THE C. BOYDEN GRAY
Center *for the* Study
of the Administrative State
ANTONIN SCALIA LAW SCHOOL • GEORGE MASON UNIVERSITY

Algorithmic Accountability in the Administrative State

David Freeman Engstrom
Daniel E. Ho

CSAS Working Paper 19-34

Technology, Innovation, and Regulation, November 15, 2019

GEORGE MASON UNIVERSITY | ANTONIN SCALIA LAW SCHOOL

ALGORITHMIC ACCOUNTABILITY IN THE ADMINISTRATIVE STATE[1]

David Freeman Engstrom[2] Daniel E. Ho[3]

Abstract

How will artificial intelligence (AI) transform government? Stemming from a major study for 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 algorithmic accountability literature, our analysis centers around a detailed technical account of a pair of current applications in adjudication and enforcement 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 likewise push past the literature's narrow focus on constitutional law and instead train much of our analysis on 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 administrative law might adapt in response. Finally, we ask how to build a sensible accountability structure around public sector use of algorithmic governance tools while maintaining incentives and opportunities for salutary innovation. Reviewing and rejecting commonly offered solutions, we propose a 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

[1] 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, and the C. Boyden Gray Center for the Study of the Administrative State's Research Roundtable on Technology, Innovation, and Regulation for helpful conversations.

[2] 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

[3] 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

about, validate, and correct errors in algorithmic decision making. The analysis is motivated throughout by a conviction that 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 an accountability infrastructure around their use.

INTRODUCTION 3

I. THE ALGORITHMIC TREND IN ADJUDICATION AND ENFORCEMENT 8

A. ***Process as Product: Social Security Adjudication*** 9

B. ***Process as Punishment: Securities Enforcement*** 13

II. ADMINISTRATIVE LAW AND THE PUZZLE OF ALGORITHMIC ACCOUNTABILITY 22

A. ***Ex Post Review of Algorithmic Decisions*** 25

B. ***The Limits of Ex Ante Review*** 32

C. ***Informational Difficulties*** 36

III. REGULATING THE NEW ALGORITHMIC GOVERNANCE 39

A. ***Retrofitting the APA*** 39

1. ***Notice and comment*** 39

2. ***Reviewability*** 40

B. ***Mixing Ex Ante and Ex Post Review: An Oversight Board*** 41

C. ***Prospective Benchmarking*** 42

CONCLUSION 44

INTRODUCTION

In 2018, IBM published a white paper touting artificial intelligence (AI) as a way to "reinvent[] the business of government."[4] With IBM's help, governments can undergo a digital transformation, becoming more client-oriented, and "recogniz[ing] each citizen as a whole individual, with a personalized set of needs, interests, capabilities, and vulnerabilities."[5] Citizens will "know that their government has their interests at heart."[6] Moreover, new AI-based tools can "[i]mprove the decision making of civil servants for maximum impact,"[7] empowering agency administrators to "apply digital insights to predict and intervene for better citizen outcomes." "[D]igital reinvention" will yield a government that is not only more effective at performing its duties, but also one that is more responsive to citizens and operates with "[g]reater transparency."[8]

These claims should have a familiar ring. Twenty-five years ago, President Clinton made comparable promises to "reinvent government."[9] Speaking near a Sunnyvale community center in Silicon Valley, Clinton and Vice President Gore lauded the city as a model for reinvention.[10] 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 *Reinventing Government*, Sunnyvale was "the performance leader,"[11] transforming government into a lean, responsive, customer-oriented business. 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."[12] The new digital toolkit would also enable government to "empower citizens to shape the marketplace according to their own needs and values"[13] and, as Gore put it, "earn back the trust of Americans."[14]

Yet Sunnyvale floundered. When its performance index dropped, it changed the weights. When weights did not fix matters, it abandoned the overall measure. By 1999, employees quit in droves and accused municipal leadership of mismanagement.[15] So went the beacon of public sector performance measurement. When agency administrators can define and game performance measures and lack clear baselines for judging gains, such systems can undermine rather than promote regulatory goals.[16]

---

[4] IBM, Digital Transformation: Reinventing the Business of Government (2018).

[5] Id. at 13.

[6] Id.

[7] Id. at 5.

[8] Id. at 7.

[9] Remarks by President Clinton Announcing the Initiative to Streamline Government, March 3, 1993.

[10] Paul Richter, Clinton, Gore Hail Sunnyvale's City Efficiency, L.A. Times, Sep. 11, 1993.

[11] David E. Osborne & Ted Gaebler, Reinventing Government: How the Entrepreneurial Spirit Is Transforming the Public Sector 142 (1992).

[12] Seth Mydans, Where Trouble Is Rare And Governing Is Easy, N.Y. Times, Sep. 10, 1993.

[13] Osborne & Gaebler, supra note [_], at 306.

[14] Al Gore, Creating a Government That Works Better and Costs Less: Status Report of National Performance Review, Sep. 1994, at. 14.

[15] 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.").

[16] 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

Moreover, poorly constructed performance measures can reduce external accountability and oversight by burying government action under a crush of numbers and self-serving, misaligned metrics.

What should we make of current calls to reinvent government, this time using AI?[17] Can AI make good on a twenty-five-year-old promise to remake government? Will it, as IBM and many others suggest, yield a more nimble, effective, 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? Perhaps most important of all, how can law manage these opportunities and risks?

In 2019, we led a unique, interdisciplinary team of three dozen lawyers, law students, 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 federal regulatory agencies. We canvassed the roughly 150 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 re-imagine 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. But our project also 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 only "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

---

for Public Sector Quality Improvement, 13 Ann. Rev. L. & Soc. Sci. 251 (2017); see also See John Buntin, 25 Years Later, What Happened to 'Reinventing Government'?, Governing (Sept. 2016), available at https://www.governing.com/topics/mgmt/gov-reinventing-government-book.html (noting tendency of performance management systems to ossify and encourage agencies to "post good numbers" rather than develop innovative solutions to problems).

[17] 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 Business of Government (2018); Franco Amalfi, Building Government for the 21st Century, Oracle (2018); Hila Mehr, Artificial Intelligence for Citizen Services and Government, Harvard Ash Center for Democratic Governance and Innovation (2017); Miguel Carrasco et al., The Citizen's Perspective on the Use of AI in Government, BCG (2019).

[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. L. Rev. 1249 (2008) (offering a "new framework for administrative and constitutional law designed to address the challenges of the

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

This Article seeks to shift the debate onto a more concrete footing by providing a more grounded account of the new algorithmic governance tools and the challenges they raise, and by advancing a novel proposal for their regulation that balances the imperatives of internal administration with the legal demands of external accountability. In so doing, we make four distinct contributions.

First, drawing on extensive in-depth interviews and research into technical and operational details, we offer rich descriptive insight into particular applications of AI, highlighting their likely evolution and the key normative and distributive implications of their adoption. We gain needed traction in performing that task by focusing in on AI tools that support two modes of government decision-making at the heart of the redistributive and coercive power of the state: adjudication of benefits and privileges and enforcement of regulatory mandates. The use of algorithmic tools in both areas implicate profound value choices. In the adjudication context, process is the product, such that 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.[23] In the enforcement context, being singled out and made to

---

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 (2019) (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 Georgetown L.J. 1147 (2017) (but 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 Joel Tito, Destination Unkown: Exploring the Impact of Artificial Intelligence on Government, Centre for Public Impact (2017), *available at* nation-Unknown-AI-and-government.pdf (exploring criminal justice use cases only); Sandra G. Mayson, Bias In, Bias Out, 128 Yale L.J. ___ (forthcoming 2019); Solon Barocas & Andrew D. Selbst, Big Data's Disparate Impact, 104 Cal. L. Rev. 671 (2016); Andrew D. Selbst, Disparate Impact in Big Data Policing, 52 Ga. L. Rev. 109 (2017); Andrew Guthrie Ferguson, Predictive Prosecution, 51 Wake Forest L. Rev. 705 (2016); Aziz Z. Huq, Racial Equity in Algorithmic Criminal Justice, __ Duke L.J. __ (forthcoming 2019).

[21] See, e.g., Citron, supra note __; Ananny & Crawford, supra note __; Andrew Guthrie Ferguson, Big Data and Predictive Reasonable Suspicion, 163 U. PA. L. REV. 327, 329–30 (2015); Kate Crawford & Jason Schultz, Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms, 55 B.C. L. Rev. 93 (2014).

[22] *See, e.g.,* Citron, supra note __; 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); Cary Coglianese & David Lehr, Regulating by Robot: Administrative Decision Making in the Machine-Learning Era, 105 Geo. L.J. 1147 (2017).

