Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies

REPORT SUBMITTED TO THE ADMINISTRATIVE CONFERENCE OF THE UNITED STATES

David Freeman Engstrom, Stanford University
Daniel E. Ho, Stanford University
Catherine M. Sharkey, New York University
Mariano-Florentino Cuéllar, Stanford University and Supreme Court of California

February, 2020

DISCLAIMER

This report was commissioned by the Administrative Conference of the United States in furtherance of its mission to "study the efficiency, adequacy, and fairness of . . . administrative procedure"; "collect information and statistics from . . . agencies and publish such reports as it considers useful for evaluating and improving administrative procedure"; and to "improve the use of science in the regulatory process." 5 U.S.C. §§ 591, 594. The opinions, views, and recommendations expressed are those of the authors. They do not necessarily reflect those of the Conference or its members.

Table of Contents

LEAD AUTHORS

David Freeman Engstrom. David Freeman Engstrom is the Bernard D. Bergreen Faculty Scholar and an Associate Dean at Stanford Law School. He is an elected member of the American Law Institute and a faculty affiliate at the Stanford Institute for Human-Centered Artificial Intelligence (HAI), CodeX: The Stanford Center for Legal Informatics, and the Regulation, Evaluation, and Governance Lab (RegLab). He received a J.D. from Stanford Law School, an M.Sc. from Oxford University, and a Ph.D. in political science from Yale University and clerked for Chief Judge Diane P. Wood on the U.S. Court of Appeals for the Seventh Circuit.

Daniel E. Ho. Daniel Ho is the William Benjamin Scott and Luna M. Scott Professor of Law, Professor of Political Science, and Senior Fellow at the Stanford Institute for Economic Policy Research at Stanford University. He directs the Regulation, Evaluation, and Governance Lab (RegLab) at Stanford, and is a Faculty Fellow at the Center for Advanced Study in the Behavioral Sciences and Associate Director of the Stanford Institute for Human-Centered Artificial Intelligence (HAI). He received his J.D. from Yale Law School and Ph.D. from Harvard University and clerked for Judge Stephen F. Williams on the U.S. Court of Appeals for the District of Columbia Circuit.

Catherine M. Sharkey. Catherine Sharkey is the Crystal Eastman Professor of Law at NYU School of Law. She is an appointed public member of the Administrative Conference of the United States, an elected member of the American Law Institute, and an adviser to the Restatement Third, Torts: Liability for Economic Harm and Restatement Third, Torts: Remedies projects. She was a 2011-12 Guggenheim Fellow. She received an M.Sc. from Oxford University and a J.D. from Yale Law School. She clerked for Judge Guido Calabresi of the U.S. Court of Appeals for the Second Circuit and Justice David H. Souter of the U.S. Supreme Court.

Mariano-Florentino Cuéllar. Mariano-Florentino Cuéllar is a Justice on the Supreme Court of California, the Herman Phleger Visiting Professor of Law at Stanford University, and a faculty affiliate at the Stanford Center for AI Safety. A Fellow of the Harvard Corporation, he also serves on the boards of the Hewlett Foundation, the American Law Institute, and the Stanford Institute for Human-Centered Artificial Intelligence (HAI), and chairs the boards of the Center for Advanced Study in the Behavioral Sciences and AI Now. He received a J.D. from Yale Law School and a Ph.D. in political science from Stanford University and clerked for Chief Judge Mary M. Schroeder of the U.S. Court of Appeals for the Ninth Circuit.

CONTRIBUTORS

Many dedicated professionals contributed to this report. To acknowledge these contributions, we list here the principal authors and contributors for each chapter and section.

Executive Summary and Introduction
Authors: Mariano-Florentino Cuéllar, David Freeman Engstrom, Daniel E. Ho, Catherine Sharkey, Liza Starr

Taking Inventory: A Survey of Federal Agency Use of AI
Authors: David Freeman Engstrom, Daniel E. Ho, Liza Starr
Contributors: Kinbert Chou, Shushman Choudhury, Madeline Levin, Coby Simler, Stephen Tang

Regulatory Enforcement at the Securities and Exchange Commission
Author: David Freeman Engstrom
Contributors: Sandhini Agarwal, Alex Duran, Michael Fischer, Joseph Levy, Sunny Kang

Law Enforcement at Customs and Border Protection
Authors: Nitisha Baronia, Cristina Ceballos, Mariano-Florentino Cuéllar, Daniel E. Ho
Contributors: Matthew Agnew, Peter Henderson, Geet Sethi, Stephen Tang

Formal Adjudication at the Social Security Administration
Authors: Daniel E. Ho, Derin McLeod
Contributors: Urvashi Khandelwal, Liza Starr, Emma Wang

Informal Adjudication at the U.S. Patent and Trademark Office
Authors: Daniel E. Ho, Urvashi Khandelwal, Alex Yu

Regulatory Analysis at the Food and Drug Administration
Author: Catherine Sharkey
Contributors: Cassi Carley, Shushman Choudhury, David Freeman Engstrom, Zach Harned, James Rathmell, Liza Starr, Chase Weidner

Public Engagement at the Federal Communications Commission and Consumer Financial Protection Bureau
Authors: Nitisha Baronia, Mariano-Florentino Cuéllar, David Freeman Engstrom, Daniel E. Ho
Contributors: Clint Akarmann, David Hoyt, Patrick Reimherr, Florian Tramer

Autonomous Vehicles for Mail Delivery at the United States Postal Service
Author: Shawn Musgrave

Building Internal Capacity
Authors: Nitisha Baronia, David Freeman Engstrom, Daniel E. Ho, Shawn Musgrave, Catherine Sharkey
Contributors: Cassi Carley, Ben Morris, Nate Tisa

Transparency and Accountability
Author: David Freeman Engstrom

Bias, Disparate Treatment, and Disparate Impact
Author: Daniel E. Ho

Hearing Rights and Algorithmic Governance
Authors: David Freeman Engstrom, Amit Haim, Daniel E. Ho

Gaming and Adversarial Learning
Authors: David Freeman Engstrom, Daniel E. Ho, Liza Starr

The External Sourcing Challenge: Contractors and Competitions
Authors: David Freeman Engstrom, Reed Sawyers

Several individuals also contributed to the report as a whole, including Ryan Azad, Jami Butler, Mikayla Hardisty, Alexandra Havrylyshyn, Luci Herman, and Liza Starr.

Executive Summary

Artificial intelligence (AI) promises to transform how government agencies do their work. Rapid developments in AI have the potential to reduce the cost of core governance functions, improve the quality of decisions, and unleash the power of administrative data, thereby making government performance more efficient and effective. Agencies that use AI to realize these gains will also confront important questions about the proper design of algorithms and user interfaces, the respective scope of human and machine decision-making, the boundaries between public actions and private contracting, their own capacity to learn over time using AI, and whether the use of AI is even permitted. These are important issues for public debate and academic inquiry.

Yet little is known about how agencies are currently using AI systems beyond a few headline-grabbing examples or surface-level descriptions. Moreover, even amidst growing public and scholarly discussion about how society might regulate government use of AI, little attention has been devoted to how agencies acquire such tools in the first place or oversee their use.

In an effort to fill these gaps, the Administrative Conference of the United States (ACUS) commissioned this report from researchers at Stanford University and New York University. The research team included a diverse set of lawyers, law students, computer scientists, and social scientists with the capacity to analyze these cutting-edge issues from technical, legal, and policy angles. The resulting report offers three cuts at federal agency use of AI:

- a rigorous canvass of AI use at the 142 most significant federal departments, agencies, and sub-agencies (Part I)
- a series of in-depth but accessible case studies of specific AI applications at seven leading agencies covering a range of governance tasks (Part II); and
- a set of cross-cutting analyses of the institutional, legal, and policy challenges raised by agency use of AI (Part III).

Taken together, these analyses yield five main findings.

First, the government's AI toolkit is diverse and spans the federal administrative state. Nearly half of the federal agencies studied (45%) have experimented with AI and related machine learning (ML) tools. Moreover, AI tools are already improving agency operations across the full range of governance tasks, including:

- *Enforcing* regulatory mandates centered on market efficiency, workplace safety, health care, and environmental protection;
- *Adjudicating* government benefits and privileges, from disability benefits to intellectual property rights;
- *Monitoring and analyzing* risks to public health and safety;
- *Extracting* useable information from the government's massive data streams, from consumer complaints to weather patterns; and
- *Communicating* with the public about its rights and obligations as welfare beneficiaries, taxpayers, asylum seekers, and business owners.

The government's AI toolkit spans the full technical scope of AI techniques, from conventional machine learning to more advanced "deep learning" with natural language and image data.

The government's AI toolkit is diverse and spans the federal administrative state. Nearly half of the federal agencies studied (45%) have experimented with AI and related machine learning (ML) tools.

Second, and despite wide agency embrace of AI, the government still has a long way to go. In canvassing agency use of AI, Stanford computer scientists evaluated the techniques deployed in each use case and, while limited public details precluded clear conclusions as to many, rated only 12% as high in sophistication. This is concerning because agencies will find it harder to realize gains in accuracy and efficiency with less sophisticated tools. This result also underscores AI's potential to widen, not narrow, the public-private technology gap.

Third, AI poses deep accountability challenges. When public officials deny benefits or make decisions affecting the public's rights, the law generally requires them to explain why. Yet many of the more advanced AI tools are not, by their structure, fully explainable. A crucial question will be how to subject such tools to *meaningful accountability* and thus ensure their fidelity to legal norms of transparency, reason-giving, and non-discrimination. The case studies presented in the report highlight several vital aspects of that challenge:

- Transparency's costs, benefits, and feasibility will vary across policy areas, governance tasks, and AI techniques. Open-sourcing of technical details might be appropriate when agencies are allocating social welfare benefits but can undermine agency use of valuable enforcement tools because of gaming by regulatory targets.
- One key area for future inquiry is how to adapt existing principles of administrative law, which is more likely to modulate agency use of AI than the constitutional constraints that occupy much current debate.
- Policymakers should also consider other interventions. A promising candidate is to require agencies to engage in prospective "benchmarking" of AI tools by reserving a random hold-out sample of cases for human decision, thus providing critical information to smoke out when an algorithm has gone astray or "automation bias" has led decision-makers to excessively defer to an algorithm.

To achieve meaningful accountability, concrete and technically-informed thinking within and across contexts—not facile calls for prohibition, nor blind faith in innovation—is urgently needed.

Fourth, if we expect agencies to make responsible and smart use of AI, technical capacity must come from within. While many agencies rely on private contractors to build out AI capacity, a majority of profiled use cases (53%) are the product of in-house efforts by agency technologists. This underscores the critical importance of internal agency capacity building as AI continues to proliferate. In particular:

- In-house expertise promotes AI tools that are better tailored to complex governance tasks and more likely to be designed and implemented in lawful, policy-compliant, and accountable ways. Sustained collaboration between agency officials and in-house technologists facilitates identification of appropriate questions, seizing new innovations, and evaluating existing tools, including contractor-provided ones.
- Fully leveraging agency use of AI will require significant public investment to draw needed human capital and update outmoded data and computing systems. Given fiscal and labor market constraints, agencies should also explore non-commercial sources of valuable technical capacity, including collaborations with universities, NGOs, and industry and agency-sponsored competitions.

In-house expertise yields AI tools that are better tailored to complex governance tasks and more likely to be implemented in a lawful, policy-compliant, and accountable fashion.

Fifth, AI has the potential to raise distributive concerns and fuel political anxieties. Growing agency use of AI creates a risk that AI systems will be gamed by better-heeled groups with resources and know-how. An enforcement agency's algorithmic predictions, for example, may fall more heavily on smaller businesses that, unlike larger firms, lack a stable of computer scientists who can reverse-engineer the agency's model and keep out of its cross-hairs. If citizens come to believe that AI systems are rigged, political support for a more effective and tech-savvy government will evaporate quickly.

To achieve meaningful accountability, concrete and technically-informed thinking within and across contexts—not facile calls for prohibition, nor blind faith in innovation—is urgently needed.

