remain. In cases in which the regulatory target was wrongly accused, the question of the propriety of the upstream use of the algorithm will, as a practical matter, merge with the substantive liability

128. FTC v. Standard Oil Co. of Cal., 449 U.S. 232, 242 (1980).

129. *Id*. at 243.

130. Petroleum Exploration, Inc. v. Pub. Serv. Comm'n, 304 U.S. 209, 222 (1938). Around the time of *Standard Oil*, the Court reiterated: "Mere litigation expense, even substantial and unrecoupable cost, does not constitute irreparable injury." Renegotiation Bd. v. Bannercraft Clothing Co., 415 U.S. 1, 24 (1974). At other times, the Court has paid lip service to the notion that being targeted for investigation or other enforcement action is costly. *See* Marshall v. Jerrico, Inc., 446 U.S. 238, 249 (1980) (noting that enforcement decisions can "result in significant burdens on a defendant or a statutory beneficiary, even if he is ultimately vindicated"). But the Court has never suggested that these costs are legally cognizable.

131. The analogy to interlocutory review is an apt one, as the *Standard Oil* Court noted: because an agency's issuance of a complaint will ultimately merge with an eventual decision on the merits, it would not qualify for interlocutory review under the collateral order doctrine. *Standard Oil*, 449 U.S. at 246.

132. 5 U.S.C. § 704 (2018).

question. Moreover, *Standard Oil*'s rejection of litigation costs as cognizable legal injury negates any possible recourse other than reversal on liability.[133] As a result, it is only cases in which a court upholds the agency's finding of substantive liability that will proceed to the question of the propriety of the agency's upstream use of the algorithm. But here, given *Standard Oil*'s clear rejection of litigation costs as a legally cognizable injury, a finding that the agency used an illegitimate means to reach a legitimate end can be dismissed as harmless error or would not provide grounds for recovery of damages. In short, neither scenario is likely to yield systematic review of an agency's algorithmic enforcement toolkit.

* * *

In short, conventional ex post judicial review of agency action under the APA is unlikely to generate systematic or even consistent review of the government's new algorithmic toolkit in either the enforcement or adjudication context.

C. Current Administrative Law: The Limits of Ex Ante Review

Another potential avenue for challenging agency use of algorithmic tools lies in characterizing the adoption of AI as a rule instead of a step in the adjudicatory or enforcement process. This path opens up two further potential mechanisms of accountability. One is the APA's requirement that "legislative" rules be subjected to notice and comment, which could apply to algorithms used for adjudication or enforcement. The other is pre-enforcement judicial review of an algorithmic tool in the enforcement context more specifically, before the tool is applied in a particular case and without the necessity of a violation. As with conventional ex post judicial review of agency action, however, these mechanisms still amount to a patchwork of accountability under current administrative law.

1. Algorithms as Legislative Rules Requiring Notice and Comment

A cornerstone of the APA's accountability regime is the requirement that "legislative" rules must be ventilated via notice and comment. That process requires that an agency explain what a proposed regulation is designed to achieve, solicit comments from interested parties, "consider[] . . . the relevant matter presented," and provide a "concise general statement of th[e] basis and

133. Other potential avenues of recourse are likewise unavailable. For instance, the Federal Tort Claims Act specifically withholds the Act's general waiver of sovereign immunity for malicious prosecution or abuse of process claims, carving out "[a]ny claim based upon an act or omission of an employee of the Government, exercising due care, in the execution of a statute or regulation, whether or not such statute or regulation be valid, or based upon the exercise or performance or the failure to exercise or perform a discretionary function or duty on the part of a federal agency or an employee of the Government, whether or not the discretion involved be abused." *See* 28 U.S.C. § 2680 (2018). This removes any possibility of common-law remedies.

purpose" of the rule.[134] As a practical matter, the "concise general statement" is often neither concise nor general due to judicial scrutiny under an arbitrary and capricious standard.

Not all rules,[135] however, qualify as legislative in nature. Lower courts have worked out a complicated doctrinal structure for sifting agency pronouncements that deserve the "legislative" label from those that are mere policy statements, rules of agency procedure or practice, or interpretative rules clarifying an agency's prior regulations.[136] Painting with a broad brush, these line-drawings variously distill to: (i) whether the rule has a binding effect on the agency, particularly line-level staff,[137] (ii) whether the rule substantially alters the rights and interests of regulated parties,[138] and (iii) the amount of regulatory work the rule does relative to the governing statute or prior agency-promulgated rules.[139] The resulting tangle of doctrines has been described as "tenuous," "blurred," "baffling," and "enshrouded in considerable smog."[140]

These characterizations alone may be enough to establish that notice and comment is unlikely to provide a systematic source of accountability, but a brief examination of cases implementing the tests helps drive home the point. As just one example, the extent to which an algorithm binds will turn in significant part on the degree to which there is a human in the loop—a question that is itself a highly subjective one and also likely to change with informal shifts in agency practice. But courts regularly characterize policies as legislative rules, even where substantial discretion remains with the agency and its line-level prosecutors.[141] An illustrative case is *McLouth Steel Prod. Corp. v. Thomas*,[142] where the court had to characterize a model used by the EPA to predict a company's levels of hazardous waste. The EPA argued that the model was not subject to notice and comment rulemaking because it was "not solely determinative of EPA's action" and was instead "one of many tools" used. Despite finding that the rule was "not ironclad" and that it in fact permitted exercise of agency discretion, the court found the model, upon close review, to

134. 5 U.S.C. § 553(c) (2018).

135. The APA capaciously defines a rule as "an agency statement of general or particular applicability and future effect designed to implement, interpret, or prescribe law or policy or describing the organization, procedure, or practice requirements of an agency." *See id.* § 551(4).

136. An agency might also claim that a rule is exempt from notice and comment under the APA's good cause exemption. *See id.* § 553(d)(2).

137. Cmty. Nutrition Inst. v. Young, 818 F.2d 943, 947 (D.C. Cir. 1987) (per curiam).

138. Nat'l Mining Ass'n v. McCarthy, 758 F.3d 243 (D.C. Cir. 2014) ("The most important factor concerns the actual legal effect (or lack thereof) of the agency action in question on regulated entities.").