[23] 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).

defend against a regulatory action, even if ultimately vindicated, is costly. The process itself, as the saying goes, can be the punishment.[24]

Second, we push past the abstractions of the existing literature by surfacing key technical and operational details of frontier use cases at two federal agencies. In agency adjudication, a novel application at the Social Security Administration (SSA) is an algorithmic tool that identifies disability benefits cases that are likely to be full grants, enabling the SSA to conserve resources required for a full hearing. A second algorithmic tool in use at the SSA identifies errors in draft decisions by administrative judges, thus potentially avoiding costly appeals and reversals and improving the consistency of agency decisions. Turning to agency enforcement, a wide range of key agencies, from the Securities and Exchange Commission (SEC) and Internal Revenue Service to the Environmental Protection Agency and Centers for Medicare and Medicaid Services, are developing and deploying machine learning applications that help focus scarce agency resources on high-risk individuals and entities.[25] We focus on the SEC's suite of algorithmic enforcement tools that predict, among other things, instances of insider trading and also which investment advisors are violating their obligations under federal securities laws. Looking across adjudication and enforcement in this way 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.

Third, 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,[26] we argue that much, if not most, of the hard work of regulating the new algorithmic governance tools will come not in the clouds of constitutional doctrine but in the statutory streets of administrative law. Administrative law's virtual absence in academic and policy discussions is concerning not just because of its centrality, but also because how current doctrine will resolve the most pressing cases seems far from certain. As we show, current case law is unclear whether adjudication or enforcement algorithms 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 itself a problem because it is unlikely to translate into a consistent and comprehensive approach to regulating public sector AI use that consciously balances competing concerns.

Fourth, we offer a novel and generalizable solution for monitoring, oversight, and accountability. We begin by spelling out limitations of several of the more prominent prescriptions. A minimalist option would be to retrofit the APA to enable prudent *ex ante*

---

[24] The notion that process is punishment comes from Malcolm Feeley's classic work of socio-legal research. See Malcolm M. Feeley, The Process Is the Punishment: Handling Cases in a Lower Criminal Court (1979).

[25] 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 recent effort to give off monitoring and enforcement in order to understand currents in administrative law and governance, see Rory Van Loo, Regulatory Monitors: Policing Firms in the Compliance Era, 119 Colum. L. Rev. __ (forthcoming 2019) (defining a "regulatory monitor" as "an agency actor whose core power is to regularly obtain nonpublic information from businesses outside the legal investigatory process," but also conceding that "[i]n many agencies, regulatory monitors combine prosecutors' enforcement and adjudicatory authority with the patrol function of police officers and the investigatory function of detectives").

[26] See notes __-__, infra, and accompanying text.

review of algorithmic tools through the notice and comment process or judicious *ex post* review by courts. We offer some suggestions in this regard but ultimately conclude that front-end rulemaking and back-end judicial review of the usual sort authorized by the APA are ill-suited to wrestle with the systemic considerations relevant to the adoption of AI. Forcing algorithms into the current template of notice and comment is over-inclusive and will likely retard 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 poses a substantial mismatch in judicial capacity and the technical demands of algorithmic oversight. A common but similarly limited solution looks to an oversight board staffed with technologists, academics, lawyers, and agency representatives to monitor, investigate, and recommend adjustments to agency adoption and use of AI.

We argue that a more promising intervention than either of these options would require agencies to engage in prospective human benchmarking. In a nutshell, agency administrators would reserve and then analyze a random sample of decisions using the agency's conventional, non-algorithmic approach, thus providing 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" has 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.

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 hence used interchangeably with machine learning, but excludes forms of process automation (*e.g.*, a case management system to process benefits applications digitally) and conventional forms of statistical analysis (*e.g.*, regression analysis with the aim of drawing a causal inference). Second, we focus on AI tools used to augment core agency decisions, and hence exclude forms of pure research (*e.g.*, papers published in an academic capacity by economists at the Federal Reserve Board). Third, 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 governance tool's machine learning methodology, in part because there is substantial reliance on third-party contractors to develop systems and also out of understandable concern about gaming by the regulated community. By focusing our analysis on a set of 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 as the best way to gain analytic leverage on the 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 effective internal administration and external accountability. Readers who are interested in uses of AI to support types of action other than adjudication and enforcement or who seek to understand other

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

challenges to public sector AI use—from machine learning's technical limits to adversarial learning and capacity building—are directed to the report itself and related work.

Our article proceeds as follows. Part I provides an in-depth view of two salient 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 current 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 final 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 for 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. Surfacing the full set of 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 of AI-based governance technologies. In adjudication, AI holds the promise of solving the accuracy challenge that has bedeviled the SSA for generations. More efficient and accurate processing of cases might even revive the lost constitutional value of dignity by freeing up judges to provide hearings independent of their accuracy benefits. As Jerry Mashaw put it, in adjudication "the process is the product."[28] 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. Yet because enforcement may impose serious costs on regulated parties, courts and even line-level enforcers themselves may heighten the demand for intelligibility of models.

Finally, our look beneath the hood at the SSA and SEC offers concrete confirmation of the transparency concerns that feature heavily in, and indeed dominate, the emerging algorithmic accountability literature. But our descriptive portrait also introduces a number of issues that have barely registered in that literature at all. First, there are steep technical challenges of automating government tasks that trade in large amounts of unstructured text and legalese. Second, internal capacity-building will be central to the AI transition, given the iterative process of developing useable tools and the ever-present threat of gaming. Last, the demand for intelligible models may not solely be driven by

[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).

regulated parties or courts, but rather from within the agency itself. From staff attorneys at the SSA and line-level enforcers at the SEC, the demand for intelligibility – and a form of internal due process – looms large in the face of complex AI tools.

A. *Process as Product: Social Security Adjudication*

1. *The Problem of ALJ Arbitrariness*

Consider a classic problem of formal adjudication: decisional independence risks arbitrariness.[29] 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. 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.[30] Yet while Professor Jerry Mashaw famously highlighted the problem of inconsistency some 40 years ago, decisional arbitrariness persists to the present day.[31]

[29] See Jerry L. Mashaw et al., Social Security Hearings and Appeals: A Study of the Social Security Administration Hearing System 21 (1978); Harold J. Krent & Scott Morris, Achieving Greater Consistency in Social Security Disability Adjudication: An Empirical Study and Suggested Reforms 15 (2013).

[30] 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); Kathleen G. Noonan et al., Legal Accountability in the Service-Based Welfare State: Lessons from Child Welfare Reform, 34 L. & Soc. Inq. 523 (2009); William H. Simon, Legality, Bureaucracy, and Class in the Welfare System, 92 Yale L.J. 1198, (1984); 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).

[31] See David Ames, Cassandra Handan-Nader, Daniel E. Ho & David Marcus, Due Process and Mass Adjudication: Crisis and Reform, 72 Stan. L. Rev. (forthcoming, 2019); 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).

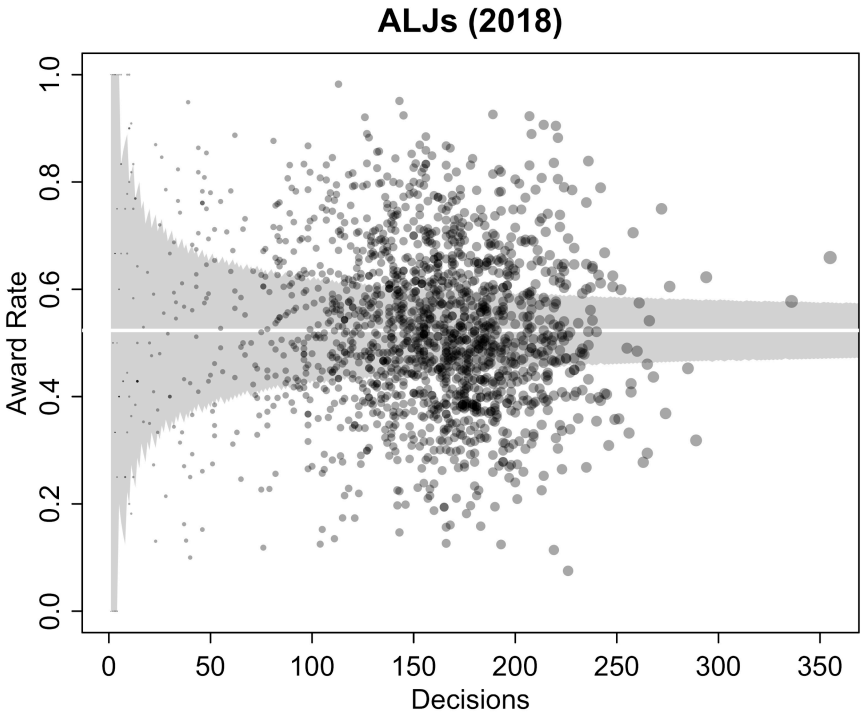

Figure 1: Number of decisions on *x*-axis against the award rate on the *y*-axis for all ALJs in 2018. The grey interval indicates the pointwise 95% interval under the null hypothesis that ALJs have the same underlying grant rate.

2. *Pioneering Applications of AI*

Can AI change this state of affairs? The SSA Appeals Council has developed three applications of AI in adjudication.[32] 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 2000 possible paths of disability cases.[33] 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 or origin) to create clusters of comparable cases. Due to labor-management concerns, clustering only re-ordered 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

[32] 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).