In sum, 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 hollow out the human expertise inside agencies with few compensating gains, widen the public-private technology gap, increase undesirable opacity in public decision-making, and heighten concerns about arbitrary government action and power. Given these stakes, agency administrators, judges, technologists, legislators, and academics should think carefully about how to spur government innovation involving the appropriate use of AI tools while ensuring accountability in their acquisition and use. This report seeks to stimulate that thinking.

Introduction

Americans depend on the federal government not only to provide for the common defense and promote general welfare, but to protect the environment, advance public health, promote innovation, and implement labor and employment standards. As federal agencies develop new rules and guidance and adjudicate, enforce, and otherwise implement statutory policies, they encounter a constantly changing economic, social, and technological context. The growing sophistication of and interest in artificial intelligence (AI) and machine learning (ML) is among the most important contextual changes for federal agencies during the past few decades.

While many scholars and commentators have speculated about how government should regulate AI, we know precious little about how government agencies themselves use AI. Getting an accurate picture of such use today is critical for developing a national AI strategy that can help guide the country's approach to AI in the future, for modernizing the public sector, and for instituting appropriate safeguards to govern the adoption and use of AI. The recently proposed AI in Government Act,[1] for instance, aims to "improve the use of AI across the federal government by providing access to technical expertise and streamlining hiring within the agencies."[2] At the same time, there is mounting resistance against some of the most controversial uses of AI. As the most visible example, a number of jurisdictions have recently moved to ban use of facial recognition systems, and similar efforts have begun to percolate in Congress.[3] Without understanding how government agencies develop and deploy emerging AI technologies, it is difficult to craft sensible and workable prescriptions.

We realize that crafting those prescriptions—and, indeed, addressing any of the most important issues involving government adoption of AI—can prove contentious, and rightly so. Some observers are concerned that AI will further enhance government power, enabling surveillance that could threaten privacy and civil liberties. Others express concern that AI will further disempower marginalized groups. And still others take the view that the power of AI in the private sector, without appropriate knowledge in the public sector, can undermine agencies' capacity to achieve regulatory goals. Ultimately, our goal here is not to take any single categorical position on the normative desirability of any specific government use of AI or ML tools. Instead, our aim is to understand how agencies are currently using this technology and identify the most important legal and policy implications it presents.

The New Algorithmic Governance

The use of AI-based tools to support government decision-making, implementation, and interaction—what could be called "algorithmic governance"—already spans the work of the modern administrative state. Table 1 previews some of the use cases explored in this report and advances a typology of governance tasks to which agencies are applying AI. Among these are two core tasks of modern government: enforcing regulatory mandates ("enforcement") and adjudicating benefits and privileges ("adjudication"). However, federal-level use cases span well beyond enforcement and adjudication to other critically important governance tasks, such as regulatory analysis, rulemaking, internal personnel management, citizen engagement, and service delivery.

TABLE 1. ALGORITHMIC GOVERNANCE TOOLS BY USE CATEGORIES

Use Type	Description	Examples
Enforcement	Tasks that identify or prioritize targets of agency enforcement action	• Securities and Exchange Commission, Centers for Medicare and Medicaid Services, and Internal Revenue Service predictive enforcement tools • Customs and Border Protection and Transportation Security Administration facial recognition systems • Food Safety and Inspection Service prediction to inform food safety site testing
Regulatory research, analysis, and monitoring	Tasks that collect or analyze information that shapes agency policymaking	• Consumer Financial Protection Bureau analysis of consumer complaints • Bureau of Labor Statistics coding of worker injury narratives • Food and Drug Administration analysis of adverse drug events
Adjudication	Tasks that support formal or informal agency adjudication of benefits or rights	• Social Security Administration system for correcting adjudicatory errors • U.S. Patent and Trademark Office tools for adjudicating patent and trademark applications
Public services and engagement	Tasks that support the direct provision of services to the public or facilitate communication with the public for regulatory or other purposes	• U.S. Postal Service autonomous vehicles project and handwriting recognition tool • Department of Housing and Urban Development and U.S. Citizenship and Immigration Services chatbots • Agency analysis of submitted rulemaking comments
Internal management	Tasks that support agency management of resources, including employee management, procurement, and maintenance of technology systems	• Department of Health and Human Services tool to assist procurement decision-making • General Services Administration tool to ensure legal compliance of federal solicitations • Department of Homeland Security tool to counter cyberattacks on agency systems

Table 1 also provides important context for current AI innovation by situating the new algorithmic governance tools in the context of past government innovation. In one sense, the new algorithmic governance tools build on several decades of federal government experimentation with data mining—to identify criminal suspects, monitor suspicious banking practices, and administer transportation security—that created flash-points around government privacy and cybersecurity practices in the 2000s.[4] Other tools harken back even further, to efforts in the 1990s to "reinvent government" through data-based performance management and oversight.[5] Finally, the new algorithmic governance tools have plain antecedents in "expert systems" championed throughout the 1960s and 1970s by Herbert Simon to rationalize and evaluate administrative behavior.[6] Such systems relied on input by domain experts to craft logical rules to automate decision-making.

Yet these new algorithmic governance tools differ from prior technological innovation in three important ways. First, these tools are more *inscrutable* in that even a system's

engineers may not fully understand how it arrived at a result.[7] In an expert-based system, the logical rules are written as conditional (if-then) statements. In traditional statistical analysis, outcomes are modeled with relatively few explanatory variables and the resulting models remain relatively simple. By contrast, state-of-the-art machine learning deploys far more complex models to learn about the relationship across hundreds or even thousands of variables. Model complexity can make it difficult to isolate the contribution of any particular variable to the result.

Second, and relatedly, machine learning outputs are often *nonintuitive*—that is, they operate according to rules that are so complex, multi-faceted, and interrelated that they defy practical inspection, do not comport with any practical human belief about how the world works, or simply lie beyond human-scale reasoning.[8] Even if data scientists can spell out the embedded rule, such rules may not tell a coherent story about the world as humans understand it, defeating conventional modes of explanation.[9]

Because machine learning can yield counter-intuitive results with flaws that can be difficult to detect, observers may not consider the results fully "accountable," [10] even when they have a detailed indication of how an algorithmic system works.[11] To be sure, some of these concerns may diminish over time with continued advances in "explainable AI"—a term that describes an emerging set of techniques that has shown promise in rendering machine learning models more interpretable by ranking, sorting, and scoring data features according to their pivotalness in the model or by using visualization techniques or textual justifications to lay bare a model's decision "pathway."[12] But technical challenges remain, especially with more complex algorithmic models. For the moment, surprisingly little is known, for example, about how and why the most advanced neural networks work.[13]

Third, the new algorithmic governance tools differ from past rounds of public sector innovation in the sense that they are often more *deeply embedded* in the work of government. As Table 1 illustrates, more powerful analytic methods have made possible the automation of a wider range of government tasks than before.[14] Importantly, the expanding menu of applications, particularly those that perform enforcement and adjudication tasks, is rapidly moving the new algorithmic governance tools to the center of the coercive and (re-) distributive power of the state.[15] In addition, the growing sophistication and power of AI is nudging agencies toward fully automated decision-making, leaving progressively less to human discretion and judgment.[16] Government officials who use those tools may, to borrow from the AI lexicon, be increasingly left "out of the loop." Finally, leaps in analytic power mean more displacement of discretion at all levels of bureaucracy. Growing sophistication may permit algorithmic tools to continue "steadily climb[ing] up the bureaucratic ladder," shaping, and in some cases displacing, the decisions of more senior agency decision-makers.[17] At the same time, the impact of AI systems on administrative government also goes in the opposite direction: What could be called an "IT-level bureaucracy" is in some cases increasingly displacing the smaller-scale and more numerous decisions of the "street-level bureaucrats" that perform much of the visible, citizen-facing work of government.[18]

Goals

Understanding these features of the new algorithmic governance toolkit is critical. To aid that understanding, this report has three principal goals.

First, this report aims to inform the trajectory of AI use in government by understanding whether, how, and why agencies are beginning to use these tools. Agencies often face daunting constraints. Innovation and continuous improvement––including through the use of AI/ML––can help agencies achieve their challenging missions with integrity, efficiency, and fairness. A variety of risks and opportunities exist in domains ranging from cybersecurity to public engagement in the regulatory process. We can better understand those risks and opportunities, as well as the challenges agencies will face as they adapt to an increasingly AI-driven world, if we know how agencies are experimenting with these technologies. Over time, more specific metrics of agency use of AI tools can help policymakers identify opportunities for improvement and remedy deficiencies.

Second, our report aims to spell out how these new tools raise new and challenging questions in law and policy about fairness, transparency and accountability, due process, and capacity building. While the new algorithmic governance tools hold the promise of more accurate and consistent government decisions, their opacity also creates myriad legal puzzles because of administrative law's core commitment to transparency and reason-giving when government takes actions that affect rights. In the years ahead, judges, lawyers,

agency administrators, and legislators will have to face these legal quandaries. Not only does the law require an answer to those questions, but the continuing development of AI/ML systems can benefit from engagement with the various legal and governance issues raised by the use of such technology in administrative agencies. This report begins to map and assess these issues.

Third, our analysis sketches out promising directions for future research. If that research is to inform a robust understanding of how federal agencies can better achieve the many (often contradictory) demands placed on them, it must engage with agencies' actual practices and legal responsibilities. By canvassing agency practices and then offering more detailed case studies at an important moment in the history of government use of AI, we hope to catalyze further work in this area and perhaps even parallel efforts on state and local agencies as well as international entities.

Scope

AI technologies and the federal government are both vast. To be clear about the scope of our work, it is worth defining terms and delineating what this report covers and what it does not.

By "artificial intelligence," we limit our scope to the most recent forms of machine learning, which train models to learn from data. These include a range of methods (*e.g.,* neural networks, random forests) capable of recognizing patterns in a range of types of data (*e.g.,* numbers, text, image)—feats of recognition that, if undertaken by humans, would be generally understood to require intelligence. The definition includes both "supervised learning," where "training data" is used to develop a model with features to predict known "labels" or outcomes, and "unsupervised learning," where a model is trained to identify patterns in data without labels of interest. Conceptually, AI includes a range of analytical techniques, such as rule-based or "expert" symbolic systems,[19] but we limit our focus to forms of machine learning. Our scope also excludes conventional forms of statistical inference (*e.g.,* focused on causal, as opposed to predictive, inference) and forms of process automation that do not involve machine learning (*e.g.,* an online case management system).[20] We often use the shorthand "AI/ML" to describe the family of tools and techniques falling within the above definition.

By "federal agencies" we mean executive departments and their sub-components as well as independent agencies. While the project aspired to the widest possible scope in investigating the growing role of AI in the federal administrative state, we limit the set of agencies investigated in two ways. First, we do not examine military and intelligence agencies (*e.g.,* the National Security Agency; the Department of Defense; agencies working on cyber-defense) because it is difficult to obtain reliable publicly available information from such agencies. Second, to make our inquiry tractable given a rapidly changing landscape, we limit ourselves to the 142 largest and most prominent federal agencies, as set forth below.

By "use case," we focus on the use of AI for core agency functions. We do not examine agency use of traditional regulatory methods to monitor or regulate industries and other private sector actors who are themselves deploying AI systems (*e.g.,* the SEC's regulation of high-frequency trading algorithms or NHTSA's regulation of autonomous vehicles using conventional rulemaking). That said, substantial industry reliance on AI technology will serve as a useful flag or marker in identifying actual and prospective agency use of AI in administrative decision-making, so the report offers some preliminary insights on the topic.

Roadmap

The rest of this report proceeds as follows.

Part I provides the results of a systematic survey of federal agency use of AI. We examine a wide range of public evidence—including agency websites, news articles, press releases, congressional testimony, and mandated data mining reports—to develop a portrait of AI adoption at the largest 142 federal administrative agencies (measured by full-time equivalent employees). We examine whether there is evidence that the agency has experimented or adopted AI/ML, the policy area and task to which such use cases are devoted, how such use cases were developed (*e.g.,* in-house vs. contractor vs. competitions or other non-commercial sources), underlying technology (*e.g.,* supervised vs. unsupervised machine learning models), data source, and level of sophistication. We cannot claim perfect comprehensiveness from this analysis, particularly given our reliance on publicly available sources. Moreover, the technology is rapidly changing, so much so that the landscape has surely changed at numerous agencies since we embarked on this study. However, in adhering to a common protocol, our aim is to provide a rigorous portrait of AI use

that can help policymakers, agency officials, academics, and other interested persons to understand where the federal administrative state is and where it might be heading.