139. Paralyzed Veterans of Am. v. D.C. Arena L.P., 117 F.3d 579 (D.C. Cir. 1997); Am. Mining Cong. v. Mine Safety & Health Admin., 995 F.2d 1106, 1110-11 (D.C. Cir. 1993).

140. *Community Nutrition*, 818 F.2d at 946.

141. *See, e.g.*, Guardian Fed. Sav. & Loan Ass'n v. FSLIC, 589 F.2d 658, 666–67 (D.C. Cir. 1978) (describing a legislative rule as a rule which "narrowly limits administrative discretion").

142. 838 F.2d 1317, 1319 (D.C. Cir. 1988).

be a legislative rule requiring notice and comment.[143] Other courts, however, refuse to apply a legislative label even where an agency pronouncement leaves no discretion at all. For instance, the D.C. Circuit refused to apply the legislative label to a series of agreements the EPA entered into with animal feeding operations in which the EPA promised not to bring enforcement actions pending the development of a methodology for measuring emissions.[144] Despite what amounted to a total cabining of enforcement discretion, the court reasoned that a narrow focus on discretion would extend the rule to nearly every consent agreement between an agency and a regulated entity.[145]

Against this uncertain backdrop, consider the SSA's expedited grant process (QDD). Recall that the adoption of QDD in fact went through notice and comment, because it required amendment of existing procedural rules.[146] Yet whether the proposal provided sufficient notice of the algorithmic decision tool is unclear.[147] SSA stated that the "predictive model . . . will score claims by taking into account such factors as medical history, treatment protocols, and medical signs and findings."[148] Claims would be subject to QDD if the model found a "high degree of probability" of a disability. No more detail was provided. On the one hand, key aspects of the model would seem to fit under the legislative rule rubric: the probability threshold would bind lower level officials (in the sense of removing cases from standard review to the QDD team) and a quick grant "substantially alter[s] the rights and interests" of regulated parties in light of the counterfactual delay of receipt of benefits.[149] On the other hand, discretion would still rest (a) in the QDD review team to decide a recommended quick grant, and (b) in adjudicators for all other cases. And one might argue that, despite the value of an earlier benefits determination, there is no alteration of rights in the sense that eligibility criteria are unchanged and claimants may receive backdated benefits payments if ultimately found eligible. In that sense, the QDD adoption resembles the medical-vocational guidelines (sometimes referred to as "the grid"), replacing case-by-case vocational expert judgment. The grid still allowed ALJs to deviate under certain circumstances, but were promulgated via notice and comment.[150] Under current administrative law, it is unclear whether SSA should have provided greater clarity about the QDD algorithm, but such operational details are critical to understanding its impact on the rights of beneficiaries.

143. *Id.* at 1319-23.

144. Ass'n of Irritated Residents v. EPA, 494 F.3d 1027, 1034 (D.C. Cir. 2007).

145. *Id*.

146. United States *ex rel*. Accardi v. Shaughnessy, 347 U.S. 260 (1954); Ariz. Grocery Co. v. Atchison, Topeka & Santa Fe Ry., 284 U.S. 370 (1932).

147. K.W. *ex rel*. D.W. Armstrong, 789 F.3d 962 (9th Cir. 2015) (finding notice lacking when the statistical budget calculation was altered).

148. Administrative Review Process for Adjudicating Initial Disability Claims, 71 Fed. Reg. 16,430 (Mar. 31, 2006).

149. *See* Am. Hosp. Ass'n v. Bowen, 834 F.2d 1037, 1041 (D.C. Cir. 1987).

150. *See* Heckler v. Campbell, 461 U.S. 458, 462 n.5 (1983).

Finally, requiring notice and comment for all algorithmic tools would be suboptimal. As we have shown above, the range of algorithmic decision tools is considerable. The SSA tool profiled previously that clusters cases for processing falls much more squarely within the ambit of a rule of internal agency organization, and there is nothing about the use of unsupervised learning in that setting that mandates notice and comment.[151] Moreover, our research into agency adoption of AI confirms that there is a considerable gap between private and public sector innovation. Notice and comment is a protracted process and, when combined with pre-enforcement review, can stymie innovation and prevent dynamic government responses to a changing policy problem or regulatory landscape. The use of technology itself is not a per se indicator of the kind of rule that necessitates notice and comment.

2. Algorithms and Pre-Enforcement Review

Pre-enforcement review of agency rules is available if a litigant can meet the familiar two-pronged test of fitness for judicial resolution and hardship.[152] Fitness is determined by whether the disputed claims raise a purely legal question and also the finality of the agency's decision, defined as whether the rule is the consummation of an agency process from which legal consequences will flow.[153] Hardship boils down to whether a rule's impact is sufficiently direct and immediate, which in turn asks whether the rule requires an immediate and significant change in the plaintiff's conduct of affairs with substantial penalties for noncompliance or otherwise imposes an injury that cannot be remedied upon review of an individual action.[154]

Some parts of the fitness inquiry do not pose a barrier to pre-enforcement review of algorithmic tools of the sort deployed by the SSA and SEC. So long as an agency's use of an algorithmic tool has advanced beyond the pilot stage, it plainly represents a final and settled agency position. Likewise, an agency's potential initiation of an enforcement action plainly rises to the level of a legal consequence. Whether an algorithm's propriety is a purely legal question, however, is a closer question. On one view, the output of an algorithmic tool is a prediction as to an ultimate legal outcome—for the SSA, whether a disability benefits case is a likely grant, or for the SEC, whether a broker is likely to be violating the securities laws. Facts serve solely as model inputs—the data features that drive the model—in generating that conclusion. Given this, the most common question upon review of an algorithmic tool—whether the tool's legal predictions fit within the substantive law that governs the agency's action—entails a purely legal comparison of the encoded, algorithmic rule and

151. *See* 5 U.S.C. § 553(b)(3)(A) (2018) (exempting from notice and comment "rules of agency organization, procedure, or practice").