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

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.[34] The model would use information about medical history, treatment protocols, medical symptoms, and findings to predict easy grants, to be reviewed by a state QDD examiners. Similarly, SSA developed a pilot program for expediting claims at the ALJ hearing level. The model uses 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-cum-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 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.[35] Figure 2 presents a 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.

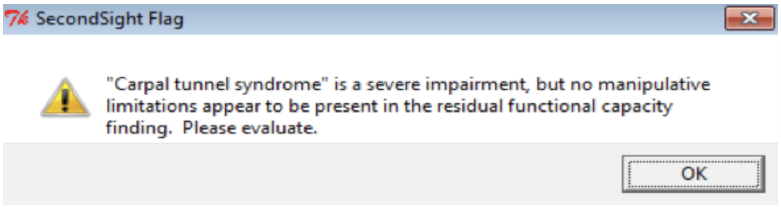

Figure 2: Screenshot from Insight system flagging a potential inconsistency in a draft decision.

3. *Trajectory*

SSA's adoption of AI has been more advanced than at other adjudicatory agencies. While it remains unclear what effect they will have on hearing-level decisions by ALJs, 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, dispositions forecasts may improve the accuracy and consistency of decisions by attorneys and ALJs. For instance, each adjudicator might be presented

[34] Administrative Review Process for Adjudicating Initial Disability Claims, 71 Fed. Reg. 16,242 (Mar. 31, 2006); 20 CFR 404.1619, 416.1019.

[35] https://www.oalj.dol.gov/LIBDOT.HTM

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 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 predict the likely language 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 lynchpin of due process. As Justice Brennan reasoned in Goldberg, the "hearing has one function only . . . to protect a recipient against an erroneous termination."[36] 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 "social malaise" that Goldberg worried about.[37] Instead, 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 was completely nervous but, after speaking and listening to him talk with kindness, I felt relief. . . . He was truly a great Judge even though I was denied."[38]

On the other hand, the adoption of AI, particularly in light of the trajectory of use cases, raises serious questions. First, each of the use cases may increasingly displace the exercise of judicial discretion, even when manual review remains nominally present. The predicted disposition might allow an ALJ to compare her inclination to the wisdom of the crowd, potentially threatening notions of decisional independence. The search tool may allow an ALJ to spend less time reviewing the record, eroding de novo review. The decision template might 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 may be exported into willingness to deviate from the automated default.

---

[36] 397 U.S. 254, 267 (1970).
[37] Id. at 265.
[38] https://www.disabilityjudges.com/state/virginia/norfolk/james-j-quigley

Second, if these tools enable centralization of policy control, they raise deep questions about separation of powers and functions within agencies. In immigration adjudication, for instance, the exemption from performance reviews was only secured by letter, not statute or regulation. As a result, the exemption was later removed, enabling greater forms of presidential control of immigration adjudication. 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."[39] 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 it 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.

Last, despite these fundamental questions, we lack even the most basic understanding of the impact of these 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 Insight system itself. If more motivated employees adopted the Insight system, the performance differences may stem simply from different levels of motivation. 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."[40] 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. *Process as Punishment: Securities Enforcement*

1. *The Challenge of Enforcement*

Agency enforcement poses a classic tradeoff between discretion and accountability.[41] Discretion is necessary because agency resources are finite but regulatory targets, and the

[39] Eugene Volokh, Chief Justice Robots, 68 Duke L.J. 1135, 1191 (2019).

[40] Social Security Administration Office of the Inspector General, The Social Security Administration's Use of Insight Software to Identify Potential Anomalies in Hearing Decisions 5 (April 2019).

[41] 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

monitoring and search costs that can be paid to identify them, are nearly limitless.[42] Monitoring and search costs can quickly eat up agency enforcement 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.[43] 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.[44] It can also conceal arbitrary selection of enforcement targets, which is itself socially costly.[45] 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 vindicated.[46] Process, as we have repeatedly noted, can be a costly and undue form of punishment.[47]

2. *Pioneering Applications of AI*

AI-based applications hold the promise of substantially reducing the agency search costs that can hamstring agency enforcement operations while making agency deployment of scarce enforcement resources more precise and less arbitrary. Less clear is whether growing agency use of algorithmic enforcement tools will render enforcement decision-making more or less transparent and thus legally and politically accountable.

The Securities and Exchange Commission's (SEC) development and deployment of a suite of algorithmic enforcement tools provide a window into the possibilities and limits of these technologies in securities regulation and beyond. While the SEC has deployed as many as half a dozen enforcement-related algorithmic tools, three tools in particular illustrate the agency's approach. The first two tools target trading-based market-based misconduct. One of these, in use at the Division of Enforcement and known as the 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

---

institutions in a way that promotes accountability while preserving a role for independent, professional judgment.").

[42] See Robert A. Kagan, Editor's Introduction: Understanding Regulatory Enforcement, 11 Law & Pol'y 89, 110 (1989) ("Most regulatory agencies feel chronically understaffed and underbudgeted relative to their caseload.").

[43] 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).

[44] 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").

[45] 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.").

[46] 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.

[47] See Feeley, supra note __.

equities and options trading records."[48] 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 in use at the SEC, called ATLAS, complements the ARTEMIS tool by focusing on first-time, rather than serial, insider trading activities. Developed in the Philadelphia Regional Office by the Office of Compliance Inspections and Examinations in collaboration with the Division of Enforcement, ATLAS is the newest of the SEC's algorithmic enforcement tools.

Both ARTEMIS and ATLAS first require a hypothesis to be generated before more targeted data collection and analysis can begin. That process typically starts with automated analysis of the public filings of a company that has experienced a significant stock movement. While companies announce important events in scheduled 10-K and 10-Q 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 sort filings into categories of reported events[51]—for instance, M&A targeting, bankruptcy, or FDA approval decisions.[52] The second is a supervised learning algorithm trained on past cases that triggered elevated review in order to flag current filings that may warrant further investigation. Note, however, that this process is only partially automated. A key factor that analysts consider is the trading volume leading up to an event, a type of inquiry that is susceptible of straightforward human review of buy volumes.

An agency examiner who concludes that trading around a specific company's stock warrants further investigation issues a "bluesheet request" and thus begins the more targeted collection and analysis of data. 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.[53] Upon deciding the target of a bluesheet request, the SEC identifies which broker/dealers traded the security at issue by obtaining the clearing reports submitted to FINRA.[54] SEC analysts must decide how far back in time to request data, up to the three-year limit authorized under the Securities Exchange

---

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

[51] 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).

[52] 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.

[53] 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. § 200, § 240 (2001). 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. Industry Reg. Authority, *Electronic Bluesheet Submissions - Attachment A Record Layout for Submission of Trading Information*, Regulatory Notice (Jan. 29, 2018), http://www.finra.org/sites/default/files/notice_doc_file_ref/Regulatory-Notice-18-04.pdf.

[54] Sec. Pub. & Priv. Offerings Appendix J7 (2d ed.), *3.2.2 Bluesheets* (Nov. 2018).

Act.[55] In order to ensure that bluesheet-derived data is high-quality, the SEC and FINRA[56] regularly bring charges against brokerage firms for inaccurate or incomplete submissions.[57]

Once bluesheet data have been collected, the SEC uses its ARTEMIS tool to analyze those data alongside data from every previous bluesheet request to determine whether the trading activity in question constitutes a suspicious anomaly.[58] The SEC has not disclosed the precise features the agency uses to make this determination, but the features 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.[59] These features are used in a one-class support vector machine to determine if a particular trade is suspicious. Automated bluesheet analysis is validated on a game theory concept called Shapley values, which attributes success across a group when contributions are unequal. If the AI/ML model identifies an outlier, the Shapley values help indicate what is driving the outlier position and what distinguishes it from other suspicious cases.

The ATLAS tools uses a similar approach. Once bluesheet data have been collected, the data is pre-processed to extract a half dozen hand-crafted data features. These features are not in the public record, but SEC staff report that they were developed using the design team's domain knowledge about insider training and are said to have an "intuitive explanation."[60] Likely candidates include how often the trade normally 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. These data features are then fed into a one class support vector machine ("SVM") to determine if the trade is suspicious.[61] The potential regulatory targets who are fed into the model are then split into two categories: those who lost money on a trade, and those

---

[55] *Id.*; *see also* Telephone Interview with Scott Bauguess, former Deputy Director and Deputy Chief Economist, Sec. & Exchange Comm'n (Feb. 15, 2019).

[56] 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 https://www.finra.org/about.

[57] For instance, in June 2016, FINRA fined Deutsche Bank Securities Inc. USD 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 1 million trades. See Press Release, Sec. & Exchange Comm'n, Citigroup Provided Incomplete Blue Sheet Data for 15 Years (Jul. 12, 2016), https://www.sec.gov/news/pressrelease/2016-138.html. And in July 2016, Citigroup Global Markets Inc. was fined USD 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. Industry Reg. Authority, FINRA Fines Deutsche Bank Securities Inc. $6 Million for Submitting Inaccurate and Late Blue Sheet Data (Jun. 29, 2016), http://www.finra.org/newsroom/2016/finra-fines-deutsche-bank-securities-inc-6-million-submitting-inaccurate-and-late-blue.