Part II offers a set of rich case studies of AI innovation at specific federal agencies. One limitation of Part I's survey is that it is based on publicly available sources. Such sources rarely provide sufficient technical detail on AI systems and offer little insight on the process of generating such use cases. Part II's case studies overcome this limitation by relying on extensive interviews with federal officials. The particular case studies were chosen to illustrate the range of agencies, use cases, and types of technology being deployed. In the enforcement context, we study the tools developed by the Securities and Exchange Commission (SEC) and Customs and Border Protection (CBP). With respect to agency adjudication, we focus on the Social Security Administration (SSA) and the U.S. Patent and Trademark Office (PTO). For regulatory analysis, we examine a pair of pilots at the Food and Drug Administration (FDA). For citizen engagement, we examine the role of emerging tools for computationally assisted processing of complaints and comments in rulemakings, exemplified in the context of the Federal Communications Commission (FCC) and the Consumer Financial Protection Bureau (CFPB). And to assess the potential for improving citizen services, we examine a pilot in automated mail delivery by the U.S. Postal Service (USPS).

Part III turns to implications and recommendations that cut across particular tools and governance tasks. We cover six major areas: (1) the challenges of building AI capacity in the public sector, including data infrastructure, human capital, and regulatory barriers; (2) the difficulties inherent in promoting transparency and accountability; (3) the potential for unwanted bias and disparate impact; (4) potential risks to statutory hearing rights and due process; (5) risks and responses associated with gaming and adversarial learning, and (6) the use of contracting and procurement to supplement agency technical expertise and capacity.

Because so little is known about federal agency usage of AI/ML, the case studies contained in this report are lengthy and richly detailed and can each be read independently. For readers interested in the highlights, we provide a short summary of takeaways at the beginning of each case study. Readers short on time are advised to read the canvass in Part I, the case study highlights in Part II, and then Part III's cross-cutting implications.

Acknowledgments

We are grateful for the willingness of ACUS and its leadership to support this project. Given ACUS's mission of fostering improvement in federal agencies' procedures while engaging lawyers, scholars, and government officials, we consider this project to fit well with the agency's broad, nonpartisan mission.

Many others contributed time and resources to this effort. We appreciate generous support from Stanford Law School, Stanford's Institute for Human-Centered Artificial Intelligence, NYU Law School, and the Stanford Institute for Economic Policy Research.

We are likewise grateful to the many dedicated agency officials and staff who shared their knowledge and views with us, during countless meetings, site visits, phone calls, and a roundtable meeting at NYU Law School in February 2019. They are too many to name here, and some agencies specifically requested that we not reference by name any personnel. We nonetheless wish to direct special thanks to the following present and former public servants for their time, advice, and help connecting us with resources at various stages of the project: Scott Bauguess, Jeff Butler, Kurt Glaze, Lynne Parker, Gerald Ray, James Ridgway, Paul Verkuil, and Matthew Wiener.

We received valuable feedback from participants at the American Bar Association's Administrative Law Conference, an ACUS Plenary Session, the Ethics, Policy and Governance Conference by Stanford's Institute for Human-Centered Artificial Intelligence, the Technology, Innovation, and Regulation Conference at the C. Boyden Gray Center for the Study of the Administrative State, a meeting of the Regulation, Evaluation, and Governance Lab at Stanford, and the faculty workshop at the School of Law at the University of Texas at Austin. Numerous colleagues provided exceptionally helpful comments, including Scott Hemphill, David Marcus, Lisa Ouellette, Arti Rai, and Chris Walker.

Particular thanks go to the remarkable students from throughout Stanford University and NYU Law School who enrolled in the policy practicum that catalyzed work on this project in the winter of 2019. These students included 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 Weidner, and Alex Yu. Luci Herman provided sage advice for the student teams in the practicum. The efforts of this group, along with what's been done by the many thoughtful public officials with whom we've interacted over the past year, are a testament to the value of integrating legal, policy, and technical knowledge and showcase some of the possibilities for continuing to improve the work of the public sector in the years to come. We also appreciate superb editorial assistance and dedicated research support from three members of Justice Cuéllar's staff: Ryan Azad, Alexandra Havrylyshyn, and Mikayla Hardisty. Jami Butler helped with the graphic design of the report.

Finally, we are grateful to a core team of stellar research assistants who played critical roles in helping us synthesize the material for the report: Nitisha Baronia, Kinbert Chou, Amit Haim, Zach Harned, Maddie Levin, Shawn Musgrave, Reed Sawyers, Coby Simler, and Liza Starr.

All of these participants in the work associated with this report, along with innumerable agency officials who engaged with us, helped us gain understanding of how an important technology is poised to reshape the venerable administrative state. The possibilities to improve agency performance and benefit the public are as striking as some of the risks. Making the most of the former while minimizing the latter will take resources, time, and wisdom. We hope the pages to come—replete with rich information about what's currently occurring and what agencies will seek to do in the years to come—make a lasting contribution on that score.

Part I. Taking Inventory: A Survey of Federal Agency Use of AI

Where, and for what purposes, are federal agencies developing and deploying algorithmic governance tools? What are the principal types of AI techniques federal agencies are developing and deploying? And what are the primary sources of AI-based governance tools—in-house agency technologists, the procurement process, or some other channel?

To answer these questions, our research team of lawyers, social scientists, and computer scientists identified and characterized possible AI use cases at the 142 most significant federal departments, agencies, and subagencies—collectively referred to hereafter as "agencies." This Part presents our empirical results and provides a broad overview of how federal agencies are using AI-based tools to perform the work of governance. By situating federal government use of AI in a wider context, we provide a first-of-its-kind snapshot of the current state of federal government development and deployment of AI.

Methodology

To generate a rigorous portrait of government use of AI, we began by identifying the most significant federal agencies. We started with the ACUS Sourcebook of U.S. Executive Agencies, which lists roughly 300 agencies, bureaus, and offices, including independent agencies.[1] To focus on the most substantial agencies, we trimmed this list to agencies with at least 400 employees, removing 125 agencies.[2] We also excluded 21 active military and intelligence-related organizations (*e.g.,* the Defense Information Systems Agency and the Missile Defense Agency), leaving us with 142 agencies overall.

We then relied on a wide range of sources to search for evidence that the agency had considered deploying an AI/ML use case. We defined a use case as an instance in which an agency had *considered using* or had *already deployed* AI/ML technology to carry out a core function. We did not count instances where agencies demonstrated no intent to operationalize a given tool—for example, a pure research paper using AI/ML conducted by an economist at the Federal Reserve with little direct connection to the agency's regulatory or other duties. Sources included industry and nonprofit reports, congressional testimony, press releases, agency websites, mandated data mining reports, and academic studies.[3]

After substantial piloting (including a survey sent to agency officials via ACUS), the most reliable protocol for identifying AI use cases was an agency-by-agency, web-based search protocol, augmented by a range of third-party sources.[4] We compiled these results with agency use cases as distinct observations.

We note at the outset that this methodology is limited in several respects. First, our results reflect searches conducted during January through August of 2019. The technology is developing rapidly, so the aggregated results should only be considered a snapshot-in-time. Second, it was not always easy to determine the boundary between AI use cases to perform core agency functions and pure research. Third, use cases are defined by what information is publicly available. It is possible that access to non-public, pan-government information

would yield a different picture. A final challenge we faced was determining what constituted a use case based on often limited technical and operational documentation. Numerous agencies touted their use of tools to "automate" functions or their application of "predictive analytics," but many of these tools would not necessarily be considered a form of modern machine learning and were excluded.[5] We attempted to resolve boundary issues as well as we could through multiple rounds of quality control.[6]

Contrary to popular perceptions presuming government agencies uniformly rely on antiquated systems and procedures, many agencies have in fact experimented with AI/ML.

Results

The results of this survey shed significant light on the state of AI/ML in federal administrative agencies.

First, contrary to popular perceptions presuming government agencies uniformly rely on antiquated systems and procedures, many agencies have in fact experimented with AI/ML. Nearly half (64 agencies, or 45%) of canvassed agencies have expressly manifested interest in AI/ML by planning, piloting, or implementing such techniques. To offer a flavor, the National Oceanic and Atmospheric Administration is using AI to refine high-impact weather tracking systems to improve decision-making in real-time. The Transportation Security Administration is exploring the use of image recognition to screen passenger luggage for explosive devices. The Centers for Medicare and Medicaid Services is developing AI-based tools to predict health care fraud. And the Department of Housing and Urban Development deployed a prototype chatbot to enable citizens to acquire information about rental assistance, agency programs, and civil rights complaint procedures.

Second, many agencies have pioneered multiple AI/ML use cases. We documented 157 use cases across 64 agencies. AI usage is heavily concentrated in a small number of agencies, with about 7% of canvassed agencies responsible for 70% of all identified use cases. Table 2 lists the number of use cases at the top 10 adopters. A large number of use cases fell under health- and law-enforcement-focused subagencies such as the Food and Drug Administration, the Office of Justice Programs, and the Transportation Safety Administration and Customs and Border Protection. As a result, the Department of Health and Human Services, the Department of Justice, and the Department of Homeland Security account for a collective 51 use cases. Perhaps unsurprisingly, NASA has also rapidly adopted AI. For instance, NASA developed a prototype cockpit advisor system based on IBM's Watson, which would enable pilots to query a knowledge base for situationally relevant information.

TABLE 2. TOP TEN AGENCIES AND SUBAGENCIES BY NUMBER OF USE CASES

Agency Name	Number of Use Cases
Office of Justice Programs	12
Securities and Exchange Commission	10
National Aeronautics and Space Administration	9
Food and Drug Administration	8
United States Geological Survey	8
United States Postal Service	8
Social Security Administration	7
United States Patent and Trademark Office	6
Bureau of Labor Statistics	5
Customs and Border Protection	4

Table 2: The above list excludes overarching department-level agencies. For example, the Department of Health and Human Services (19 use cases), the Department of Justice (16 use cases), and the Department of Homeland Security (16 use cases) have been refactored into respective sub-agencies (e.g., the Food and Drug Administration, the Office of Justice Programs, and Customs and Border Protection). In addition, note that three other agencies or subagencies other than CBP have four use cases: the Board of Governors of the Federal Reserve System, the National Oceanic and Atmospheric Administration, and the Federal Bureau of Investigation.

FIGURE 1. AI USE CASES BY POLICY AREA

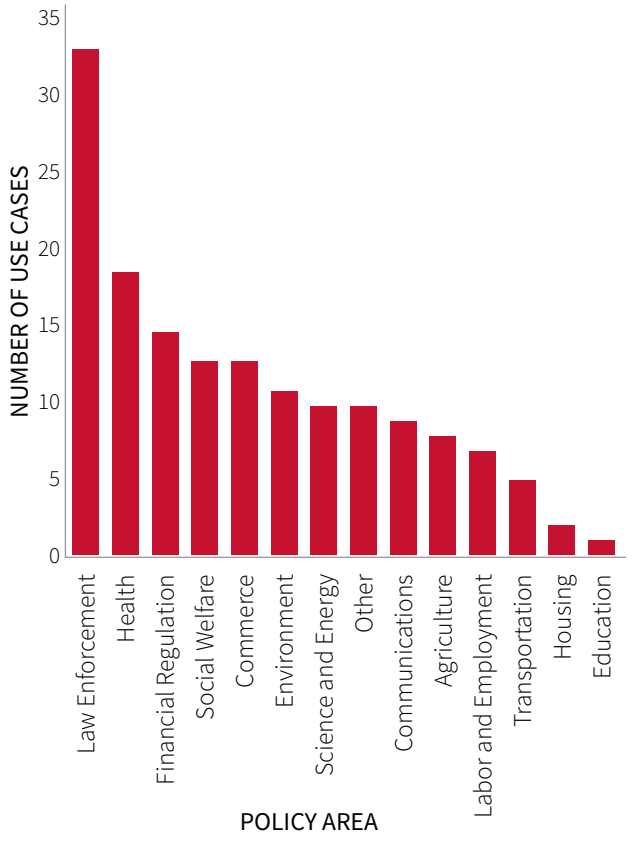

Figure 1: Each bar represents the count of the number of use cases by policy area. For simplicity, each agency was coded as falling into one primary policy area. Science and Energy were combined into one category. Some agencies that were coded as occupying two primary fields were collapsed (*e.g.,* EPA was classified as 'Environment'; the National Institute of Food and Agriculture as 'Agriculture'; the International Trade Commission as 'Commerce'; the Consumer Financial Protection Bureau as 'Financial Regulation'; and the Railroad Retirement Board as 'Social Welfare').