152. Abbott Labs. v. Gardner, 387 U.S. 136 (1967).

153. Bennett v. Spear, 520 U.S. 154, 177-78 (1997).

154. *Abbott Labs*, 387 U.S. 136.

the statute's substantive liability standard. If, by contrast, the propriety of the rule turns on details of its bureaucratic implementation—for instance, the degree to which front-line enforcement or adjudicatory staff rely on it, and thus the extent to which a human remains "in the loop"—then the question is likely one of mixed law and fact, thus undermining the required fitness showing.

Other contours of the doctrine deepen the risk of a checkerboard of accountability. For instance, the hardship question as articulated by the Court in the *Abbott Labs*/*Toilet Goods* duo makes industry characteristics, not features of the rule itself, the most salient part of the analysis.[155] Algorithmic tools used to regulate the publicity-sensitive pharmaceutical industry will be more reviewable than tools used to regulate other industries. Still more variation in accountability is likely to arise out of the fierce debate among lower courts about whether ripeness doctrine should permit pre-enforcement challenges to non-legislative guidance documents[156] and procedural rules.[157] Several courts, for instance, have held that procedural challenges to policy pronouncements that were not promulgated as rules must await an agency effort to enforce the policy.[158] The famously blurry line dividing legislative rules and other types of agency pronouncements adds another way in which some algorithmic tools will qualify for pre-enforcement review while others will not.

D. The Informational Challenges of Conventional APA Review

Even if an algorithm were subjected to judicial review or notice and comment, substantial informational barriers impede review of algorithmic decision tools. The existing algorithmic accountability literature, we noted previously (in Section II.A), has begun to sketch an account of the regulatory challenges posed by the technical opacity of more sophisticated algorithmic tools. But Part I's concrete consideration of the technical and operational details of specific algorithmic governance tools, when combined with an understanding of the current landscape of administrative law, brings into relief a more concrete set of accountability challenges that go well beyond standard concerns about technical opacity.

---

155. For instance, the result in *Abbott Labs* arguably turns on the unique public relations vulnerability of a pharmaceutical company facing an FDA enforcement action.

156. Generally speaking, guidance documents are not open to pre-enforcement review. *See* Nat'l Park Hosp. Ass'n v. Dep't of Interior, 538 U.S. 803, 810 (2003) (policy statement not ripe); Florida Power & Light Co. v. EPA, 145 F.3d 1414, 1421 (D.C. Cir. 1998) (interpretative rule held not ripe.); Ciba-Geigy Corp. v. EPA, 801 F.2d 430, 434 (D.C. Cir. 1986) (policy statement not ripe); *see also* First Nat. Bank of Chi. v. Comptroller of Currency of U.S., 956 F.2d 1360 (7th Cir. 1992), *cert. denied*, 506 U.S. 830 (1992). However, there are exceptions. A leading case is *Aviators for Safe & Fairer Regulation, Inc. v. FAA*, 221 F.3d 222, 225 (1st Cir. 2000). There, the court found that where a notice was "final in a procedural sense," it could be ripe for pre-enforcement review. *Id.* This conclusion turned on the same ripeness analysis advanced in *Abbott Labs.* 387 U.S. at 148-49. It's sensible to think a similar approach could be taken towards the adoption of a new artificial intelligence program.

157. Abbs v. Sullivan, 963 F.2d 918 (7th Cir. 1992).

158. Pub. Citizen, Inc. v. U.S. Nuclear Regulatory Comm'n, 940 F.2d 679, 681-82 (D.C. Cir. 1991); Nat. Res. Def. Council, Inc. v. EPA, 16 F.3d 1395 (4th Cir. 1993).

First, public sector AI use implicates a different set of legal frameworks than applies to purely private sector use, occluding access to the technical and operational details of their use. For instance, when agencies have contracted with third parties to provide algorithmic governance tools, code and other technical details may be protected by the same patent, copyright, or trade secrecy rules that apply in the private sector context. Government use provides it no further right to distribute code.[159] When produced in-house, however, code may instead be protected under FOIA's law enforcement or trade secrecy exemptions.[160] And when produced in-house for adjudication, the status of such software remains unclear. Some agencies affirmatively exclude software in their FOIA implementing regulations.[161] Others, like the U.S. Digital Service, have open sourced their code. Even when code is available, however, parties may be unable to understand how the algorithm works in practice, or fully identify errors and bias, without access to underlying data. A facial recognition model, for instance, may appear flawless in code, but gender and racial disparities can emanate from training data that underrepresent individuals with darker complexion.[162] Yet in many agency domains, the underlying training data cannot be fully disclosed by law. In the SSA context, individual data is protected under the Privacy Act of 1974.[163] And in the SEC context, while raw disclosures are available, data from prior investigations used in supervised learning models (e.g., which filings triggered elevated review) is likely protected under FOIA's exemption for law enforcement purposes.

Second, even if data and code were made available, courts performing judicial review of the APA sort remain poorly situated to review the accuracy of the machine learning model as a whole. As a preliminary matter, litigants typically seek to remedy the specific error in their case. A court might therefore find that the algorithm wrongly flagged a benefits applicant as undeserving and order the agency to correct the error. But it is much harder to probe and provide a remedy for systematic sources of algorithmic error within the confines of a single APA challenge. Consider the case of *Ledgerwood v. Arkansas Department of Public Health*,[164] where Medicaid recipients challenged the method of allocating caregiver hours to recipients with disabilities under state law. Prior to 2015, nurses assessed individual need to assign caregiver hours. After publishing a notice of proposed rulemaking to merge two programs, the state switched to deploying an algorithm to assess needs. In 2018, a state trial

159. 48 C.F.R. § 12.212 (2019).

160. 5 U.S.C. § 552(b)(4), (7) (2018).

161. *See, e.g.*, 32 C.F.R. § 291.3(b)(2)(ii) (2020) ("Normally, computer software, including source code, object code, and listings of source and object codes, regardless of medium are not agency records.").

162. Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, CONF. FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 77 (2018).

163. 5 U.S.C. § 552(a) (2018).

164. Ark. Dep't of Human Servs. v. Ledgerwood, 530 S.W.3d 336 (Ark. 2017).

court found that the failure to notify individuals of the algorithmic change was a statutory violation. The legal aid attorney Kevin De Liban obtained the algorithm in a twenty-one-page printout, making it extremely difficult to scrutinize. While the court enjoined the agency from using the algorithm, it resorted to relying on the procedural defect of failure to notify parties of the algorithmic change. This move reflects the lack of capacity of courts and litigants to engage with such tools. To be sure, expert witnesses could be hired, but, as *Ledgerwood* illustrates, this would likely have substantial distributive effects on what kind of errors can be corrected.