[58] *Id.*

[59] 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 Daniel Koster, Complex Financial Instruments Unit, Securities and Exchange Commission, and Jonathan Vogan, Quantitative Research Analyst, Securities and Exchange Commission (Feb. 15, 2019).

[60] Koster Interview, *supra* note **Error! Bookmark not defined.**.

[61] 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*, MACHINE LEARNING 101 (May 3, 2017), https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72.

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 are treated as suspicious.[62] Finally, as with ARTEMIS, Shapley values are used for interpretability and to help determine how each of the 6 features contributed to the determination of the particular trader being marked as suspicious or unsuspicious. They also help rule out outliers who may not be suspicious but are simply outliers.[63]

As with the determination of whether to issue a bluesheet request, the use of ARTEMIS and ATLAS to analyze trading data against prior bluesheet-derived data is only one of many tools and systems that line-level SEC enforcers use to build a case. Finding instances of insider trading is an iterative process that requires shifting through many sources of data, understanding situational context—for instance, a retail investor with only index funds who suddenly leverages in a biotech company just before an FDA approval—and being able to corral concepts into higher-level syntheses.

Moving beyond ARTEMIS and ATLAS, a third AI-based enforcement application seeks to identify investment advisors who should be subjected to more stringent treatment under the agency's examination program. 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.[64] The sheer scope of the program creates significant opportunities to economize on scarce agency resources by concentrating examination efforts on a subset of registrants.

The application aims to predict which investment advisors may be violating the federal securities laws based on disclosures made in Form ADV filings.[65] That form contains two parts. The first concerns the investment advisor's "business, ownership, clients, employees, business practices, affiliations, and any disciplinary events of the adviser or its employees."[66] 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."[67]

---

[62] Koster Interview, *supra* note **Error! Bookmark not defined.**.

[63] *Id.*

[64] 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.

https://www.sec.gov/news/press-release/2018-299

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

[66] SEC. & EXCHANGE COMM'N, FAST ANSWERS: FORM ADV (Mar. 11, 2011), https://www.sec.gov/fast-answers/answersformadvhtm.html.

[67] *Id.*

Because these forms are comprised of free text, NLP algorithms are used to normalize the inputs for analysis.[68] That process consists of three steps: (i) text 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 prior referrals to the agency's enforcement arm based on earlier investigation to classify each investment advisor as "high," "medium," or "low" risk.[71] Entities flagged as "high" risk are passed on to an SEC official, 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 has developed AI-based enforcement tools that surpass that of most federal agencies. But the SEC is by no means alone in its efforts to leverage powerful new analytic techniques and computing power in conducting enforcement. In searching for evidence of AI use cases across the top 170 federal agencies, we found that the modal use case was in enforcement, suggesting that these methods are likely to spread across many other enforcement domains. 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.

In addition, more aspects of the examination process are likely to be automated in the near future. A model could be developed, for instance, to predict whether stock trades were sufficiently anomalous to request bluesheet data from a particular company. In the medium term, advances 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"[73] – cutting edge language models (e.g., Google's BERT model) will facilitate transfer learning to adapt large-scale models to domain-specific task, requiring much less training data. Similarly, while state-of-the-art NLP methods work well with short texts—*e.g.*, the several-hundred-word IMDb movie reviews that are a fixture of CS research—more methods will be developed to deal with complex, lengthy, and jargon-filled legal documents. In the long-term, the most audacious application of AI would be to automate each step of an investigation (e.g., sending letters of inquiry, compiling answers) all the way to the filing for an enforcement

---

[68] Telephone Interview with Austin Gerig and Marco Enriquez, Office of Research and Data Services, Division of Economic and Risk Analysis, Sec. & Exchange Comm'n, and David Saltiel, Office of Analytics and Research, Sec. & Exchange Comm'n (Feb. 20, 2019).

[69] *Id.*

[70] The first is called the Latent Dirichlet allocation (LDA), which uses the "bag of words" model. 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}.

[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. https://en.wikipedia.org/wiki/Random_forest

[72] Feature importance is calculated by calculating Gini importance. https://medium.com/the-artificial-impostor/feature-importance-measures-for-tree-models-part-i-47f187c1a2c3

[73] This is based on word2vec trained on the English GoogleNews Negative300 corpus. http://bionlp-www.utu.fi/wv_demo/

action. Just as the SSA's tools could become auto-complete for adjudication, these tools could become auto-complete for enforcement.

4. *Implications*

Development and deployment of algorithmic tools by the SEC and other agencies holds significant implications for the current operation and structure of the regulatory state. AI-based enforcement tools can, by reducing agency search costs, facilitate more robust enforcement activity, whether by permitting agencies to identify more regulatory targets more efficiently or by allowing agencies to shift scarce resources away from regulatory search and toward actual prosecution of violations. Algorithmic enforcement tools can also serve as force-multipliers that narrow the public-private technology gap and thus help to level the playing field between underfunded enforcement agencies and well-resourced regulated parties.

The coming of algorithmic enforcement also augurs other, more substantial shifts in the regulatory landscape. Algorithmic enforcement tools could halt or even reverse the decades-long shift away from public enforcement and toward private litigation as a regulatory mode.[74] Indeed, one explanation for the shift from public to private enforcement, achieved largely via legislative creation of private rights of action and whistleblower schemes throughout the post-war era, was not just a recognition that public enforcers can be poorly situated to surface privately held information about misconduct. It was also a more fiscally focused legislative desire to move enforcement costs from on-budget to off-budget forms.[75] A significant reduction in regulatory search costs could alter that core legislative calculus. In addition, more public enforcement can mean better calibration of enforcement effort and, through agencies' more synoptic view of the regulatory landscape, a more certain approximation of a socially optimal level of enforcement effort.[76]

But algorithmic enforcement tools also give cause for concern. First, and most obvious is bad or biased data, particularly where model inputs are past, analog enforcement patterns. Bluesheet data, for instance, is plainly unrepresentative across capital markets, as requests are not randomly generated.[77] Supervised models depend

---

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

[75] 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).

[76] To be sure, it is also possible that if data is publicly available (e.g., SEC filings), algorithmic targeting may empower private enforcement as well. But because many of the inputs are only observable to the agency, the net effect may be to shift the mode of enforcement toward public enforcement.

[77] Sec. Pub. & Priv. Offerings Appendix J7 (2d ed.), *3.2.2 Bluesheets* (Nov. 2018). 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 Exchange Comm'n, SEC Approves Plan to Create Consolidated Audit Trail (Nov. 15, 2016), https://www.sec.gov/news/pressrelease/2016-240.html. 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. Sec. and Exchange Comm'n, Rule 613 (Consolidated Audit Trail), https://www.sec.gov/divisions/marketreg/rule613-info.htm 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

accurately labeled data, but identifying true positives and false negatives for a representative sample can be prohibitively expensive. IRS solves this problem in part with random audit data, but no such gold standard data exists for the SEC. This lack of ground truth reduces the accuracy of models and makes validation difficult. The concern hence arises that NLP-based detection may be driven by superficial features from past enforcement decisions, replicating heuristics deployed by beleaguered line-level enforcers rather than building a richer and more precise model of noncompliance.[78] Algorithmic tools may exacerbate, rather than mitigate, the risk of arbitrary agency action.

Second, to date, AI "shrink[s] the haystack" of potential offenders, but a human, and typically a lawyer, remains "in the loop." Use of these tools by line-level staff also remains, for the moment, entirely voluntary. These operational details have plainly shaped the development of the technology in legally and normatively significant ways, as agency technologists must, in effect, sell skeptical line-level staff on the tool. Voluntary status has pressed agency technologists to, among other things, develop user-friendly interfaces that enforcement staff distributed throughout the agency, including the regional enforcement offices, can access and readily use. 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." 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 building their models—thus, the focus on Shapley values within the ARTEMIS and ATLAS systems to help isolate which data features may be driving an algorithmic output. Staff skepticism and demand for explainable outputs raise the interesting possibility that governance of public sector algorithmic tools will at times come from "internal" due process, not the judge-enforced, external variety.[79] Of course, SEC officials could change either feature of the SEC's current approach by making use of the tools by line-

---

to every trade quote and order, origination, modification, execution, routing, and cancellation. *Id.; see also,* Deloitte, *Perspectives: Consolidated audit trail: The wait is over,* https://www2.deloitte.com/us/en/pages/financial-services/articles/sec-rule-613-consolidated-audit-trail-national-market-system-nms-plan-banking-securities.html#.