Third, agency AI/ML use is spread across a wide range of policy areas. As reflected in Figure 1, the top three policy areas were in law enforcement,[7] health, and financial regulation. But Figure 1 also shows that use cases span virtually all other substantive policy areas, such as environment, energy, social welfare, and communications. This highlights the breadth of AI use and impact.

FIGURE 2. AI USE CASES BY GOVERNANCE TASK

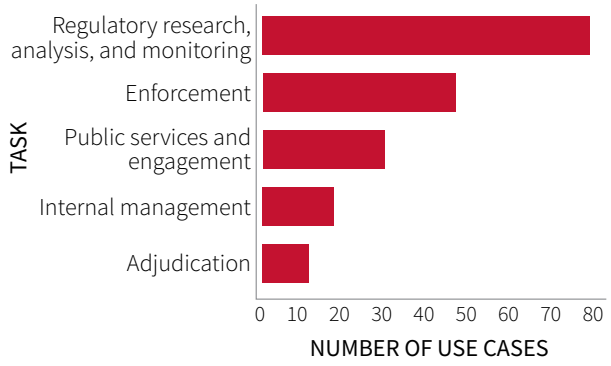

Figure 2: Each bar represents the count of the number of use cases for each task. The 'Regulatory research, analysis, and monitoring' category includes tools (or research) that involve collecting and analyzing information to inform agency policymaking. The 'Enforcement' category includes use cases that support or lead to enforcement actions, including monitoring tasks for finding and tracking violations. The 'Public services and engagement' category includes tools that facilitate the provision of services to or communication with the public for regulatory or other purposes. The 'Internal management' category includes tools to support all other internal agency management functions, including employee management and procurement. The 'Adjudication' category includes tools that aid in formal or informal adjudication of benefits or rights. Coding was keyed to the primary purpose of each use case. Twenty-four use cases received multiple codings (*e.g.,* both 'Regulatory research, analysis, and monitoring' and 'Public services and engagement' for a tool that analyzes consumer complaints).

Fourth, agency AI/ML use serves diverse regulatory functions. As shown in Figure 2, agencies use AI to prioritize enforcement (*e.g.,* prediction of potential violators of the federal securities laws at the SEC), engage with the public (*e.g.,* a United States Citizenship and Immigration Services chatbot to provide assistance answering immigration questions), conduct regulatory research, analysis, and monitoring (*e.g.,* Department of Health and Human Services tools to predict adverse drug events and unplanned hospital admissions), and adjudicate rights and benefits (*e.g.,* United States Patent and Trademark Office tools to support patent and trademark determinations).[8]

Agency AI/ML use is spread across a wide range of policy areas and serves diverse regulatory functions.

FIGURE 3. AI USE CASES BY IMPLEMENTATION STAGE

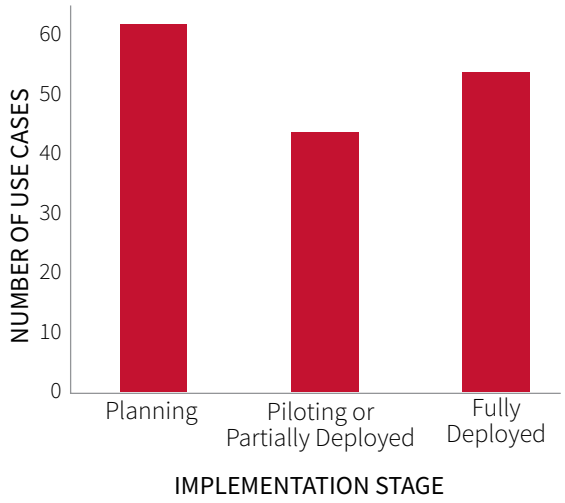

Figure 3: The 'Planning' category includes cases where the AI tool is not yet built, though an agency or one of its representatives has expressed interest or committed to it being deployed. The 'Piloting or Partially Deployed' category includes all use cases currently under technical development or testing, and 'Fully Deployed' entails a use case that has been adopted and integrated into the agency's governance pipeline.

Fifth, agency AI/ML uses vary in their stage of development. Figure 3 allocates use cases to three implementation stages—planning, piloting/partially deployed, and fully deployed. On the one hand, only one-third (53 use cases, or 33%) of the applications are fully deployed. On the other hand, the sheer amount of planning and piloting is a testament to how much attention is currently being devoted to scoping out usage of AI/ML.

FIGURE 4. AI USE CASES BY DEVELOPER TYPE

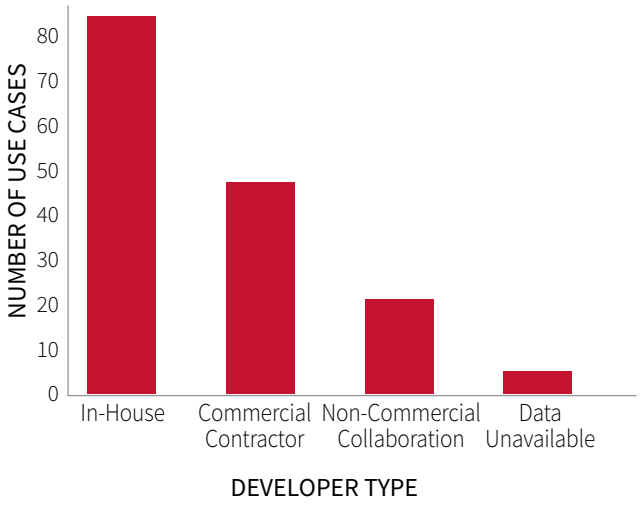

Figure 4: Each bar represents the number of use cases procured through the respective category of developer.

Sixth, agency AI/ML use cases vary in their development source. Figure 4 presents the primary developer of the application. We denote whether the use case was built in-house by agency staff, by third-party (commercial) contractors, by non-commercial collaboration (*e.g.,* collaboration with an academic lab, public-facing competitions), or a mix. Contrary to much of the literature's fixation on the procurement of algorithms through private contracting, over half of applications (84 use cases, or 53%) were built in-house, suggesting there is substantial creative appetite within agencies.

Contrary to much of the literature's fixation on the procurement of algorithms through private contracting, over half of applications (84 use cases, or 53%) were built in-house, suggesting there is substantial creative appetite within agencies.

FIGURE 5. AI USE CASES BY MACHINE LEARNING METHOD

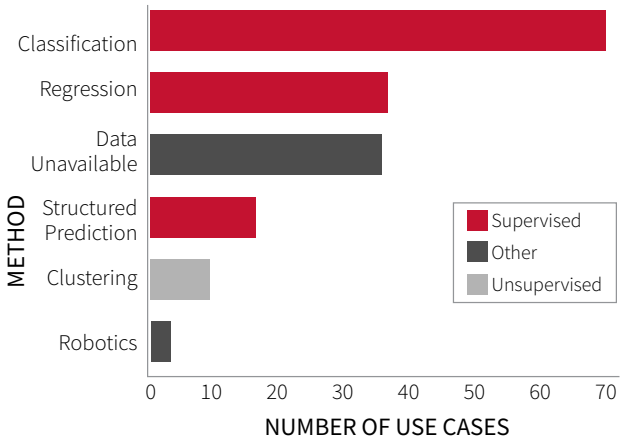

Figure 5: Each bar represents the number of use cases utilizing the respective method. For four instances of use cases with multiple methods (*e.g.,* clustering and classification), each applicable method is counted and reported. Colors indicate broad typology of supervised (red), unsupervised (gray), or other type (charcoal) of learning.

Seventh, administrative agencies are experimenting with a range of AI/ML methods. To understand the distribution of methods, we classified each use case as (a) supervised versus unsupervised learning (or other), (b) type of supervised learning (regression for continuous outcomes, classification for categorical labels, structured prediction (*e.g.,* chatbots)), and (c) type of unsupervised learning (clustering or dimensionality reduction). Figure 5 displays the breakdown of methods. The dominant method by far is supervised learning (red), comprising nearly 71% of all agency use cases.[9] Unsupervised learning and robotics are much less prevalent in the administrative state, at least for the time being.

Eighth, agency AI/ML use leverages a range of data. Figure 6 provides a breakdown of the types of datasets being used. The vast majority (78%) of applications rely on structured or text data. Despite rapid advances in computer vision, for instance, fewer agencies have begun to rely on image, sound, or other forms of unstructured data.

FIGURE 6. AI USE CASES BY DATA TYPE

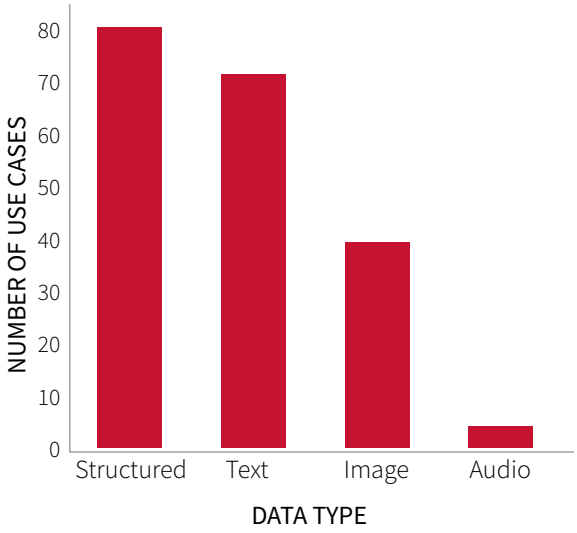

Figure 6: 'Structured' data includes numerical information and other factored variables. In 32 use cases, multiple data types formed the basis for an application (*e.g.,* one use case might include both textual and image data). In those instances, we counted each application-data type.

Last, our team of computer scientists assessed the technical sophistication of each use case. To illustrate the scale used, we considered: (a) logistic regression using structured data to be of lower sophistication; (b) a random forest with attention to hyperparameter tuning to be of medium sophistication; and (c) use of deep learning to develop "concept questioning" of the patent examination manual to be of higher sophistication. Here lies the most sobering finding: For most government applications (61%), there is insufficient publicly available technical documentation to determine with precision what methods are deployed. In some cases, the agency's description appears more like marketing language or concerns a tool still under development. In other cases, agencies describe use of neural networks, natural language processing, or facial recognition technologies but do not provide enough technical details to discern whether a use case is a simpler or more sophisticated version thereof. We did not make judgments solely based on task or incantation

FIGURE 7. AI USE CASES BY LEVEL OF SOPHISTICATION

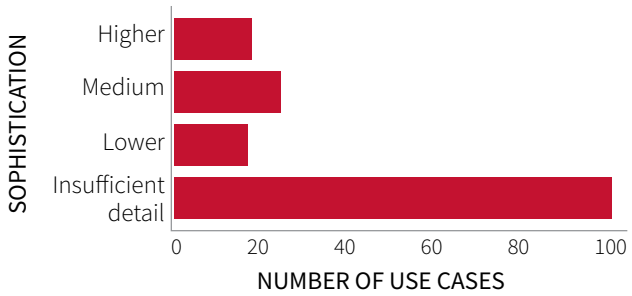

Figure 7: Use cases with an 'Insufficient Detail' sophistication rating either do not divulge sufficient technical details or are still under development.

of method. The result, as presented in Figure 7, is that only a minority of agency use cases (12%) could clearly be rated as higher in sophistication by our team of computer scientists. Even among the sample of use cases for which sophistication could be discerned, only 31% would rate as 'Higher' in sophistication. While the deep learning revolution has rapidly transformed the private sector, it appears to have only scratched the surface in public sector application.

While the deep learning revolution has rapidly transformed the private sector, it appears to have only scratched the surface in public sector application.