Third, data and algorithms may change dynamically, particularly in the enforcement context. Consider the SEC's supervised learning model for Form ADV disclosures. The model is trained on past referrals to the SEC's enforcement arm, but the set of referrals grows over time, with different forms of human input for each referral. This means that each model might be distinct, so that the model reviewed at one stage (notice and comment) may already be substantively quite different when deployed. Conversely, problematic predictions at one point (the initiation of a specific enforcement action) might vanish as the model is updated. These challenges become more acute as agencies adopt more advanced forms of machine learning that are dynamic in nature (e.g., active learning or reinforcement learning). By nature, the notice-and-comment process and APA-type challenges are static and fail to generate the kind of information required to understand an algorithm in action.

Fourth, and as noted previously, even full access to source code and data does not necessarily achieve interpretability, particularly for more sophisticated models. Explainable and interpretable AI is a frontier challenge in computer science research. And if the engineers cannot understand it, the ability of parties during a sixty-day commenting period or a judge in an adversarial judicial proceeding will be even more limited. Compounding this problem is the possibility that regulated parties can deploy adversarial learning to fool models. Figure 3 displays a well-known example of the brittleness of prevailing deep learning approaches: adding seemingly random noise can fool a deep learning model into misclassifying an image, when both images are indistinguishable to human eyes. Computer scientists are actively researching defensive protocols, but the basic finding to date has been that it is remarkably easy to fool these models.

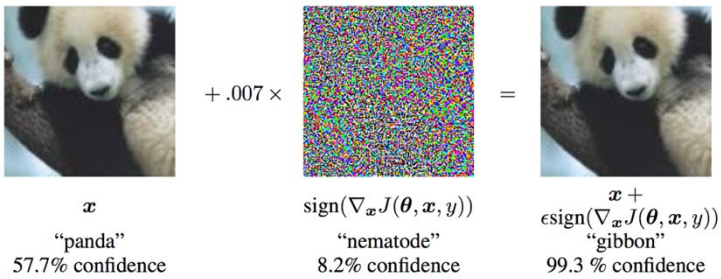

Figure 3. Prior published example of "adversarial learning" to fool image recognition model into misclassifying object.[165] By adding imperceptible noise to an image, its classification can change from "panda" to "gibbon."

Consider an example of image similarity search piloted by the Patent and Trademark Office and Word Intellectual Property Organization in Figure 4. These models deploy state-of-the-art deep learning (convolutional neural networks trained on a large set of image data). The four images are the most similar images based on a search for the World Wildlife Fund panda logo. If implemented, this image similarity tool would displace the current manual search efforts that trademark examiners engage in, based on classification codes. Yet adversarial learning can fool the similarity search into failing to retrieve existing trademarks that are visually similar to a human, thus undermining the goals of the trademark system. Moreover, because well-resourced parties are more likely to have the capacity to develop adversarial models, such developments could cause unwarranted disparities between the haves and have-nots.

165. Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy, *Explaining and Harnessing Adversarial Examples* (2014), https://arxiv.org/abs/1412.6572 [https://perma.cc/XN5P-NB7B].

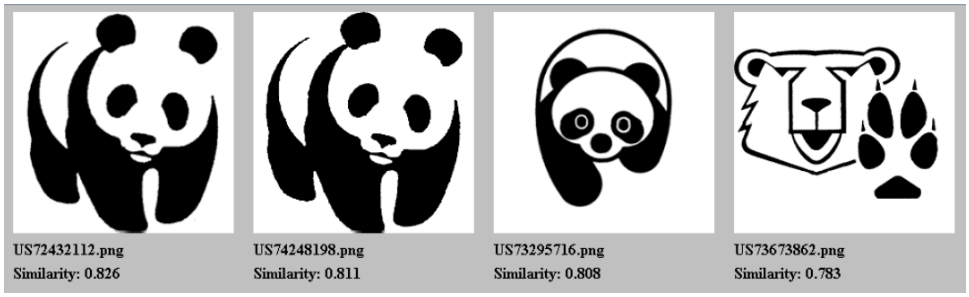

Figure 4. Example of prototype trademark similarity search model.[166] Images provide the first four search results based on a search of the World Wildlife Fund trademarked panda logo.

Similar adversarial examples exist for NLP, where adding random text that results in no meaningful change for a human reader, may fool an NLP model into misclassifying the text. Just as in the trademark example, sophisticated parties may be able to develop models to fool the SEC's NLP model into classifying a registrant's disclosure as "low risk," hence evading enforcement efforts. While the trademark example provides a potential backstop under the Lanham Act, inadvertent underenforcement due to adversarial learning has no easy solution—and indeed might never be detected—by the SEC.

Last, even if the model is completely transparent, its usage may not be. As noted previously, when a line-level prosecutor retains the ultimate authority to initiate an enforcement action, automation may (a) displace investigative resources away from false negatives, and/or (b) crowd out the exercise of discretion with suspected positives. The result can be an idiosyncratic focus on a subset of violations or, worse, a runaway feedback loop. In the SEC context, AI tools may lead the agency to fight the last war, at the expense of spotting new trends in the evasion of the securities laws by sophisticated actors. In adjudication, formal authority for adjudicators may be functional abdication. Agency adjudicators face crushing caseloads and high production quotas, so the temptation to quickly ratify model-based predictions is high. This behavior might generate the appearance of improvement in the sense of higher consistency across adjudicators. The system might thus appear to have solved the problem of arbitrariness, but only because of the fiction of adjudicator review.

Having canvassed the APA landscape and its applicability to concrete AI use cases, it is hard to escape the conclusion that the current APA mechanisms

166. Christophe Mazenc, *Machine Learning Applied to Trademarks Classification and Search*, WIPO 39 (2018), https://www.wipo.int/edocs/mdocs/globalinfra/en/wipo_ip_itai_ge_18/wipo_ip_itai_ge_18_p17.pdf [https://perma.cc/N8M9-LG5C].

remain ill-suited for providing meaningful accountability over rapid advances in AI.