[78] See Erik Hemberg, Jacob Rosen, Geoff Warner, Sanith Wijesinghe, & Una-May O'Reilly, Tax Non-Compliance Detection Using Co-Evolution of Tax Evasion Risk and Audit Likelihood, ICAIL '15, June 8-12, 2015), *available at* http://dx.doi.org/10.1145/2746090.2746099; see also See Levmore & Fagan, supra note __, at __ (making the related point that automated decision tools will work best in "stable legal environments"). A more general version of the problem is that, when a line-level enforcer retains the ultimate authority to initiate enforcement, automation may displace investigatory resources away from false negatives and/or crowd out the exercise of discretion with suspected positives. Still another somewhat similar phenomenon—"runaway feedback loops"—is well-documented in the predictive policing context. When a predictive model is used to deploy police, and subsequent arrest data is used to re-train the model, a "runaway feedback loop" occurs: regardless of the crime rate, police may be sent to the same neighborhood over and over. See Danielle Ensign et al., Runaway Feedback Loops in Predictive Policing, Conference of Fairness, Accountability, and Transparency, 2018, https://arxiv.org/abs/1706.09847.

[79] 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 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); Neal Kumar Katyal, Internal Separation of Powers: Checking Today's Most Dangerous Branch from Within, 115 Yale. L.J. 2314 (2006); Gillian E. Metzger, The Interdependent Relationship Between Internal and External Separation of Powers, 59 Emory L.J. 423 (2009); Amanda Leiter, Soft Whistleblowing, 48 Ga. L. Rev. 425, 429 (2014).

level staff mandatory or by making agency decisions turn primarily or even entirely on algorithmic outputs. To that extent, the SEC could quickly increase the centrality and significance of the tools. But internal pressure to make the tools user-friendly and intelligible—an open research frontier in NLP—could still remain.

A third concern is a resource catch-22: The same resource constraints that drive agencies to automate agency operations in the first place can also cause agencies to cut corners on validation and testing, increasing the risk of low-quality decisions and arbitrary enforcement efforts. Moreover, even if an agency's enforcement operation keeps "humans in the loop," those humans may gradually pull back from engagement because of automation bias, injecting less and less human judgment into an algorithmically dominated process.[80] Without systems of political and legal accountability in place, algorithmic enforcement can be ineffective or worse.

A fourth concern centers on the possible distributive effects of algorithmic enforcement, particularly from gaming.[81] 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.[82] Worse, agency adoption of algorithmic enforcement tools may be slow or haphazard relative to the private sector, and the current trend toward algorithmic governance merely the start of an unwinnable arm's race, with public investment in technology met by equal or greater investments by the private sector. The dystopic result is a digitized government that is no more effective and yet far more expensive, both for taxpayers who foot the bill for government build-up and for society at large as all sides divert valuable social resources into a race for regulatory advantage.[83]

A final implication, though it remains unclear whether it will prove to be a virtue or a vice, is the effect of algorithmic enforcement tools on the transparency of agency enforcement efforts, and thus the degree to which the tools will improve or degrade legal and political accountability. Algorithmic tools, by distilling an agency's enforcement plan to a single encoded set of criteria, may help satisfy calls for agencies to establish concrete enforcement criteria that cabin discretion and facilitate review of agency

---

[80] "Automation bias" refers to the tendency of humans to unreasonably defer to automated outputs over time. See Citron, supra note __, at 1272; Linda J. Skitka, Kathleen L. Mosier, & Mark Burdick, Does Automation Bias Decision-Making?, 51 Int. J. Human-Computer Studies 991 (1991); R. Parasuraman and D.H. Manzey, Complacency and Bias in Human Use of Automation: An Attentional Integration, 52 Hum. Factors 381 (2010).

[81] See Engstrom et al., Enforcement by Algorithm, at __; 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).

[82] 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.

[83] A coarse analogy is the debate about whether and how robust judicial review of agency rulemaking has "ossified" the system, impairing the capacity of agencies to achieve regulatory goals with little benefit in accountability. See, e.g., Thomas O. McGarity, Some Thoughts on "Deossifying" the Rulemaking Process, 41 Duke L.J. 1385 (1992); Richard J. Pierce, Jr., Seven Ways to Deossify Agency Rulemaking, 47 Admin. L. Rev. 59 (1995); Mark Seidenfeld, Demystifying Deossification: Rethinking Recent Proposals to Modify Judicial Review of Notice and Comment Rulemaking, 75 Tex. L. Rev. (1997).

enforcement activities for consistency and equity.[84] But the opacity of many of the more sophisticated tools may well render agency enforcement policies more, rather than less, opaque, exacerbating concerns about a lack of legal and political accountability. We return to this concern momentarily in Part III's exploration of administrative law's response to the new algorithmic governance tools.

II. 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.[85] When government takes action that affects rights, it must explain why. On the other hand, the algorithmic tools that agencies are increasingly using to make and support public decisions are not, by their structure, fully explainable.[86]

A rapidly growing academic literature explores this clash, 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 a fast-growing literature with 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 the well-established idea that machine learning outputs are *inscrutable* in the sense that even their own engineers cannot necessarily understand how

---

[84] See Lisa Schultz Bressman, Judicial Review of Agency Inaction: An Arbitrariness Approach, 79 N.Y.U. L. Rev. 1657, 1693 (2004) (arguing that up-front standards will "prevent, or at least minimize, corrupting influences from pervading administrative enforcement decisionmaking"); Barkow, supra note __, at 1154 (surveying strategies for "combatting selective enforcement" and advocating a requirement that an agency "make clear the criteria it will use to make enforcement decisions"); id. at 1173 ("The key is to get agencies to better publicize their enforcement practices and the relevant metrics."); Kate Andrias, The President's Enforcement Power, 88 N.Y.U. L. Rev. 1031, 1105 (2013) (calling for greater presidential control over enforcement, including a requirement that agencies submit regular reports outlining their enforcement priorities, metrics, and results). Many such calls come via the debate over the propriety of agency use of "guidance" documents. See, e.g., Jessica Mantel, Procedural Safeguards for Agency Guidance: A Source of Legitimacy for the Administrative State, 61 Admin. L. Rev. 343, 393 (2009) (arguing that guidance documents "provide an effective means by which agencies can ensure more accurate, consistent, and predictable decisions by agency personnel"). However, it is important to note that many guidance documents are, unlike enforcement algorithms, publicly available, limiting the usefulness of the analogy.

[85] In the American context, this norm pervades administrative law, both in the Administrative Procedure Act, see 5 U.S.C. § 557(c)(3)(A) ("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 (2019). For primary materials, see, e.g., Danish Act on Public Administration, §§ 22-24; Administrative Procedure Code of 1976, 39 VwVfG (Germany); Conseil Constitutionnel 1 julliet 2004, no. 2004-497 DC (France); Charter of Fundamental Rights of the European Union (CFR) Art. 41 (EU).

[86] See, e.g., J. Burrell, How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms, 3 Big Data and Society 1 (2016);

the machine got to a given result,[87] but they are also often *non-intuitive* in that the rules they derive to make predictions are so complex, multi-faceted, and interrelated that they defy practical inspection or do not comport with any practical human belief about how the world works.[88] As 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"[89]—may not yield accountability in the sense of rendering decisions fully legible to data subjects or surfacing all of a system's flaws.[90] Instead, desired transparency may only be approximated by mixing and matching multiple, partial modes of explanation, including a "decision-level" accounting of a given decision's "provenance" via the machine's inputs and outputs, and also a "system-level" accounting of the tool's "purpose, design, and core functioning,"[91] such as data descriptions, modeling choices, and the like.[92]

---

[87] See Andrew Selbst & Solon Barocas, The Intuitive Appeal of Explainable Machines, 87 Fordham L. Rev. 1085, __ (2018). For a highly accessible version, see Judea Pearl & Dana McKenzie, The Book of Why: The New Science of Cause and Effect 359 (2018).

[88] See Selbst & Barocas, Intuitive, supra note __, at __.

[89] See Aaron Rieke, Miranda Bogen, & David G. Robinson, Public Scrutiny of Automated Decisions: Early Lessons and Emerging Methods 19 (Upturn and Omidyar Network, date?).

[90] Kroll et al, supra note __, at 661 (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 __, at 980 ("Seeing inside a system does not necessarily mean understanding its behavior or origins."); 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 Annual Meeting of the Int'l Communicaton Ass'n (May 22, 2014). 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 Center 2019). However, while emerging techniques are rendering machine learning models more interpretable by ranking, sorting, and scoring data features according to their pivotalness 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 (2018). Another approach to interpretability uses visualization techniques or machine-based textual justifications to lay bare a model's decision "pathway." See Chris Olah, et al., *The Building Blocks of Interpretability*, DISTILL (Mar. 6, 2018), https://distill.pub/2018/building-blocks; L. A. Hendricks, et al., *Generating Visual Explanations*, EUR. CONF. ON COMPUTER VISION, Springer, 2016, at 3-19. 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, Intuitive, supra note __, at 38.

[91] Selbst & Barocas, Intuitive, supra note __, at 43, 64 (offering an accessible explanation of the debate over "outcome- and logic-based explanations"). For similar efforts to categorize explanation types, see 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"); 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).

[92] See Selbst & Barocas, Intuitive, supra note __, at 64; Rieke et al., supra note __, at 18; Coglianese & Lehr, Transparency, supra note __, at __ (distinguishing between individual- and group-level explanations).