These results on sophistication should be taken with a grain of salt. Reasonable people can disagree about comparative assessments of sophistication, particularly given different domains and availability of documentation. Moreover, available documentation likely skews toward older technology. Finally, as just noted, in the most common case it was difficult to discern the level of sophistication ("insufficient detail"). There may be additional highly sophisticated use cases that our team of computer scientists could not clearly categorize as such, though these are likely relatively few in number. Overall, the results suggest there is considerable room for improvement and development.

*The Appendix is available at www.law.stanford.edu/ACUS-AI-Report

To sum up, this survey provides much needed grounding of the landscape of federal government use of AI. To that end, we provide summary information about all use cases in the online Appendix* to this report. However, while these descriptive statistics paint a rich portrait of government innovation, they do not fully capture the process, technical detail, and institutional setting of AI in government. For that, we turn to in-depth case studies.

Part II. Case Studies of Federal Agency Deployment of AI

This Part performs a deeper dive into the algorithmic governance tools deployed or under development at federal agencies. We aim to describe in detail a handful of especially impactful or promising use cases without losing generality and without discounting the rich diversity of applications across agencies, governance tasks, and policy areas. To that end, we showcase seven algorithmic governance tools spanning seven governance tasks: civil enforcement, hybrid civil / criminal enforcement, formal adjudication, informal adjudication, regulatory analysis, public engagement, and public service provision. This approach, administrative lawyers will recognize, partially tracks categories of agency action within contemporary American administrative law.[1] Further, each chapter hews to a common format, first spotlighting one or more use cases at a single agency and thoroughly exploring their technical and operational details and their future trajectory. Each chapter then closes by framing the agency-specific implications of the tool and the wider family of applications in use at other agencies. We reserve discussion of legal and policy issues that cut across use cases and governance tasks for Part III.

Regulatory Enforcement at the Securities and Exchange Commission

As Part I's inventory revealed, AI has made some of its most substantial inroads in the context of agency enforcement activities. These efforts are especially important, for enforcement is the tip of the spear of the modern regulatory state. It is the primary way government gives real-world effect to legal mandates, thus converting "law on the books" into "law in action."[2] It is also one of the main ways government delivers to the polity a wide range of policy benefits, from clean air and water and safe food, drugs, and workplaces to capital and labor markets that are efficient and fair. And it is the principal means by which the government protects its own interests against those who would abuse it by underpaying taxes or defrauding the government when it buys needed goods or services from the private sector.

KEY TAKEAWAYS

- The Securities and Exchange Commission (SEC) is using a suite of algorithmic tools to identify violators of the securities laws.
- A challenge is training data that accurately reflects ground truth, avoids narrow feedback loops based on prior decisions, and accounts for dynamic changes in wrongdoing.
- Accountability of AI-based enforcement runs into a tradition of enforcement discretion protected under law and the risk of gaming by regulatory targets.
- Agency investigators demand explainable AI-based tools, not just risk predictions, facilitating "internal" due process.

Not only is enforcement central to governance, it also embodies one of the core dilemmas of modern administration: the trade-off between discretion and accountability. On the one hand, agencies vested with enforcement authority need discretion because agency resources are finite, and the costs agencies would incur in identifying all violators of a law and pursuing them to a conclusion are virtually infinite.[3] Moreover, the cost of prosecuting an enforcement action may exceed its social benefit by stifling socially productive activity, or it may simply not be a sound use of scarce agency resources given other policy priorities and imperatives.[4] On the other hand, the exercise of prosecutorial discretion—and an agency's decision not to wield the state's coercive power at all—brings risks. Agency forbearance can yield arbitrary selection of regulatory targets that undermines the legitimacy of the regulatory state by treating similarly situated violators differently. In addition, an agency's decision to forego an enforcement action may mask infidelity to Congress's command or, worse, patterns of political influence and agency "capture" that threaten rule of law.[5] In short, prosecutorial discretion is both a necessary component of good regulatory governance and an ever-present threat to the regulatory state's legitimacy.

In what follows, we focus much of our attention on a suite of enforcement tools in use at the Securities and Exchange Commission that navigate these pressures by helping agency staff "shrink the haystack" of potential violators and better allocate scarce agency resources. While we focus on the SEC, a number of other federal agencies with significant enforcement mandates, among them the Internal Revenue Service (IRS), the Centers for Medicare and Medicaid Services (CMS), and the Environmental Protection Agency (EPA), are also developing or deploying similar tools.

I. THE SECURITIES AND EXCHANGE COMMISSION

The mission of the Securities and Exchange Commission (SEC) is to "protect investors, maintain fair, orderly, and efficient markets, and facilitate capital formation."[6] To achieve these regulatory objectives, the SEC issues rules governing securities exchanges, securities brokers and dealers, investment advisors, and mutual funds.[7] The SEC not only has the authority to issue rules under various of the federal securities laws but can also bring enforcement actions against those who violate them. The SEC brings hundreds such enforcement actions each year. The SEC's wide-ranging regulatory and enforcement duties are reflected in its structure and organization. The Commission is headed by five Presidentially-appointed Commissioners, one of whom serves as chairperson.[8] The Commission is further organized into five divisions[9] and several standalone offices.[10]

II. AI USE CASES

The SEC is currently developing or deploying multiple algorithmic enforcement tools across all five of its divisions and several of its standalone offices. We here profile four such tools. One targets fraud in accounting and financial reporting, two target trading-based market misconduct, particularly insider trading, and a fourth targets unlawful investment advisors and asset managers.

A. Accounting and Financial Reporting Fraud: CIRA

To detect fraud in accounting and financial reporting, the SEC has developed the Corporate Issuer Risk Assessment (CIRA). CIRA is a dashboard of some 200 metrics that are used to detect anomalous patterns in financial reporting of corporate issuers of securities.[11] Today, there are over 7,000 corporate issuers who must submit financial statements, such as annual 10-K and quarterly 10-Q forms, to the SEC for oversight.[12] These reports can be hundreds of pages long, containing general business information, risk factors, financial data, and so-called MD&As (Management's Discussion and Analysis of Financial Condition and Results of Operations).

Analyzing this immense body of reports is a resource-intensive process, and, as with any agency, the SEC has limited resources with which to do it. CIRA's goal is to help the agency more efficiently utilize its finite resources by identifying corporate filers that warrant further investigation. One way SEC staff have sought to manage large data flows is through use of a machine learning tool that helps identify which filers might be engaged in suspect earnings management.[13] The tool is trained on a historical dataset of past issuer filings and uses a random forest model to predict possible misconduct using indicators such as earnings restatements and past enforcement actions.[14] Enforcement staff scrutinize the results, thus maintaining a human eye, and consider them alongside a range of other metrics and materials.[15] Though the algorithmic outputs are only part of a broader analysis, SEC staff report that CIRA's algorithmic component improves the allocation of scarce enforcement resources.[16]

B. Trading-Based Market Misconduct: ARTEMIS and ATLAS

A further pair of tools target trading-based market misconduct: the Advanced Relational Trading Enforcement Metrics Investigation System (ARTEMIS) and the Abnormal Trading and Link Analysis System (ATLAS).[17]

ARTEMIS identifies and assesses suspicious trading by "analyz[ing] patterns and relationships among multiple traders using the Division's electronic database of over six billion electronic equities and options trading records."[18] The tool aims to catch all instances of insider trading in the market and enhances the SEC's monitoring and surveillance powers. ARTEMIS's focus is serial cheaters. This is said to be an easier demographic of offenders to identify compared to first-time insider traders, the target of the ATLAS tool.

The ARTEMIS process is not automated. It first requires the agency to identify a suspected offender before targeted data collection and a full-fledged investigation can proceed. The first step is a preliminary analysis of publicly available information and corporate filings. Companies often announce important events in scheduled 10-K and 10-R filings, but for material events that fall outside these scheduled filings, companies are required to make an announcement using an 8-K form, which is submitted to the SEC's Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system, a public database of corporate filings and voluntary shareholder reports. SEC staff have developed in-house natural language processing (NLP) tools to analyze submitted 8-K forms.[19] A standard NLP process is used to process these forms, stemming the words and then applying a "bag of words" model[20] to classify documents according to the significance of an event, language changes in the disclosures, scheduled and unscheduled earnings announcements, and events not necessarily related to earnings, such as CEO terminations, FDA (dis)approval announcements, court judgments, and clinical trials, among others. The labeled data is then pushed through a supervised learning algorithm to identify trigger events and market changes that may warrant investigation. Once the data has been sifted and analyzed, a human examiner reviews

the results. This is a disciplined process to ensure that the agency is not overburdening brokerage companies or others within the broker/dealer community without a firm basis for doing so.[21]

Once an SEC examiner hypothesizes that there was insider trading on a stock, agency staff prepare a "bluesheet" request to the relevant parts of the broker/dealer community to obtain a comprehensive trading record for the stock or related options within a set time period.[22] Staff decide which securities merit review and the time period for which to obtain trading data.[23] Staff must also identify which broker/dealers traded the security at issue by obtaining the clearing reports submitted to FINRA.[24] To ensure that the data in bluesheets is high-quality, the SEC and FINRA bring charges against brokerage firms for inaccurate or incomplete submissions.[25]

Data obtained via bluesheet requests are sorted and categorized according to the event that triggered elevated review.[26] Next, the SEC uses bluesheet data alongside a database of every previously requested bluesheet to judge whether the trading behavior is anomalous in the context of the trader's identity and historical behavior.[27] ARTEMIS currently uses an unsupervised learning model for anomaly detection. The working assumption is that suspicious activity is an outlier and will not match the patterns of most other data. Because of the difficulty of getting labeled data for this task—it is hard to identify all true positives in past data and impossible to identify the false negatives— the SEC uses an unsupervised learning approach. While accurate ground truth[28] data is difficult to obtain, the models make predictive inferences about fraud by viewing data points in relation to one another.[29]

ATLAS complements the ARTEMIS tool by focusing on first-time, rather than serial, insider trading. It is the newest of the SEC's algorithmic enforcement tools, and there is much less publicly available information about its technical and operational details. Like ARTEMIS, the ATLAS tool uses bluesheet data, from which a half dozen hand-crafted data features are extracted. Such features were defined using domain knowledge about insider training and are described as having an "intuitive explanation."[30] These data features are then fed into a supervised machined learning model (a one-class support vector machine or SVM) to determine if the trade is suspicious.[31] The potential regulatory targets fed into the model are then split into two categories: those who lost money on a trade, and those who made money. The SVM is trained on the former, then fit to the latter. The assumption is that the behavior of those who made money should not differ significantly from those who lost money over time. Outliers are treated as suspicious.

It is important to note that ARTEMIS and ATLAS are only two of many tools and systems SEC staff use to build insider trading cases. As SEC staff emphasized, the process of identifying and investigating insider trading is an iterative process that requires sifting through many sources of data, knowing the context of the situation, and synthesizing evidence and concepts into higher-order judgments.

C. "Registrant" Misconduct: The Form ADV Fraud Predictor

A fourth and final tool, the Form ADV Fraud Predictor, helps SEC staff predict which financial services professionals may be violating federal securities laws.[32] The tool parses so-called Form ADVs—also known as the Uniform Application for Investment Adviser Registration and Report by Exempt Reporting Adviser—a filing that investment advisors who manage more than $25 million in assets must submit to the SEC annually. Form ADVs contain two parts. The first part requires disclosure of the investment advisor's "business, ownership, clients, employees, business practices, affiliations, and any disciplinary events of the adviser or its employees."[33] The second elicits information regarding services offered, fee schedule, as well as an array of information relating to disciplinary information, conflicts of interest, and the educational and business background of the advisor and key supporting management and staff.[34]

Because Form ADVs are composed of free text, NLP algorithms are used to normalize the inputs in order to detect instances of fraud. Because it is difficult to observe fraud directly,[35] the SEC has developed a multi-step process to automate the fraud detection pipeline. After a pre-processing step that algorithmically converts PDF forms into useable blocks of text,[36] an unsupervised NLP technique (Latent Dirichlet allocation or LDA[37]) generates topics that best describe the words in each document.[38] This approach identifies topics in the documents without prior knowledge about what the topics will be.