III. Regulating the New Algorithmic Governance

This Part steps back and asks how current oversight of administrative action may be adapted and reformed to address the basic gaps we identified above. We spell out several possible reforms, ranging from a minimalist retrofitting of the APA to a maximalist creation of a comprehensive oversight board. We consider these options, as well as a middle-ground approach that would require agencies to engage in prospective benchmarking of AI governance tools, empowering agency administrator and external overseers alike to assess, diagnose, and correct for deviations between AI-augmented and human decisions.

A. Retrofitting the APA

Retrofitting the APA would likely entail one of two moves: subjecting algorithmic tools to the APA's procedures for notice and comment and relaxing the APA's limitations on reviewability in the enforcement or other contexts.

1. Notice and Comment

One move would be to provide greater clarification for when the adoption of AI constitutes a legislative rule, given the novel questions presented by AI use cases. We suggest several factors that may guide courts and agencies in this analysis.

First, the more humans remain "in the loop," the less notice and comment should be triggered. The QDD process, for instance, still ultimately leaves it to a QDD review team to decide whether to grant benefits for expedited cases. Human review, however, cannot be a mere formality. The review process would have to be designed to permit genuine exercise of human discretion—e.g., the QDD review team would need sufficient time and decisional independence to review proposed grants. Otherwise, the adoption of AI may functionally bind officials and "substantially alter[] the rights and interests of regulated parties," counseling in favor of notice and comment. An additional indicator of the extent of displacement of human discretion is the *threshold* for human review. In the enforcement context, a supervised learning algorithm that flags a case as "high risk" necessarily sets a (probability) threshold for the risk classification. The lower the threshold, the greater the chance of false positives and the lower the chance of false negatives. Indeed, when the threshold is zero, that is equivalent to using no algorithm at all; all cases would have to be processed by a human reviewer. The more the threshold approaches one, the

greater the risk that human discretion is displaced by the algorithm.[167] This fundamental tradeoff—between false positives and false negatives—may affect the amount of human discretion and the optimal threshold cannot be determined absent a weighing of the social costs of each type of error.[168] The proper level at which to set the threshold is precisely where public participation via notice and comment may be most useful.

Second, notice-and-comment is more appropriate when AI adoption involves considerable distributive consequences. For instance, when QDD expedites benefits to a distinct demographic group, the decision presents larger policy questions best suited for notice and comment. How distinct are applicants that apply in paper form (along geography, age, race, or gender)? If so, is there a way to deploy resource savings from QDD to provide comparable benefits to these applicants? In the enforcement context, could machine learning inadvertently perpetuate prior enforcement priorities? These broader questions may benefit from notice and comment, even if the model remains at the development stage. For instance, research around predictive policing has yielded useful approaches to the runaway feedback loop: allow only new arrests to enter the training data when the arrest was surprising relative to the model.[169] Here notice and comment genuinely allows agencies to secure input on how to design more robust AI tools.

Third, the desirability of notice and comment of the algorithm differs for enforcement and adjudication. In enforcement, for the same reasons that FOIA exempts enforcement data, notice and comment of an algorithmic adoption may do more to impede than improve the tool. In contrast, there is value to beneficiaries of understanding the method and criteria of benefits in the adjudicatory context. Because the process is itself product, algorithmic changes matter for claimants. It would be relevant, for instance, if claimant groups opposed expedited grants because of the omission of hearings. The adoption of AI for adjudication—when it implicates hearings and decisional independence—should hence be more likely to be subjected to notice and comment. Adoption of AI may "encode[] a substantive value judgment" that acts as more than a mere procedural rule.[170]

While these factors help to clarify when the adoption of AI should be subjected to notice and comment, they merely provide guidance. Specific applications remain far from clear.

---

167. This is based on the assumption that human reviewers are much less likely to pay attention to the large pool of predicted negatives.

168. For instance, in the QDD setting, a low threshold for expedited resolution may mean that more agency resources are diverted, therefore lengthening the decision time for non-expedited decisions.

169. Danielle Ensign et al., *Runaway Feedback Loops in Predictive Policing*, CONFERENCE OF FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY (2018), https://arxiv.org/abs/1706.09847 [https://perma.cc/4APU-N8QK].

170. Air Transp. Ass'n v. Dept. of Transp., 900 F.2d 369 (D.C. Cir. 1990), *judgment vacated and remanded for mootness*, 498 U.S. 1077 (1991).

2. Reviewability

Our suggestions for reviewability are distinct across adjudication and enforcement. In adjudication, claimants can challenge the denial of disability benefits in district court. Yet jurisdiction channeling—whereby the remedy for an improper denial is to reverse the agency's decision—makes it more difficult for claimants to challenge systematic sources of error, which are much more likely to be prevalent when they stem from algorithmic decisions. Due process counsels in favor of enabling claimants to challenge algorithmic decision tools. For instance, if the SSA's Insight system fails to parse a particular functional impairment that is contested for a subgroup (e.g., balancing for individuals with chronic back pain[171]), litigants should be able to seek a remedy that goes beyond the granting of benefits, namely remedying the systematic error of the Insight program.

In the enforcement context, Congress or courts may wish to relax the presumption against reviewability of enforcement prioritization under *Heckler v. Chaney*.[172] Alternatively, liberal characterization of algorithms as rules combined with pre-enforcement review would potentially enable parties to determine when an algorithm has deviated substantially from the formal goals of enforcement.

While these APA fixes would ensure greater accountability of AI tools, significant underenforcement against bad algorithms is likely to remain.[173] Judicialization of agency decision making also brings well-known costs, introducing delay, diverting scarce agency resources, and disrupting agency priority setting.[174]

B. Mixing Ex Ante and Ex Post Review: An Oversight Board

Given the limitations of ex ante and ex post review, an institutional solution might be an oversight board of AI strategy within the agency.[175] Congress could mandate this by statute or agencies could create an oversight board by rule. The charge to such an oversight board would be to (a) provide

171. Sara Kersnoveske, Libby Gibson & Jenny Strong, *Item Validity of the Physical Demands from the Dictionary of Occupational Titles for Functional Capacity Evaluation of Clients with Chronic Back Pain*, 24 WORK 157, 165-66 (2003).