The second track in the literature tours the regulatory 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 many of the options available in any regulatory context. These include individual, rights-based measures (*e.g.*, private lawsuits; whistleblower schemes that incentivize those with 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 (*e.g.*, mandatory impact assessments).[93]

This literature has generated an initial set of insights about the accountability challenges of algorithmic governance. An example is Danielle Citron's point 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, misses the fact that algorithmic tools are designed to operate at scale. Lost in case-level balancing is the possibility that a one-time but costly increase in procedural scrutiny of an algorithmic tool can yield massive social benefits across the thousands or millions of cases to which the tool is applied.[94] But the existing literature also falls short on several fronts. Most treatments, as noted previously, abstract away from the technical and operational details of actual algorithmic tools, and many also commingle public and private sector AI use despite the very different logics and legal imperatives governing each.[95] 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 more a 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. Constitutional avoidance—which holds that courts should avoid ruling on constitutional issues in favor of other, often statutory, grounds—means that administrative law, and the Administrative

---

Beyond the transparency isue, 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 __, at 983; see also Citron, supra note __, 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, Transparency, 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. of 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 Boracas & Selbst, at 678; Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson & Harlan Yu, *Accountable Algorithms,* 165 U. PA. L. REV. 633, 679-82 (2017).

[93] See Kaminski, supra note __; see also Andrew Tutt, An FDA for Algorithms, 69 Admin. L. Rev. 83 (2017); Deven R. Desai & Joshua A. Kroll, Trust But Verify: A Guide to Algorithms and the Law, 31 HARV. J.L. & TECH. 1, 46 (2017).

[94] Citron, *supra* note __, at 1249.

[95] See notes __-__, supra, and accompanying text.

Procedure Act, not constitutional law, will very often be the legal constraint of first resort. The virtual absence of administrative 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.

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 judge-overseen administrative law, at least in its current guise, is unlikely to yield systematic, as opposed to pocketed and even idiosyncratic, review of agency use of algorithmic decision-making tools. Left to its own devices and judicial efforts to apply 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. *Ex Post Review of Algorithmic Decisions*

Under current administrative law, agency use of AI is unlikely to be subjected to systematic scrutiny via ex post judicial review. In the enforcement context, a thicket of reviewability and related doctrines insulate algorithmic decision-making. In the adjudication context, reviewability provides less of a shield, but the chance of ex post review of something like the QDD algorithm remains slim. We note that while we explain these difficulties to securing judicial review, 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.

1. *Enforcement Decisions*

Modern administrative law erects substantial barriers to judicial review of enforcement decisions.[96] 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.[97] Moreover, under the APA, courts generally lack

[96] Bressman, supra note __; see also Van Loo, supra note, 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).

[97] 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

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.[98]

*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.[99] 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.[100] The first exception, articulated a decade before *Heckler* in *Dunlop v. Bachowski*,[101] 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.[102] 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. And the resulting review can be thorough: Courts regularly require agencies to make a full explanation,[103] including how the agency assessed the importance of specific

---

denial' of equal protection of the law") (quoting Yick Wo v. Hopkins, 118 U.S. 356, 373 (1886)); 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); Marshall v. Jerrico, Inc., 446 U.S. 238, 249 (1980) (noting that "traditions of prosecutorial discretion do not immunize from judicial scrutiny 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. Am. Elec. Power Serv. Corp., 258 F.Supp 2d 804, 806 (S.D. Ohio 2003).

[98] The relevant part of the APA is Section 701(a)(2)'s prohibition of review of questions that are "committed to agency discretion."

[99] 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).

[100] Heckler v. Chaney, 470 U.S. 821 (1985). Lower courts have extended the *Heckler* 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) (__); 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).

[101] 421 U.S. 560 (1975).

[102] Chaney, 470 U.S. at 832–33, 105 S.Ct. at 1656) (noting lack of review unless Congress "has provided guidelines for the agency to follow in exercising enforcement powers"). See also Dunlop.

[103] *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 Defense 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.").

considerations[104] or why it departed from prior practice.[105] 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.[106]

The remaining exceptions, however, are even narrower and offer only weak and irregular prospects of rebutting *Heckler*'s presumption. One such exception triggers where an agency has adopted a policy of nonenforcement that rises to the level of an "abdication" of its statutory responsibilities.[107] Note, however, that this is not 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.[108] The focus, as in *Dunlop*, remains at all times on whether there are sufficient indicia of legislative intent to rebut the presumption of non-reviewability. Courts have been reluctant to find abdication in cases where the agency is engaged in at least some enforcement activity.[109] So long as an algorithmic tool does not foreclose enforcement entirely, and merely pares down the universe of targets, the exception is not triggered.

A creative route around *Heckler* would exploit a possible ambiguity in the case's normative foundation. Beyond *Heckler*'s separation-of-powers framing of enforcement

---

[104] See, e.g., Southwestern Public 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. Savings & 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).

[105] *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").

[106] 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.*

[107] Chaney, 470 U.S. 833 n.4.

[108] [CITE]

[109] *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. CIV. A. 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").

as a core executive responsibility,[110] the doctrine is generally understood to be driven by the complexity and technical nature of enforcement decisions.[111] 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.[112] That reading enjoys substantial support in *Heckler* itself,[113] and it is hard to see 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 *Heckler* 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 reading of indeterminacy also draws support in *Heckler* 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.[114] Note, however, that re-centering *Heckler* around indeterminacy could lead courts in either direction on the reviewability of an agency's use of algorithmic enforcement tools. On the one hand, algorithmic tools must typically specify an objective function, potentially converting an opaque, all-things-considered weighing of factors into a rule-bound and tractable calculus. On the other hand, an NLP-based

---

[110] 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 Baltimore 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 Horner*, 854 F.2d 490 at 497.

[111] See Baltimore Gas & Elec. Co. v. F.E.R.C., 252 F.3d 456, 459 (D.C. Cir. 2001). 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 Gen. of U.S.*, 793 F.2d 473, 477 (2d Cir. 1986), which notes that non-reviewability 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.), *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").

[112] 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 N.w. U. L. Rev. 1689, 1703 (2013) (noting different agency objective functions and "maximands" when engaged in enforcement-related decision-making).

[113] As the *Heckler* Court itself noted, an agency's enforcement decision "involves a complicated balancing of a number of factors which are peculiarly within [agency] expertise," 470 U.S. at 831, 105 S.Ct. at 1655, 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, 105 S.Ct. 1649 at 1655–56.

[114] Chaney at 832. See also Texas v. U.S., 809 F.3d 134 (5th Cir. 2015), which leans heavily on the fact that deferred action is "affirmative agency action" because it confers lawful presence and employment authorization on a large class of people who would otherwise be removeable.

machine learning tool of the sort the SEC is utilizing may be more indeterminate than even a gauzy, multi-factor written protocol that guides line-level staff working up cases in the analog way. The result is 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 other innovative paths around *Heckler*. 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.[115] 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.[116] 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 *Heckler* 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.[117] There may, 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.[118] Note as well that the designation of an algorithm as a rule takes the case outside *Heckler* entirely and raises further and distinct questions of reviewability, particularly the availability of pre-enforcement review. The possibility that an algorithmic enforcement tool constitutes a

---

[115] GoJet Airlines, LLC v. FAA, 743 F.3d 1168 (8th Cir. 2014).

[116] See Massachusetts Public Interest Research Group, Inc. v. U.S. Nuclear Regulatory Commission, 852 F.2d 9 ((1st Cir. 1988).

[117] *Edison Elec. Inst. v. U.S. E.P.A.*, 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); *see 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 Employees 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, Inc. v. United States*, 132 F.3d 808, 812 (D.C. Cir. 1998); *Crowley Caribbean Transp., Inc. 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 __, at 933 (noting that enforcement shares features of both – and involves both wholesale and retail decisions).

[118] See Norton v. Southern Utah Wilderness Alliance, 542 U.S. 55, 64 (2004); Lujan v. National Wildlife Federation, 497 U.S. 871, 891 (1990). For an argument that administrative unduly forestalls challenges to agency failures of "systemic administration," see Metzger, supra note __.

rule, with all that such a designation would entail under the APA, is taken up more fully below.

In short, save situations in which the mandatory framing of an agency's organic statute brings it within *Dunlop*'s domain, or a judicial willingness to re-center *Heckler* around agency self-cabining or inscrutability, agency use of algorithmic enforcement tools will be largely insulated from judicial challenges by non-targets seeking to compel agency enforcement.

Regulatory Targets—Challenging Agency Enforcement. *Heckler*'s presumption against reviewability is flipped when judicial review of an algorithmic enforcement tool is sought by an enforcement target itself.[119] But even here, current administrative law erects substantial reviewability barriers that block the most likely avenues for judicial challenge, foiling anything resembling systematic review.

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 not immediately challengeable.[120] 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 Socal had no similar impacts "other than the disruptions that accompany any major litigation."[121] Litigation costs, as the Court had put it several decades earlier, are "part of the social burden of living under government."[122]

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.[123] At that point, a regulatory target who has mounted an unsuccessful defense, and thus been found liable, could attempt to argue that even an agency enforcement action that is unassailable as a substantive matter is nonetheless

---

[119] 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, Inc. v. Vellop, 401 U.S. 402 (1971) (articulating strong presumption of reviewability of final agency action under the APA).