The final step deploys a supervised learning algorithm to flag current registrants as "high," "medium," and "low" priority for further investigation by SEC staff.[39] The algorithm is trained on a dataset of past registrants that were referred to the agency's enforcement arm. Data features include the topics in the Form ADVs as well as information collected by SEC staff during

interviews and site visits, among other sources. This step relies upon a random forest model to predict document priority, with "high" priority documents referred to relevant SEC staff. This recommendation is accompanied by an explanation of the document's flag, including a rough measure of feature importance.[40]

III. FUTURE TRAJECTORY OF AI AT SEC

The SEC's suite of algorithmic tools provides a glimpse of a potential revolution in regulatory enforcement. Here we highlight some technical challenges and opportunities that are likely to shape the trajectory of algorithmic enforcement tools at the SEC in the near- to mid-term. Among these are *input challenges*, including the need for data that accurately reflects ground truth and takes account of the dynamic nature of wrongdoing, and *analytic challenges*, which largely relate to the need for technical capacity to develop, deploy, and maintain useable tools and exploit continued advances in machine learning. Together, these challenges reveal both the limits of the new algorithmic enforcement and the rich possibilities going forward.

A. Input Challenges: Data, Ground Truth, and the Dynamic Nature of Wrongdoing

Enforcement tools can only be as good as their data inputs. Unlocking the full potential of machine learning in any regulatory context, but especially in the enforcement context, requires abundant, well-labeled data that accurately reflect "ground truth" about misconduct. Data quantity and quality are thus a key determinant of, and a significant limit on, the potential of the SEC's new algorithmic governance toolkit.

Some of the SEC's data challenges afflict any agency developing algorithmic governance tools. Many of the documents the SEC uses to power its algorithmic tools are not in machine readable formats and thus require substantial pre-processing. For this reason, there has been an internal push at the SEC to require filings in both human and machine readable formats.[41] The SEC must also navigate a welter of data laws limiting collection, storage, and use of data. We discuss the effect of data laws on the future of algorithmic governance in Part III's discussion of internal capacity building.

Beyond these more generic hurdles, data challenges in the enforcement context tend to take one of two forms, reflecting either a lack of randomization or the difficulty of finding accurate ground truth in training data. The first of these is exemplified by the bluesheet process that feeds the SEC's ARTEMIS and ATLAS tools. That process, as noted previously, is neither comprehensive nor random. Instead, it is hypothesis-driven and reflects SEC staff judgments about the likelihood of market misconduct in each case. As a result, the types of misconduct and entities targeted will reflect the assumptions, heuristics, and biases of enforcement staff. Furthermore, the ARTEMIS and ATLAS tools are trained on a pool of trading data that includes only past bluesheet requests and thus captures only a small fraction of total trading activity. When either of the SEC's tools looks for patterns suggestive of insider trading, the system compares previously flagged trading behavior to other flagged traders, not traders in the market as a whole, potentially reducing the tool's accuracy.

The second type of input challenge is finding accurate ground truth for training data. This concern pervades algorithmic enforcement tools because it is difficult to identify all true positives in past data, and one can never "know," or comprehensively identify, false negatives. A related challenge comes at the intersection of automation and human-level discretion. When a line-level investigator retains the ultimate authority to initiate an enforcement action, uncritical reliance on automation may displace investigatorial attention away from false negatives and/or crowd out the application of discretion to false positives. If prior enforcement actions are used as training data, the system may unduly confine enforcement actions to a distinct subset of all violations. This phenomenon has been well-documented in the predictive policing context: When a predictive model is used to deploy police, and the resulting arrest data is employed to re-train the model, a "runaway feedback loop" occurs.[42] Police may be sent to the same neighborhoods over and over again regardless of the underlying crime rate. In short, algorithmic detection may be dominated by superficial features from prior enforcement decision-making, replicating the idiosyncrasies of line-level enforcers rather than building richer and more precise models of noncompliance.

A final and fundamental challenge arises from the dynamic nature of wrongdoing. As already noted, effective algorithmic enforcement tools require training data that accurately reflect ground truth about misconduct. But the regulatory landscape, and the ground under it, can shift over time. For many agencies, enforcement is akin to a game of "whack-a-mole" in which regulatory subjects seek to evade regulation by developing new artifices designed to evade, or narrowly navigate between, announced rules. Tax shelters, to cite a concrete example, follow this script. As a result, algorithmic enforcement tools are rarely turnkey systems, and agencies must continually and iteratively update them to capture new

modes of wrongdoing.[43] Returning to tax, an algorithmic tool might be able to flag the complicated and choreographed set of transactions needed to implement an illegal tax shelter. But once enforcement has begun, taxpayers and the tax compliance industry shift away and develop new artifices that are identifiable to algorithmic enforcement tools only if they are sufficiently similar to the prior ones.[44] For agencies using algorithmic enforcement tools, the challenge is designing systematic methods of model optimization and updating built upon randomized case samples (*i.e.,* a sample that includes both cases identified as problematic and unproblematic), as well as careful procedures for incorporating newly discovered types of wrongdoing.[45] If data rooted in historical enforcement patterns are unreflectively used to train models and efforts to update those models are ad hoc, enforcement efforts risk focusing on an arbitrary subset of violations or fighting the last war instead of addressing new forms of misconduct.

The SEC is cognizant of these challenges and is attempting to mitigate them. Improved data systems may overcome the shortcomings of the bluesheet process. In 2016, the SEC approved a joint plan with FINRA and SROs to develop a consolidated audit trail ("CAT").[46] Adopted under SEC rules, 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, thus broadening reporting to every trade quote and order, origination, modification, execution, routing, and cancellation.[47] Once fully implemented, CAT will generate an estimated 58 billion trading records each day.[48] Granting the ARTEMIS and ATLAS systems access to this data stands to substantially improve accuracy and reliability.

In addition, the SEC has begun to pilot a range of evaluation and validation efforts. While ground truth challenges are endemic and make objective performance metrics hard to create, back-testing of the ATLAS tool revealed that its models could predict all or nearly all proven instances of past insider trading.[49] Similarly, agency technologists systematically worked with enforcement staff in one of the Commission's regional enforcement offices to test the frequency with which the Form ADV tool surfaced a problem that was then corroborated upon examination.[50] Neither of these efforts, however, rose to the level of a rigorous "benchmarking," in which agency administrators would set aside a random test sample of enforcement targets and then work up the cases in the old school, analog fashion, comparing the results to those achieved via algorithmic methods.[51] We discuss such an approach in Part III's discussion of transparency and accountability.

If data rooted in historical enforcement patterns are unreflectively used to train models and efforts to update those models are ad hoc, enforcement efforts risk focusing on an arbitrary subset of violations or fighting the last war instead of addressing new forms of misconduct.

B. Analytic Challenges: Technical Demands and Capacity Building

A second type of challenge centers on the substantial technical demands of the new algorithmic enforcement tools. As described throughout this report, developing and maintaining effective algorithmic tools across the full range of governance tasks will require significant computing and data infrastructure as well as technical expertise. Technical capacity will likely grow in importance as algorithmic governance tools become more sophisticated.

The need for technical capacity may be especially strong in the enforcement context. Many of the new algorithmic enforcement tools will, as with the SEC's Form ADV Predictor Tool, rely on NLP techniques to derive semantic meaning from unstructured texts. Enormous strides have been made in NLP in recent years due to advances in deep learning and computing power. But NLP advances have tended to focus on areas with commonly-accepted benchmark tasks (*e.g.,* GLUE, IMDb movie review datasets).[52] Advances have come more slowly in niche contexts involving more specialized, technical, and jargon-filled text with no large gold standard, labelled datasets. As a result, developing workable algorithmic governance tools may require more than off-the-shelf and third-party implementations or open-source libraries. Words used in the finance context may have different meanings than in standard parlance, requiring finance-specific corpora in order to maximize the tool's utility.[53] If a financial regulator like the SEC were to use standard datasets rather than finance-

specific ones, or if existing algorithms are not retrofitted to promote transfer between domains, the resulting system could be less effective.

Another key issue is whether agencies will meet heightened technical demands by developing internal capacity, thus choosing to "make" their own algorithmic tools, or instead "buy" needed technical capacity by acquiring new tools via the procurement process.[54] This make-or-buy choice, which we discuss in more detail in Part III, may be particularly salient in the enforcement context. The dynamic nature of wrongdoing and the subtlety and complexity of many enforcement tasks mean that the design, deployment, and maintenance of algorithmic enforcement tools may be best achieved with substantial embedded agency expertise—that is, technologists sited within the agency who understand subtle and complex governance tasks—rather than contractors working at a remove.[55] Finally, the make-or-buy choice presents an especially acute challenge in the enforcement context because of the threat of gaming and adversarial learning by regulated parties, an issue we also take up in more detail in Part III.

C. New Technical Vistas

No matter how the SEC goes about improving data inputs and acquiring needed technical capacity, doing so will allow it to pursue a number of new and promising technical vistas.

First, it is plausible—and perhaps even likely—that continued, non-trivial technical advances in the coming years will move algorithmic enforcement tools steadily closer to fully automated decision-making. This does not describe the SEC's current menu of algorithmic enforcement tools. Most of these tools use a classifier, the results of which are handed off to line-level enforcement staff who continue to work up cases themselves. Discretion to initiate enforcement action remains in human hands.[56] But this may change. Continued technological advances may eventually cause much of enforcement decision-making, from monitoring to initiation of enforcement actions to agency adjudication, to be an entirely machine-driven process. As we explain in more detail in Part III, full automation of consequential governance decisions will raise significant legal issues that courts reviewing agency action will have to resolve.

A second clear vista is agency development of ever more sophisticated analytic techniques. One likely growth area is computer vision, the branch of AI that trains computers to understand the visual world. Already, academic researchers have developed machine learning tools that analyze satellite imagery to predict which facilities are concentrated animal feeding operations and thus at greater risk of violating the Clean Water Act or other environmental laws.[57] Another clear technical frontier is use of unsupervised learning techniques to improve compliance monitoring and enforcement capabilities in the absence of data with high-quality labels.[58]

A final vista is the use of machine learning techniques that do not enhance an agency's own technical capacity, but rather allow it to harness outside talents and enterprise. Algorithms might be developed, for instance, to generate synthetic datasets that preserve the higher-order dimensions of real data without disclosing personally identifiable information or other sensitive information.[59] This data generation would be an improvement from the public use files that are currently available. By publicly releasing data of this sort and inviting collaboration, enforcement agencies like the SEC could harness private expertise and enterprise and thus obtain analytical support that they cannot currently access—a potentially happy story of technical innovation begetting technical innovation.

IV. IMPLICATIONS: THE FUTURE OF ALGORITHMIC ENFORCEMENT

The SEC is hardly alone in leveraging AI to perform enforcement-related tasks. The Internal Revenue Service (IRS) and the Centers for Medicare and Medicaid Services (CMS) have deployed algorithmic tools designed to predict illegal conduct and more precisely allocate scarce agency resources toward audit or investigation. The IRS, for instance, has responded to budget and workforce cuts by investing over $400 million to develop and operate a fraud detection algorithm, the Return Review Program (RRP), that generates fraud risk scores for all national individual tax returns claiming a refund.[60] For its part, CMS has engaged contractors to help build and implement a machine-learning-based risk assessment tool that analyzes historical and incoming claims to furnish provider-level leads to the agency's fraud investigators. And a wide range of other agencies, from the Environmental Protection Agency and Department of Labor to the Consumer Product Safety Commission, are developing or deploying tools that predict non-compliance with rules regarding the environment, workplace safety, banking, consumer product safety, food processing, disability insurance, and workers' compensation, among others. This steadily growing catalog of algorithmic enforcement tools holds significant implications for the future of regulatory governance.

A. Enforcement and Algorithmic Accountability

The proliferation of algorithmic enforcement tools at the SEC and beyond highlights especially difficult trade-offs between the efficacy of the new tools and the accountability concerns that animate administrative law. As Part III describes in more detail, algorithmic governance tools trigger a profound collision between administrative law's requirement of transparency and reason-giving and the fact that many algorithmic decision tools are not, by their structure, fully explainable. An important debate asks how much transparency, from thin system-level explanations of a tool to full disclosure of a tool's source code and data, is necessary to gauge a tool's fidelity to governing law. Some advocate deliberate impairment of an AI tool's predictive accuracy to achieve explainability.