172. 470 U.S. 821 (1985).

173. If Congress wanted to incentivize private enforcement, it could provide for attorney's fees when litigation results in a correction of government AI.

174. Nicholas Bagley, *The Puzzling Presumption of Reviewability*, 127 HARV. L. REV. 1285, 1338 (2014) (summarizing the debate).

175. The natural analogy here is Inspectors General offices or, in Margo Schlanger's framing, "offices of goodness" or other "ombudsman" approaches. *See* Margo Schlanger, *Offices of Goodness: Influence Without Authority in Federal Agencies*, 36 CARDOZO L. REV. 53 (2014). For studies of IGs, many in the civil rights context, see Mariano-Florentino Cuéllar, *Auditing Executive Discretion*, 82 NOTRE DAME L. REV. 227, 256 (2006); Katyal, *supra* note 86, at 2347-48; and Shirin Sinnar, *Protecting Rights from Within? Inspectors General and National Security Oversight*, 65 STAN. L. REV. 1027, 1035 (2013).

input for a strategic agency AI plan, (b) serve as a check for whether AI deployment comports with relevant law and policy (e.g., due process, antidiscrimination), and (c) review and issue recommendations for revising algorithmic decision tools. Board members could include senior agency staff in charge of developing the use case, the agency's Evaluation Officer (mandated under the Foundations for Evidence-Based Policymaking Act) or Chief Data Officer, academics, other stakeholders (e.g., disability rights groups, industry representatives), and representatives from other agencies.

Such a board could yield several benefits. First, a board would provide both ex ante and ex post oversight of AI deployment, without the substantial costs of notice-and-comment rulemaking or a judicial challenge. Second, by focusing on a longer-term strategic plan, the oversight board can make recommendations that touch on other agency operations that can facilitate AI. A major limitation of SSA's predictive modeling, for instance, is that much of the applicant data (e.g., previous occupation) is unstructured. Agencies have deployed significant resources to use NLP techniques to convert unstructured text into structured data, but a first order solution—one that might in fact be cheaper in the long run—would be to standardize inputs.

Third, the board would pool perspectives across levels of decision making and agencies. Current use cases are isolated across and within agencies, and the board could spark new innovation by providing perspectives from outside of the specific office. The SEC, for instance, could benefit from an agency that has considered use of "generative adversarial networks" to disclose data to enlist outside data scientists who can bring a fresh analytic eye without triggering privacy concerns. At a Roundtable we convened, over twenty agency officials attended and expressed tremendous value in sharing knowledge from what are otherwise disconnected programs. One agency official, for instance, had developed software to carry out topic modeling for comments submitted in rulemakings. Another had given a great deal of thought to inviting academics for short-term visits to foster idea generation. A board could help pool such insights across comparable agencies.

Fourth, the board could explicitly assess the potential for disparate impact. For instance, if there are serious concerns that expedited benefits would disadvantage certain demographic groups because of variations in filing capacity, the board could consider recommendations to level those differences. Fifth, the board would provide an institutional structure to determine evidence of adversarial learning to fool government AI tools (e.g., burying harmful disclosures amongst more boilerplate). Last, perhaps the most significant benefit of the board would be to foster a learning environment at the agency. The SEC represents an agency where staff were encouraged to experiment and fail. Many other agencies lack such a "sandbox" environment, which seriously impedes AI innovation within government. The board could foster such a culture of AI innovation within and across agencies by reducing administrative barriers (e.g., providing template position descriptions, developing best

practices for academic residencies, and publicly rewarding pilots regardless of result).

That said, there are considerable costs to an oversight board, most notably in time, FTEs, and resources. The solution to bad bureaucracy is not necessarily more bureaucracy. And if the prime reason for underdevelopment of AI tools in the administrative state lies in resource constraints, diverting more time to a Board may dilute already scarce AI skillsets. To be sure, the precise size and composition could be tailored to address these concerns. Agencies like SSA, EOIR, OMHA, and BVA, for instance, have structurally very similar problems, and could create a common oversight board for mass adjudication. Similarly, the SEC, EPA, and IRS each desire to learn from rich administrative data with similar ideas for enforcement targeting, and could thus benefit from information exchange.[176] Agencies may be reluctant to create such oversight boards, precisely for fear of airing the dirty laundry, but external perspectives may be important for identifying potential blind spots. Perhaps the most substantial limitation is that a Board may be limited in its capacity to engage with the operational details of these tools. Absent another mechanism for monitoring the impact of AI tools, the Board may have only limited information to support its decision.

C. Prospective Benchmarking

A third and less resource-intensive mechanism can generate critical information for accountable algorithmic adoption. We call the proposal "prospective benchmarking." The core idea is that when agencies adopt an AI decision making tool, they should subject it to *benchmarking* relative to a random hold-out set of cases that undergo conventional human review.[177] We conceive of benchmarking as *prospective* in nature as opposed to a backward-looking evaluation of how AI changed prior decision making. Such retrospective evaluations are widely acknowledged to suffer from significant inferential challenges.[178] Incorporating prospective benchmarking into deployment of AI systems enables agencies to more reliably learn about potential impact.[179] What is the accuracy gain of AI adoption? Is bias affected

176. *See* GOVERNMENT BY ALGORITHM, *supra* note 17, at 90 (cataloging benefits of inter-agency collaboration).

177. For a somewhat similar proposal in the private sector context requiring companies to show they tested a new model with and without newly available data in order to gauge potential disparate impact, see Selbst & Barocas, *supra* note 89, at 1129-38.

178. Michael Abramowicz, Ian Ayres & Yair Listokin, *Randomizing Law*, 159 U. PA. L. REV. 929 (2010); Daniel E. Ho & Donald B. Rubin, *Credible Causal Inference for Empirical Legal Studies*, 7 ANN. REV. L. & SOC. SCI. 17 (2011).