[120] F.T.C. v. Standard Oil Co. of California, 449 U.S. 232, 242 (1980).

[121] Id. at 243.

[122] Petroleum Exploration, Inc. v. Public Service 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 Board 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.

[123] The analogy to interlocutory review is an apt one, as the *Standard Oil* Court noted that, 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. Id. at ___.

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."[124] 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 question. Moreover, *Standard Oil*'s rejection of litigation costs as cognizable legal injury negates any possible recourse other than reversal on liability.[125] 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. In short, neither scenario is likely to yield systematic review of an agency's algorithmic enforcement toolkit.

2. *Adjudicatory Decisions*

While reviewability poses less of a concern in formal adjudication, the chances of judicial review of existing algorithmic decision tools like QDD 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 re-classifying the case as a "critical case" for expedited processing because the ALJ misunderstood the Hearing Appeals and Litigation Law Manual.[126] 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

---

[124] 5 U.S.C. § 704

[125] 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. This removes any possibility of common-law remedies.

[126] No. 3:12-CV-1059-O, 2013 WL 5020495, at *20 (N.D. Tex. Sep. 13, 2013) ("[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.").

process.) While back pay would ultimately be awarded, the hardship to either (a) borrow money or (b) restrict consumption while claims 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.[127] 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 (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 interrogation would 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 in Section 3.3.

B. *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 rather than a step in the agency's decision process in a particular case. This path opens up two further potential mechanisms of accountability: (i) notice-and-comment required for legislative rules; (ii) pre-enforcement judicial review of an algorithmic tool, before the tool is applied in a particular case and, in the enforcement context, without the necessity of a violation. These mechanisms, however, still amount to a patchwork of accountability under current administrative law.

1. *Legislative Rules and Notice and Comment*

An important constraint on agency discretion under the APA is the requirement that "legislative" rules must be subjected to notice and comment. That process, as most lawyers know, 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[] basis and purpose" of the rule responding to those comments.[128] As a practical matter, due to increased judicial scrutiny, the "concise general statement" is often neither. Ventilation of rules in this manner is a cornerstone of the APA's accountability regime.

---

[127] 424 U.S. 319.
[128] 5 U.S.C. 553(c).

Not all rules,[129] 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. 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,[130] (ii) whether the rule "substantially alters the rights and interests of regulated parties,"[131] and (iii) the amount of regulatory work the rule does relative to the governing statute or prior agency-promulgated rules.[132] The resulting tangle of doctrines have been described as "tenuous," "blurred," "baffling," and "enshrouded in considerable smog."[133]

These characterizations alone should be enough to establish that notice and comment is unlikely to provide a consistent or 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 also regularly characterize policies as legislative rules, even where substantial discretion remains with the agency and its line-level prosecutors.[134] An illustrative case is *McLouth Steel Prod. Corp. v. Thomas*,[135] 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 be a legislative rule requiring notice and comment.[136] 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.[137] 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.[138]

To return to the agency use cases, consider the SSA's expedited grant process (QDD). Recall that the adoption of QDD in fact went through notice and comment,

---

[129] 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 5 U.S.C. § 551(4).

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

[131] Chamber of Commerce; Air Transport Ass'n

[132] Paralyzed Veterans; AMC

[133] Cmty. Nutrition Inst. v. Young, 818 F.2d 943, 946 (D.C.Cir.1987)

[134] *See, e.g.*, *Guardian Federal Savings & 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").

[135] 838 F.2d 1317, 1319 (D.C. Cir. 1988).

[136] *Id.* at 1319-23.

[137] *Ass'n of Irritated Residents v. E.P.A.*, 494 F.3d 1027, 1034 (D.C. Cir. 2007).

[138] Id.

because it required amendment of existing procedural rules.[139] Yet whether the proposal provided sufficient notice of the algorithmic decision tool is unclear.[140] 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."[141] 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 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 alters the rights and interests of regulated parties" in light of the counterfactual delay of receipt of benefits. 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 earlier receipt, 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.[142] Under current administrative law, it remains 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.

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. Clustering cases for SSA case 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. 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. *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.[143] 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.[144] Hardship boils down to whether a rule's impact is sufficiently direct and immediate, which in turn asks whether the rule requires

[139] Arizona Grocery Co. v. Atchison, Topeka & Santa Fe Rwy., 284 U.S. 370 (1932); United States ex rel. Accardi v. Shaughnessy, 347 U.S. 260 (1954).

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

[141] 71 Fed. Reg. at 16,430.

[142] Heckler v. Campbell, 461 U.S. 458, 463 n.5 (1983).

[143] Abbott Laboratories v. Gardner, 387 U.S. 136 (1967).

[144] Bennett v. Spear, 520 U.S. 154, 177–78 (1997)

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.[145]

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—merely requires a purely legal comparison of the encoded, algorithmic rule and 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 defeating 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.[146] 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[147] and procedural rules.[148] 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.[149] The famously blurry line dividing legislative rules and other types of agency pronouncements

---

[145] Abbott Laboratories v. Gardner, 387 U.S. 136 (1967).

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

[147] 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. E.P.A., 145 F.3d 1414, 1421 (D.C. Cir. 1998) (interpretative rule held not ripe.); Ciba-Geigy Corp. v. U.S.E.P.A., 801 F.2d 430, 434 (D.C. Cir. 1986) (policy statement not ripe); see also First Nat. Bank of Chicago v. Comptroller of Currency of U.S., 956 F.2d 1360 (7th Cir. 1992), cert. denied, 506 U.S. 830, 113 S. Ct. 93, 121 L. Ed. 2d 55 (1992). However, there are exceptions. A leading case is *Aviators for Safe & Fairer Regulation, Inc. v. F.A.A.*, 221 F.3d 222, 225 (1st Cir. 2000). There, the COA 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. v. Gardner,* 387 U.S. 136, 148–49 (1967). It's sensible to think a similar approach could be taken towards the adoption of a new artificial intelligence program.

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

[149] Public Citizen, Inc. v. U.S. Nuclear Regulatory Com'n, 940 F.2d 679, 681–82 (D.C. Cir. 1991); Natural Resources Defense Council, Inc. v. U.S. E.P.A., 16 F.3d 1395 (4th Cir. 1993).

adds another way in which some algorithmic tools will qualify for pre-enforcement review while others will not.

C. *Informational Difficulties*

Even if an algorithm were proposed via notice and comment or subjected to judicial review, substantial informational barriers impede review of algorithmic decision tools. Conventional APA review is likely to be stymied by the kind of technical and operational details that are critical in each of these use cases.

First, decisions are embedded in inaccessible code. When agencies have contracted with third parties, such code may be protected by patent, copyright, or trade secrecy. Government use provides it no further right to distribute code.[150] When produced in-house, code may be protected under FOIA's law enforcement or trade secrecy exemptions.[151] And when produced in-house for adjudication, the status of such software remains unclear. Some agencies affirmatively exclude software in their FOIA implementing regulations.[152] 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. Errors and bias can originate from the training data, so the actual operation of the model may only become intelligible with the underlying data. A facial recognition model, for instance, may appear flawless in code, but gender and racial disparities can emanate from training data that underrepresents darker shade individuals.[153] Yet in many agency domains, the underlying training data cannot be fully disclosed. In the SSA context, individual data is protected under the Privacy Act of 1974.[154] 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 the data and code were made available, reviewing courts 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 the systematic source of algorithmic error. Consider the case of *Ledgerwood v. Arkansas Department of Public Health*,[155] 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 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 21-page printout, making it

---

[150] 48 C.F.R. 12.212.

[151] 5 U.S.C. § 552(b)(4) & (7) (2000).

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

[153] Joy Buolamwini & Timnit Gebru, Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, Conf. Fairness, Accountability & Transparency 77 (2018).

[154] 5 U.S.C. § 552a.

[155] https://www.arktimes.com/media/pdf/ledgerwood_v_dhs--order.pdf; https://www.arktimes.com/ArkansasBlog/archives/2018/05/14/judge-orders-dhs-to-stop-using-algorithm-for-home-care-hours-dhs-says-services-wont-be-disrupted

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, the data and algorithm may change dynamically. 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 (enforcement) might vanish as the model is updated. These dynamics become even more challenging as agencies adopt more advanced forms of machine learning that draw on 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, even with the full source code and dynamic data in hand, the black box nature of the most sophisticated machine learning algorithms can make them difficult to interpret. The fact is that we know surprisingly little about why the most advanced neural networks work.[156] 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 60-day commenting period or a judge in an adversarial judicial proceeding will be even more limited. This problem is compounded by the possibility that regulated parties can deploy adversarial learning to fool models. Figure 2 displays a well-known example of the brittleness of prevailing deep learning approaches: adding random noise to an image that to a human looks visually indistinguishable can fool a deep learning model into misclassifying the image. Computer scientists are actively researching defensive protocols, but the basic finding to date has been that it is remarkably easy to fool these models.