Algorithmic governance tools trigger a profound collision between administrative law's requirement of transparency and reason-giving and the fact that many algorithmic decision tools are not, by their structure, fully explainable.

A critical question is whether continued uptake of algorithmic tools by enforcement agencies will, on net, render enforcement decisions more or less accountable. On the one hand, the black box nature of machine learning tools may exacerbate accountability concerns. On the other hand, algorithmic enforcement tools can, by formalizing and making explicit agency priorities, render an agency's enforcement decision-making more tractable compared to the dispersed human judgments of agency enforcement staff. Algorithmic enforcement tools might thus provide a "focal point" for judicial review, undermining the normative foundation of longstanding legal doctrines, embodied by the Supreme Court's *Heckler v. Chaney* decision, hiving off agency enforcement decision-making from judicial review.[61] Algorithmic enforcement tools, by encoding legal principles and agency policies and priorities, might also qualify as "legislative rules" under the Administrative Procedure Act and thus require full ventilation via notice and comment. The result, though it runs contrary to much contemporary commentary, is that displacement of agency enforcement discretion by algorithmic tools may, on net, produce an enforcement apparatus that is *more* transparent, whether to reviewing courts or to the agency officials who must supervise enforcement staff.

But legal demands of transparency also produce further trade-offs in the enforcement context because of the risk that public disclosure of a tool's details will expose it to gaming and "adversarial learning" by regulated parties.[62] As discussed in more detail in Part III, an SEC registrant with knowledge of the workings of the SEC's Form ADV Fraud Predictor could adversarially craft its disclosures, including or omitting key language in order to foil the system's classifier. A key line of inquiry in the enforcement area will be what degree of transparency, and what set of oversight and regulatory mechanisms, can reach a sensible accommodation of interlocking concerns about efficacy, accountability, and gaming.

B. Algorithmic Enforcement and the Structure and Legitimacy of the Regulatory State

Algorithmic enforcement tools may also, in time, work a fundamental change in the structure and legitimacy of the administrative state. Algorithmic enforcement tools are force-multipliers that allow an agency to do more with less by permitting agencies to identify regulatory targets more efficiently. In this sense, the advent of algorithmic enforcement tools could halt or even reverse the decades-long shift away from public enforcement and toward private litigation as a regulatory mode.[63]

The advent of algorithmic enforcement may also supplant expertise within the federal bureaucracy, exacerbating a perceived trend toward politicized federal administration and the hollowing out of the administrative state.[64] This is especially worrying because, at least for the moment, line-level enforcers appear to play a key role in bolstering the accountability of new algorithmic tools. Because SEC enforcement staff can choose whether to use algorithmic enforcement tools, agency technologists must sell skeptical line-level staff on their value. 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 part of the disclosures triggered the classification and why. This is pressing agency technologists to focus on explainability in building their models by taking account of

frontier research on how to isolate which data features in an AI system may be driving an algorithmic output. Staff skepticism and demand for explainable outputs raise the possibility that governance of public sector algorithmic tools will at times come from "internal" due process, not the judge-enforced, external variety.[65]

Finally, as algorithmic tools move closer to the core of the state's coercive power, they may systematically shift patterns of state action in ways that raise distributive and, ultimately, political anxieties about a newly digitized public sector. As already noted, gaming reduces the efficacy of algorithmic systems and risks rendering their outputs fully arbitrary. But gaming is also likely to have a distributive cast, particularly in the enforcement context.[66] The predictions of the SEC's Form ADV Fraud Predictor as to which investment brokers are likely to be the bad apples may fall more heavily on smaller investment firms that, unlike Goldman Sachs, lack a stable of computer scientists who can reverse-engineer the SEC's system and work to keep their personnel out of the agency's cross-hairs.[67] As we explore in more detail in Part III, a narrow focus on technical and capacity-building challenges misses the profound political implications of the current algorithmic moment.

* * * *

As the SEC's experience illustrates, AI/ML tools have the potential to help enforcement agencies flag potential violations of the law and focus agency attention in a world of scarce resources. This improved accuracy and efficiency may come at a cost, however. As AI/ML tools get ever more sophisticated, they also pose real threats to the transparency and democratic accountability of enforcement agencies and of the regulatory state as we know it.

Law Enforcement at Customs and Border Protection

While civil enforcement agencies like the SEC, IRS, and EPA have begun to experiment with machine learning, pure law enforcement agencies have been quicker to adopt such tools. Nearly 100 state and local jurisdictions have replaced traditional surveillance cameras with more sophisticated AI-powered gunshot detection technology.[1] Others have employed AI-driven automatic license plate readers.[2] Police departments in Los Angeles,[3] Chicago,[4] New Orleans,[5] and Missouri[6] have deployed AI-powered predictive policing strategies to identify gang-related crimes. Federal law enforcement agencies such as the Federal Bureau of Investigations (FBI) use similar strategies, though we do not consider them here due to the information barriers inherent in examining a pure criminal law enforcement agency like the FBI. Instead, we turn in this chapter to an agency that straddles the civil and criminal divide: Customs and Border Protection (CBP).

KEY TAKEAWAYS

- Customs and Border Protection (CBP) has relied extensively on contractors to develop facial recognition technology and risk scoring of passengers.
- Reliance on contractors carries a cost: CBP was unable to explain failure rates of one biometric scanning application.
- While such tools can expedite processing at airports and borders, they also raise questions about interagency data sharing, privacy, equal protection, and rulemaking requirements.

Although CBP conducts some civilian enforcement functions such as cargo inspection, the agency also monitors criminal activity and enforces immigration and customs laws. In doing so, CBP deploys two of the most controversial AI/ML tools: facial recognition and risk prediction.

Law enforcement agencies have used facial recognition technology to match still images from crime scenes against criminal databases since at least 2009.[7] As of 2016, one in four state or local police departments had access to facial recognition databases, and sixteen states had contributed driver's license photos to the FBI's federal equivalent.[8] Today, local governments have access to an even wider suite of live facial recognition surveillance tools.[9] Despite its growing popularity with law enforcement, facial recognition raises unique privacy and due process concerns. Four cities in California and Massachusetts have banned its use,[10] and a recent Senate bill proposes to ban facial recognition technology in all public housing receiving federal funding.[11]

Risk prediction, too, is a widely contested use of AI/ML. Law enforcement risk prediction tools incorporate data about individuals—such as their age, criminal history, and gender—to identify individuals that may be at risk of committing or becoming the victims of crime. Police departments can use such tools to rank individuals prone to violence based on their likelihood of becoming involved in a future homicide.[12] Courts routinely consult AI-based risk-assessment tools, referring to automatically-generated risk scorecards in making sentencing decisions.[13] Like facial recognition, law enforcement uses of risk prediction may raise due process[14] or equal protection concerns.

In this chapter, we consider CBP's plans to "transform the way it identifies travelers by shifting the key to unlocking a traveler's record from biographic identifiers"—such as passports and visas—"to biometric ones—primarily a traveler's face."[15]

This significant shift may allow the agency to leverage a broader dataset and improve accuracy, but it also creates new vulnerabilities and raises important legal questions. We begin with a brief overview of the agency and its mandate, then examine CBP's use of facial recognition and risk prediction. We conclude by discussing the trajectory for AI use at CBP and highlight implications for the future of such technologies.

I. CUSTOMS AND BORDER PROTECTION

Customs and Border Protection sits within the Department of Homeland Security, the federal government's largest law enforcement agency.[16] The agency was established in 2003, in the aftermath of the September 11 attacks, to combine the former U.S. Customs Service with various immigration, agricultural inspection, and border patrol roles.[17] Today, CBP boasts 60,000 employees, making it "one of the world's largest law enforcement organizations."[18]

These employees are spread throughout several sub-agencies, including the Office of Field Operations—which controls ports of entry—and the U.S. Border Patrol, which patrols the areas between them.[19] The Air and Marine Operations division supports these functions,[20] and the Office of Trade coordinates with international partners and other agencies to enforce trade restrictions and customs laws at ports of entry.[21] CBP is a pure law enforcement agency, imposing both civil and criminal penalties in its enforcement of twenty-nine distinct sections of the U.S. Code, in areas ranging from intellectual property protection to environmental conservation.[22] Although CBP fulfills a wide variety of functions, the agency's overarching mission is "[t]o safeguard America's borders."[23] To achieve this mission, the agency continues to rely on specialized officer teams, such as the CBP Aircraft Search Team, to search for contraband and apprehend those engaged in trafficking or smuggling on the ground.[24] It also supplements these human teams with more advanced technological tools. For example, CBP began deploying unmanned drones as early as 2004 to monitor smuggling, human trafficking, and illegal crossings at the border.[25] The agency's use of facial recognition and risk prediction tools are consistent with this trend and are meant to support the work of its human officers.

II. AI USE CASES

Customs and Border Protection has invested significant resources in artificial intelligence. In 2018, the agency received $196 million to acquire and deploy border security technology.[26] In 2017, Executive Order 13,780 enabled the agency to "expedite the completion and implementation of a Biometric Entry-Exit Tracking System."[27] CBP's Facial Recognition and Risk Prediction Programs have grown out of the agency's emphasis on counterterrorism: The Facial Recognition Program is a "direct result" of "recommendations from the 9/11 Commission,"[28] and the Risk Prediction Program bills itself as a way to "identify potential threats."[29] The latter assesses not only human travelers, but also non-human entities such as cargo.

A. Facial Recognition

In 2004, Congress authorized CBP to collect biometric data from non-citizens entering the United States.[30] At entry points such as airports, officers would match passenger data provided by airlines with a passenger's passport. Officers would also collect the biometric fingerprints of non-citizens to ensure a match with the passports, and would run that data against criminal and terrorist watchlists.[31] In 2012, the agency began placing self-service Global Entry kiosks at airports to process biographic data for inbound U.S. citizens.[32] This enabled the agency to focus human resources on processing non-citizens. Three years later, the agency began piloting facial recognition and mobile fingerprint scanners at airports to enhance this process.[33]

Facial recognition systems match a given image or video of a face with an identity.[34] These systems work by first detecting a face—often in cluttered, crowded or obstructed settings—then matching its features against a database of known faces. A CBP port of entry kiosk, for example, may snap a photo of incoming persons, detect and crop a particular face, and create a vector representation that can more robustly match an identity in the underlying database. Biometric identification, which identifies people using physical traits such as the face, fingerprints, or voice, has several advantages.[35] In theory, each person's physical traits are unique and more difficult to falsify than written documents like passports.[36] Although CBP initially experimented with other biometric technologies such as iris and mobile fingerprint scanning,[37] it has settled upon facial recognition both because of low iris capture rates,[38] and because photos of faces are more widely available and easily accessible than iris data—which the agency would have to gather on its own.[39] The agency recognizes that "[w]hile DHS' gallery of fingerprints is large, it pales in comparison to the number of facial recognition quality photos held in existing data sources, such as the U.S. Passport and Visa databases."[40] Facial recognition also served as the least disruptive option: The

agency "found that facial recognition was intuitive for people. Everybody knows how to stand in front of a camera and have his or her picture taken."[41]

CBP launched its facial recognition program in 2017.[42] It typically obtains facial recognition software from private vendors that closely guard their intellectual property, making it difficult to determine the architecture of its program.[43] Unisys, an information technology company that is the agency's main contractor for facial recognition, seems to use deep learning in both the pre- and post-image processing stages to extract features—such as estimated age—from an image. The agency appears to have combined products from companies that likely employ different models, such as NEC's NeoFace[44]—which identifies individuals using proprietary feature-extraction methods and deep learning face-matching—and Cognitec's FaceVACs.[45] The deployment examples below illustrate the range of this technology and its applications to CBP's broader mission.