179. It bears noting that prospective benchmarking as envisioned here differs in important respects from a commonly advocated accountability mechanism: "algorithmic impact assessments." *See Directive on Automated Decision-Making*, GOV'T OF CANADA (Feb. 5, 2019), https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592 [https://perma.cc/4RQE-KD45]; Dillon Reisman et al., *Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability*, AINOW (Apr. 2018), https://ainowinstitute.org/aiareport2018.pdf [https://perma.cc/W5Z6-P8T8];

by AI adoption? How much time is saved? Does the algorithm engage in any other systematic errors relative to human review?

The proposal starts from the basic setting that in each of the use cases detailed above, machine learning is beginning to displace the exercise of human discretion in agency decision making. And because the status quo consists of a fully human decision, this adoption process provides a compelling opportunity to benchmark the tool by leveraging the central insight of machine learning: use of a random hold-out (or test) set to compare outcomes between the AI-assisted and human (status quo) decision. In the SSA context, for instance, the Insight system could be deactivated for a random hold-out set and compared to decisions made with Insight fully activated. In the SEC context, investigators could be required to fully investigate cases without the aid of risk scores for a subset of cases, with the results then compared to algorithmically assisted case decisions. In the PTO context, a trademark search system could be disabled for a random set of trademark registration applications and the results compared to decisions made in reliance on the search system. Such a proposal is easy to implement, as the agency is already in the process of transitioning from a manual to an AI-assisted system. The primary cost is that a subset of decisions would not garner the potential benefit of the new system.

The proposal would enable agencies, courts, and the public to meaningfully assess the impact of AI use cases, promoting accountability and transparency without the uncertainty of rulemaking, the overhead of an oversight board, or the inferential challenges of ex ante impact assessments.

First, benchmarking facilitates rigorous validation of the AI tool, which is sorely lacking in current practice. In the enforcement context, the NLP application for investment advisor disclosures displaces how human investigator would normally read such disclosures, and we are unaware of any serious attempt to compare the NLP flag against an investigator's independent assessment, particularly for disclosures that were not flagged.[180] Even in instances where an agency offers evidence of performance gains, it is unclear how much to attribute to the deployment of AI. Consider SSA's method to

Scientific Foresight Unit, *A Governance Framework for Algorithmic Accountability and Transparency*, EUR. PARLIAMENTARY RES. SERV. (Apr. 2019), https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU(2019)624262_EN.pdf [https://perma.cc/WCD6-4L5C]. Impact assessments call for evaluating the potential effects of algorithmic adoption on citizens (e.g., human rights), typically in advance of adoption. While such assessments can be valuable, it may also be challenging to make such predictive judgments in the absence of evidence of how the AI system operates in practice. Indeed, as Michael Greenstone has written, this is exactly the challenge facing regulatory agencies when asked to perform cost benefit analyses prior to the adoption of a regulation. *See* Michael Greenstone, *Toward a Culture of Persistent Regulatory Experimentation and Evaluation*, *in* NEW PERSPECTIVES ON REGULATION 111 (David Moss & John Cisternino eds., 2009). In contrast, prospective benchmarking acts like a rigorous evaluation (or randomized controlled trial) of the AI system itself, ensuring that the system is deployed in a way that facilitates evaluation of its impact. Policy advocates have called for algorithmic impact assessments to be renewed occasionally, and benchmarking can be conceived of as a natural extension of such goals.

180. GOVERNMENT BY ALGORITHM, *supra* note 17, at 26 (noting relatively weak validation and post-implementation testing).

compare processing times and error rates between branches that voluntarily adopt clustering and branches that refuse to adopt clustering. Nonrandom selection makes it impossible to disentangle the effects of microspecialization from the effects of a managerial change.[181] Reduction in errors and processing times could simply have been due to a renewed managerial commitment that confounds the adoption of microspecialization.

Instead, benchmarking enables decisionmakers to directly assess the impact of the AI tool in real time. Benchmark samples provide a comparison group to smoke out inaccuracies and biases. If the SEC algorithm, for instance, provided a high-risk estimate associated with an idiosyncratic network of investment advisors that was prosecuted last year, the model may perpetuate the effect of that network, but human reviewers would update based on the prior prosecution. Benchmark data would thus enable the agency to assess the impact of the AI tool, including where the model might perform poorly, whether the formal "human-in-the-loop" functionally ensures human oversight, and whether there is presence of "automation bias."

Second, ongoing benchmark data would provide invaluable information for calibrating and updating machine learning models. If adjudicators, investigators, claimants, and regulated parties change over time or due to different circumstances (known as "temporal drift" or "domain drift"), a model trained on a random retrospective test sample may not generalize prospectively.[182] Similarly, benchmark data would enable understanding when adversarial learning by regulated parties might invalidate historical models.[183] Data emerging from the benchmarked sample would provide the information to update models based on such state changes or to account for such adversarial behavior.

Third, benchmarking may be particularly valuable in instances where the government has contracted for AI services. In those instances, the government may not have access to technical details, but benchmark data can provide a performance standard to which an AI system developed by a contractor must adhere.

Last, the benchmarking process ensures that agency officials retain the resident expertise required to process cases. One particular concern might be that overreliance on AI systems could hollow out the expertise to tailor and adapt the regulatory scheme to changing context. By requiring a critical mass of human reviewers to maintain the workflow, agencies ensure that such hollowing does not occur.

---

181. *See, e.g.*, Daniel E. Ho, Sam Sherman & Phil Wyman, *Do Checklists Make a Difference? A Natural Experiment from Food Safety Enforcement*, 15 J. EMPIRICAL LEGAL STUD. 1 (2018).

182. Gerhard Widmer & Miroslav Kubat, *Learning in the Presence of Concept Drift and Hidden Contexts*, 23 MACHINE LEARNING 69 (1996).

183. Nilesh Dalvi et al., *Adversarial Classification*, PROC. 10TH ACM SIGKDD INT'L CONF. KNOWLEDGE DISCOVERY & DATA MINING 99 (2004).

The benefits of benchmarking become clear in contrast to retrofitting the APA or review by an oversight board. Ex ante examination via notice and comment will be unable to capture an algorithm's operational performance, which is particularly important given the dynamic nature of machine learning. Ex post judicial review or review by an oversight board would have access to a record of decision making, but would still fall short on understanding the impact of the algorithmic decision tool relative to the counterfactual of human decision making.