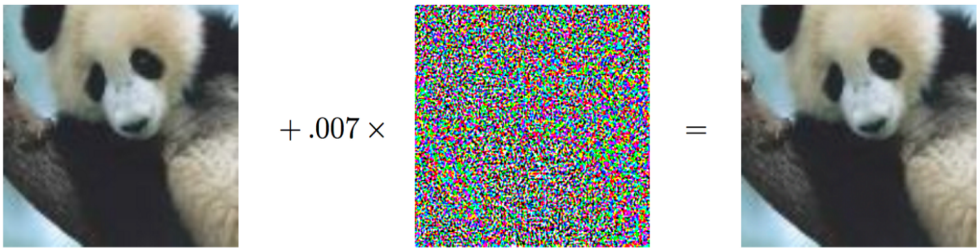

Figure 2: Example of "adversarial learning" to fool image recognition model into mis-classifying object.[157]

[156] Wojciech Samek, Thomas Wiegand & Klaus-Robert Müller, *Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models*. arXiv preprint arXiv:1708.08296 (2017).

[157] Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy, Explaining and Harnessing Adversarial Examples, arXiv preprint arXiv:1412.6572 (2014).

Consider an example of image similarity search piloted by the Patent and Trademark Office and Word Intellectual Property Organization in Figure 3. 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.

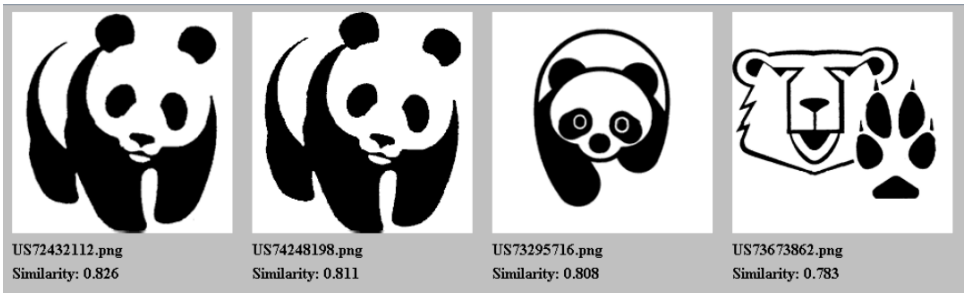

Figure 3: Example of prototype trademark similarity search model.[158] Images provide the first four search results based on a search of the World Wildlife Fund trademarked panda logo.

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 mis-classifying 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 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. 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. If prior enforcement actions are then used as training data, the system may unnecessarily confine enforcement actions to a distinct subset of all violations. This phenomenon is well-documented in the predictive policing context. When the predictive model is used to deploy police, and subsequent arrest data is used to re-train the model, a "runaway feedback loop" occurs: regardless of the crime rate, police may be sent to the same neighborhood over and over.[159] In the SEC context, AI tools may hence 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

[158] Christophe Mazenc, Machine Learning Applied to Trademarks Classification and Search, WIPO, 2018.

[159] Danielle Ensign et al., Runaway Feedback Loops in Predictive Policing, Conference of Fairness, Accountability, and Transparency, 2018, https://arxiv.org/abs/1706.09847.

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 appear to have solved the problem of arbitrariness, but only because of the fiction of adjudicator review.

In short, the current APA remain mechanisms remain ill-suited for providing meaningful accountability over rapid advances in AI.

III. REGULATING THE NEW ALGORITHMIC GOVERNANCE

As we have argued above, existing interpretations of the APA are unlikely to suitably respond to emerging uses of AI in the administrative state. In this section, we spell out several more affirmative responses, ranging from a minimalist retrofitting of the APA to a maximalist creation of a comprehensive oversight board. Reviewing and rejecting both of these options, we argue that the most compelling approach is 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 where agency uptake of algorithmic tools is apt to be most salient.

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 a human remains "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, for instance, a supervised learning algorithm that flags a case as "high risk" necessarily sets a (probability) threshold for classifying a case as "high risk." The lower the threshold, the greater the chance of false positives and the lower the chance of false negatives. Indeed, when the threshold is 0, 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 1, the greater the risk that human discretion is displaced by the

algorithm.[160] This fundamental tradeoff -- between false positives and false negatives -- determines the amount of human discretion that remains and cannot be determined absent a weighing of the social costs of each type of error.[161] 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.[162] 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.[163]

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.

2. *Reviewability*

Our suggestions for reviewability are distinct across adjudication and enforcement cases. 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 Insight system fails to parse a

[160] This is based on the assumption that humans reviewers are much less likely to pay attention to the large pool of predicted negatives.

[161] 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.

[162] Danielle Ensign et al., Runaway Feedback Loops in Predictive Policing, Conference of Fairness, Accountability, and Transparency, 2018, https://arxiv.org/abs/1706.09847.

[163] 900 F.2d 369 (D.C. Cir. 1990), judgmenl vacated and remanded for mootness, 111 S. Ct. 944 (1991).

particular functional impairment that is contested for a subgroup (e.g., balancing for individuals with chronic back pain[164]), 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*.[165] 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.[166]

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.[167] 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 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 would 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

---

[164] 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).

[165] 470 U.S. 821 (1985).

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

[167] 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 Shirin Sinnar, Protecting Rights from Within? Inspectors General and National Security Oversight, 65 Stan. L. Rev. 1027, 1035 (2013); Mariano-Florentino Cuellar, Auditing Executive Discretion, 82 Notre Dame L. Rev. 227, 256 (2006); Katyal, supra note ___, at __.

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 20 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 hence benefit from information exchange. 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 operational detail 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*

We argue that there is a less resource-intensive mechanism to produce the most important information about the operation, impact, and tradeoffs with the adoption of AI by federal regulatory agencies: prospective benchmarking.

The proposal starts from the basic setting that in each of the use cases above, machine learning is beginning to displace the exercise of human discretion in agency decisions. And because the status quo consists a fully human decision, this adoption process provides a compelling opportunity to benchmark the tool, by using the central insight of machine learning: use a random hold-out (or test) set to compare outcomes between the AI-assisted and human (status quo) decision. Whenever considering the

adoption of an AI use tool, agencies should be required to reserve a random test sample prospectively, for which conventional human decision-making would be deployed. In the SSA context, for instance, the Insight system could be deactivated for a random hold-out set. In the SEC context, investigators could be required to fully investigate cases without the aid of risk scores for a subset of cases. 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, and the only cost is that a subset of decisions would not garner the benefit of the new system.

Benchmarking would enable agencies, court, and the public to meaningfully assess the impact of AI use cases, promoting accountability and transparency without requiring the overhead of an oversight board or the uncertainty of rulemaking.

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 observe no serious attempt to compare the NLP flag against an investigator's read, particularly for disclosures that were not flagged. Even in instances where an agency offers evidence, it is unclear how much to attribute to the deployment of AI. Consider SSA's method to compare processing times and error rates between branches that voluntarily adopt clustering vs. branches that refuse to adopt clustering. This selection problem makes it impossible to disentangle the effects of micro-specialization from the effects of a managerial change.[168] Benchmarking enables decisionmakers to directly assess the impact of the AI tool in real time. The 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. The benchmark data would directly allow the agency to assess where the model needs to be calibrated. In addition, benchmarking enables agencies to assess whether the formal "human-in-the-loop" functionally ensures human oversight. It provides both a test for "automation bias" and ensures that agency officials do not lose the expertise required to process cases. Similarly, benchmarking would enable human investigators to understand when adversarial learning by regulated parties might invalidate historical models.

Second, the data generated from benchmarks provide invaluable information for 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. Data emerging from the benchmarked sample would provide the information to update models based on such state changes.

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.

How might such prospective benchmarking come about? First, Congress could statutorily either (a) mandate that agencies conduct prospective benchmarking, or (b)

[168] 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).

increase deference to algorithmic programs if benchmarking is instituted. Second, courts could find AI adoption without benchmarking evidence to be arbitrary and capricious. Third, the President could mandate benchmarking by executive order. Last, agencies themselves could institute such benchmarking under the Government Performance and Results Act. Indeed, benchmarking has a close analogue to "quality improvement" initiatives or audits that review a random sample of cases to calculate performance metrics. Current practices, however, 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.[169]

Regardless of the precise vehicle by which it is implemented, we believe benchmarking is both good machine learning practice and good governance. 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.

CONCLUSION

In this article, we have provided rich case studies of avant-garde deployments of AI in the federal government emerging out a major study for ACUS. As can be gleaned from these case studies, AI is increasingly moving to the center of administrative governance and the redistributive and coercive arms of the state. 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 AI reinvent government? The current use cases and internal innovation within agencies is promising, particularly when with rapid advances in AI research. But unless administrative law develops a coherent doctrinal and institutional approach to the governance of agency use of AI, this promise may ring as hollow as President Clinton's promise some 25 years ago.

[169] See, e.g., 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. (forthcoming, 2019).