That said, benchmarking does have weaknesses and tradeoffs. Even if human reviewers are blinded from algorithmic determinations in their test sample, their judgments may be affected by the AI system. If adjudicators trained on the Insight system, for instance, adapt by focusing on the specific quality flags thrown by the system, their manual review may fail to catch errors that are not programmed into the Insight system. As a result, high correspondence between automated and human review may not represent the actual risks of the automated system, undercutting the utility of benchmarking. This weakness may be overcome by ensuring that individuals with substantial experience prior to the adoption of the AI system participate in the benchmarking team.

Another concern is that mandatory human review could potentially diminish the benefits of AI adoption. If the AI system works, some might argue, we should *require* its usage across the board instead of reserving a sample for human review. While this reservation is most forceful for strongly validated technology, the question at the heart of AI adoption in the administrative state is about its potential effects, biases, and unanticipated performance. Because ex ante algorithmic impact assessments will necessarily be faced with uncertainty, benchmarking provides the critical epistemic benefit of enabling us to assess whether the AI system works as intended.

A final concern is that human reviewers may face conflicts of interest if they themselves have preferences, positive or negative, regarding the AI system. If there is general resistance to AI adoption, say because of fear of job security, reviewers may have an incentive to overinvest in identifying issues that the AI system cannot detect, making such human performance unrepresentative of conventional human processing. Conversely, if reviewers favor AI adoption, they may either underperform—or perform exactly as they would expect the AI system to perform—to make the automated system appear to be effective. While such perverse incentives are harder to address, at least two responses follow. One is that "overperformance" may actually be desirable. If SSA adjudicators become much more exhaustive in their benchmark review, AI adoption may feed a process of continuous improvement, enabling the AI system to better detect errors going forward. In computer science terms, humans might best be allocated to "explore" for novel fact patterns and machine learning systems may best be allocated to "exploit" known patterns by

scaling solutions.[184] Another response is that conducting these reviews in teams may make it much less likely that reviewers will "throw" the review. Staffing reviews in this fashion can effectively turn them into peer reviews that set the gold standard for how cases should be handled.[185]

In short, while there are certainly limitations to the benchmarking idea, the approach is likely to provide critical information for algorithmic accountability that ex ante impact assessments and ex post evaluations cannot provide.

How might such prospective benchmarking come about? First, Congress could statutorily promote benchmarking, either by mandating its use by agencies or, alternatively, by prescribing or increasing judicial deference to agency decisions supported by satisfactorily benchmarked algorithmic systems. Second, courts could find agency decisions made using AI systems without benchmarking to be arbitrary and capricious upon conventional APA review. Note, however, that this assumes that plaintiffs invoking judicial review of agency action can surmount the reviewability and other barriers detailed in Part II. Third, the President could mandate benchmarking by way of executive order. Last, agencies themselves could institute such benchmarking of their own accord. Indeed, benchmarking has a close analogue to "quality improvement" initiatives or audits, as prescribed by the Government Performance and Results Act, that review a random sample of cases to calculate performance metrics. That said, current practices do not inspire much confidence in the latter path. Agencies have little incentive to monitor when an AI solution has gone wrong, as the incentive may be to tout successes.[186]

We acknowledge that many details of benchmarking remain to be worked out. At what intervals should benchmarking be conducted beyond initial adoption? How much hold-out data is necessary? How should the review process be staffed? Who will make decisions about revising, updating, or even decommissioning a tool? Over time, the benchmarking system will ideally converge on a set of best practices in answer to these and other questions. In the meantime, and regardless of the precise vehicle by which benchmarking is implemented, the approach holds significant promise because it amounts to good machine learning and good governance, with substantially less overhead than alternatives. It provides a feasible and rigorous way to hold AI decision tools accountable, to increase transparency around their adoption, and to ensure that agencies themselves can ensure internal due process around their adoption.

184. Robert C. Wilson, *Humans Use Directed and Random Exploration to Solve the Explore–Exploit Dilemma*, 143 J. EXP. PSYCHOL. GEN. 2074 (2014).

185. *See* Daniel E. Ho, *Does Peer Review Work? An Experiment of Experimentalism*, 69 STAN. L. REV. 1 (2017).

186. *See, e.g.*, Ho et al., *Quality Review*, *supra* note 15.

Conclusion

In this article, we have provided rich case studies of frontier deployments of AI in the federal government emerging out of a major study for ACUS. As these case studies show, AI is increasingly moving to the center of administrative governance. Yet conventional proposals have not seriously grappled with the body of law that is most likely to negotiate the collision of technology and the administrative state. We have argued that conventional administrative law is ill-equipped for this challenge and that serious rethinking is in order to preserve principles of transparency and reasoned decision making. Our benchmarking proposal is by no means a full governance approach, but it is a simple, powerful, and eminently achievable approach.

Will the new algorithmic toolkit reinvent government, or will it instead go the way of Sunnyvale? To be sure, one can see in the current algorithmic moment some of the same overconfidence, techno-utopianism, and internal capacity-building challenges that felled the "reinventing" movement. But the analogy is also useful because it brings into stark relief the very different stakes, and the very different political context, of the current algorithmic moment. The new algorithmic governance tools are not limited to internal performance metrics. Rather, they will increasingly displace administrative discretion at the core of the redistributive and coercive power of the state. And they will do so against the backdrop of a political zeitgeist, fueled by a pronounced anti-government, populist sentiment, that is far more challenging than the calls for a more efficient and streamlined government that powered the 1990s-era "reinventing" effort.

The stakes are high. Managed well, algorithmic governance tools can modernize public administration, promoting more efficient, accurate, and equitable forms of state action. Managed poorly, government deployment of AI tools can confirm views about inefficient and arbitrary government, hollow out the human expertise inside public bureaucracies with few compensating gains, and widen, rather than narrow, the public-private technology gap. Given these stakes, policymakers, agency administrators, judges, lawyers, and technologists should think hard, and concretely, about how to spur, not stymie, government adoption of AI tools while building appropriate accountability mechanisms around their use. Unless administrative law develops a coherent doctrinal and institutional approach to the governance of agency use of AI, its promise may prove as unrealized as President Clinton's promise to reinvent government some twenty-five years ago.