1. Airports

As part of its Biometric Entry/Exit Program, CBP entered into agreements requiring airports to buy cameras to capture facial images.[46] Several airports and airlines have rolled out the agency's Traveler Verification Service, a central biometric identification system that includes facial recognition.[47] Private partners use the Traveler Verification Service for paperless boarding and processing.[48] In June 2018, Orlando International Airport became the first to use facial recognition on all travelers.[49] At least seventeen airports, from Seattle to Atlanta and Fort Lauderdale, have followed suit.[50] As passengers board an aircraft, the system captures their photos, which an algorithm then processes to ensure that the boarding passengers' faces match their registered photos. CBP claims that its facial recognition system finds passengers attempting to enter the country illegally on a daily basis.[51] CBP initially failed to clarify whether passengers could opt out of the system, generating confusion and criticism.[52] And despite CBP's later clarification that passengers could opt out, the rate at which passengers do so remains low: In one audit, only sixteen passengers across a sample of twelve flights declined to be photographed.[53]

CBP is partnering with the Transportation Security Administration to test its ability to capture and match facial images at checkpoints. But in addition to sharing its system with sister agencies, CBP also encourages third-party use of these technologies. CBP describes its database as a "device agnostic backend backbone" that will spur "private sector investment in front end infrastructure, such as self-service baggage drop off kiosks, facial recognition self-boarding gates, and other equipment."[54] As the agency states:

> Customs and Border Protection will build a backend communication portal to support TSA, airport, and airline partners in their efforts to use facial images as a single biometric key for identifying and matching travelers to their identities. This portal will provide our partners with the ability to utilize the Customs and Border Protection Biometric Pathway for identity verification at any point in the travel continuum.[55]

Commercial carriers and airport authorities send the facial images they capture to CBP's Traveler Verification Service.[56] The agency retains photos of U.S. citizens only until their identities are confirmed, but can retain photos of non-U.S. citizens for up to fifteen years.[57]

2. Land Borders

CBP has also begun testing facial and iris recognition technology at land borders. At the Otay Mesa Port of Entry, for example, the agency has tested a variety of physical deployments including kiosks and stop-and-look or on-the-move detection systems.[58] Kiosks pose the fewest technical challenges, because they scan passengers who are directed to stand in a particular pose. Stop-and-look stations are similar, but must account for different poses. On-the-move placements must process video, which may require detecting faces in a series of frames without overcounting or accounting for blur and other video artifacts. In August 2018, CBP began testing facial recognition on drivers in Anzalduas, Texas.[59] The agency has tested this technology using dedicated lanes, allowing people to opt in to the facial recognition system.[60]

B. Risk Prediction

Risk prediction modeling uses statistical methods to assess particularized future risks, such as the probability that an individual will develop a specific illness. To assess the potential security risks posed by each of the hundreds of millions of people, vehicles, and containers crossing U.S. borders, CBP employs a bundle of programs termed

the Automated Targeting System. Information about the Automated Targeting System first became publicly available in November 2006 through the Federal Register,[61] but has since been exempted from future disclosures.[62] According to agency documents, the system generates and assigns a rating to every entity that crosses U.S. borders, determining the potential threat a given entity poses and the level and priority of screening it should receive.[63] One subsystem screens passengers—including by marking airline passengers for additional screening—while another screens cargo.

By 2014, CBP had begun developing a Risk Prediction Program, which aims to improve the Automated Targeting System in collaboration with the Science & Technology Directorate of the Department of Homeland Security.[64] The Risk Prediction Program will augment existing systems with machine learning techniques to help track and assess more complex patterns.[65] Various contractors with machine learning expertise have been involved in the project. For example, the project integrates Metron's machine learning "real-time Automated Modeling & Targeting Subsystem" into the CBP passenger screening subsystem. Metron has also developed and integrated a "Link Discovery Tool" into passenger screening, aiding the "discovery of links from current passengers to past known or suspected threats."[66] In June 2018, CBP contracted with Unisys,[67] which has developed its own risk-prediction program for use by governments around the world.[68]

The program also uses information from a wide variety of government and non-government data sources. Many of these data sources are internal to DHS, such as the Advance Passenger Information System, Nonimmigrant Information System, and Enforcement Integrated Database. Others, such as the Federal Bureau of Investigation Terrorist Screening Database and the Department of Justice National Crime Information Center, come from peer agencies. But the program also reaches beyond federal government sources to collect passenger data—such as billing and email addresses, payment information, in-flight seat locations, meal preferences, and baggage data—directly from airline carriers.[69]

As in the facial recognition context, CBP has not publicly released the specific technologies its Risk Prediction Program uses to process this data. CBP has both a security and liability incentive to keep its technologies opaque: Doing so wards off adversarial learning while insulating the agency from legal challenges. But the program likely integrates a wide range of tools from contractors and parallel agency initiatives. Contractor Metron, for instance, boasts capabilities that include supervised and unsupervised machine learning and graph-mining.[70] And DHS has reported that CBP's advanced machine learning algorithms in the cargo screening context may support "research on entities and individuals of interest" for intelligence and law enforcement purposes.[71] DHS reports that CBP's passenger screening program employs only "risk-based rules that are derived from discrete data elements" and does not mention machine learning.[72] Yet DHS also contracted with Metron in 2012 to develop an "automated behavior-based screening and anomaly detection technology."[73] And in 2018, it awarded funds to DataRobot, Inc. to "begin testing a prototype of a machine learning platform" for CBP's Global Travel Assessment System.[74]

In one internal report CBP admitted that it was unable to explain the failure rates of its iris scanning technology: "Due to the proprietary technology being used, the specific cause of the failure could not be differentiated."

The Risk Prediction Program's methodology and scope remains somewhat unclear. CBP has denied using ML-powered risk assessment scores for passengers[75] and claims to instead compare each passenger's personally identifiable information "against lookouts and patterns of suspicious activity identified through past investigations and intelligence."[76] In 2009, a lawsuit by the Electronic Frontier Foundation revealed several references to "risk-scored" passenger information within the agency's internal documents.[77] Whether these risk scores come from an ML algorithm similar to those the agency is using in the cargo context or from an expert-based system remains unclear. As recently as 2018, Unisys stated in a press release that it helps CBP assess "risk associated with travelers or cargo

shipments."[78] These risk scores may be used not just by Customs and Border Protection, but also by other DHS or law enforcement agencies.

III. FUTURE TRAJECTORY FOR CBP

Although facial recognition may not be as accurate as fingerprints[79] and may be susceptible to evasion,[80] it offers a more seamless interface than even mobile fingerprint scanners. By deploying facial recognition cameras, CBP can collect biometric facial recognition data on a wider scale, and at faster speeds, than individual fingerprints. The technology also enables CBP to match travelers against a wider database of photos, while reducing the agency's "dependency on less reliable paper travel documents like passports and visas."[81] Risk prediction programs promise similar improvements in efficiency and accuracy. However, these tools remain in their early stages. As CBP expands its use, it will likely have to build greater internal agency AI/ML expertise to address the knowledge gap created by reliance on contractors. In one internal report, for example, CBP admitted that it was unable to explain the failure rates of its iris scanning technology: "[D]ue to the proprietary technology being used, the specific cause of the failure could not be differentiated."[82] If CBP fails to understand the flaws in its own technology, it can expose itself to unknown vulnerabilities and fail to detect adversarial attacks. More broadly, agencies that lack access to a contractor's proprietary technology may be unable to troubleshoot and adapt their own systems.

In addition to building internal technical capacity, CBP may have to stay abreast of advances in AI, as its programs raise security and adversarial learning concerns that law enforcement agencies continue to grapple with. Facial recognition systems can be vulnerable to adversarial attacks.[83] An attacker could, for example, craft data inputs to trick a CBP facial recognition system into incorrectly matching a face with the no-fly list or vice-versa.[84] The simplest attacks can often be the most successful: Security researchers have confused a Tesla autopilot system by strategically placing stickers in its field of view, causing the car to unexpectedly change lanes.[85] Moreover, attackers can often break adversarial defenses in short time frames.[86] Although researchers have sought to address adversarial attacks by training models with theoretical guarantees on robustness,[87] these guarantees often come at the price of accuracy. Furthermore, private contractors may exacerbate vulnerabilities. For example, many of CBP's private contractors list their research on their resumes, enabling adversaries to make educated guesses about the agency's internal models. As CBP expands its use of facial recognition and risk prediction, it must consider and address these evolving security threats.

CBP's controversial gathering and sharing of personal data highlights the importance of establishing robust consent and security procedures.

IV. IMPLICATIONS: THE CHALLENGES AHEAD FOR LAW ENFORCEMENT AGENCIES USING AI

AI/ML tools can enhance a law enforcement agency's breadth and precision. CBP's AI/ML use cases have implications for other agencies in law enforcement and beyond. The Department of Veterans Affairs and other service agencies, for example, may consider using facial recognition or other biometric identifiers to process identities, securely process private information, and reduce wait times.[88] Regulatory enforcement agencies that monitor risk on an individual level—such as the SEC and FTC—may learn from CBP's Risk Prediction Program. These insights could, for example, assist the Centers for Medicare and Medicaid Services (CMS) with its own Fraud Prevention Program.[89] At the same time, agencies will have to consider the challenges that CBP continues to face in deploying its ambitious programs.

First, CBP's controversial gathering and sharing of personal data highlights the importance of establishing robust consent and security procedures, even for interagency data sharing. Federal regulation of interagency data sharing is limited primarily to reporting requirements, and may not fully address public concerns with more modern data sharing. Prior technologies used straightforward opt-in and opt-out mechanisms: Law enforcement agencies historically collected images directly from citizens, such as during the passport, visa application, or border crossing process.[90] And while airports have long used automated passport kiosks which verify identity using facial biometrics, those technologies have been relatively easier to notice and opt out of.[91] By contrast, a U.S.

citizen who wishes to opt out of facial recognition today may find it increasingly difficult to do so: The agency has likely obtained her image from one of its many data sources and may already be using it to verify her identity. In implementing the Facial Recognition Program, for example, CBP obtained decades of passport photos from the Department of State's Consular Consolidated Database,[92] without obtaining consent from U.S. citizens.[93]

Second, data sharing raises privacy concerns. CBP shares information with various federal, state, and local authorities. It partners with the National Institute of Standards and Technology to test and evaluate vendor technologies,[94] and feeds biographic and biometric data to the shared interagency Arrival and Departure Information System.[95] This data sharing extends beyond the government. Under existing guidelines, private parties, such as airlines or airports, might be allowed to use facial images for commercial purposes. One CBP memorandum of understanding does not "preclude any Party from entering into future business agreements or commercial opportunities."[96] According to one media report, this agreement sets "no commercial limits for 'airline partners;' if they wanted to sell or somehow monetize the biometric data they collect, there was nothing stopping them."[97] Moreover, while CBP writes data usage policies into its government contracts,[98] in compliance with DHS guidelines implementing federal requirements from the Security Modernization Act of 2014 and the Cybersecurity Information Sharing Act of 2015,[99] these precautions fail to sufficiently delineate the limits on private use of agency data. To prevent the kind of criticism CBP has faced for its lack of transparency,[100] agencies may want to share clear guidelines on data gathering and data sharing at the outset.

Third, agency uses of facial recognition or risk prediction can raise equal protection concerns. As privacy advocates and academics have highlighted,[101] facial recognition systems can entrench bias into agency decision-making. Biases in the data may translate to inaccuracies for underrepresented populations, implicating disparate impact and disparate treatment concerns.[102] Many systems originally misclassified faces with darker skin tones at higher rates than faces with lighter skin tones.[103] In China, an automated policing system detected a businesswoman's face on a bus advertisement and mistakenly cited her for jaywalking.[104] Although scholars are working to develop algorithms that naturally account for these biases,[105] and vendors have significantly reduced the disparity in error rates,[106] systems continue to be far from perfect. Similarly, with respect to risk prediction, risk scores may rely on protected characteristics such race, religion, and gender.[107] The legal issues surrounding such scoring remain controversial but unresolved in the algorithmic context.[108] Agencies should consider these legal concerns when determining whether to adopt facial recognition systems.

Fourth, agency deployment of facial recognition raises Fourth Amendment issues. While current law is permissive with respect to search and seizure at the border,[109] there are some uncertainties around the use of these technologies, as well as congressional calls for greater oversight.[110] There is no judicial ruling on whether using facial recognition constitutes a search under the Fourth Amendment.[111] The use of risk scoring could implicate the Fourth Amendment if it becomes the basis for—or even merely prolongs—a search or seizure. The border exception is less likely to apply in other agency contexts deploying risk scores and facial recognition.

Fifth, agencies that implement AI/ML technologies outside of the national security context may be subject to procedural requirements under the APA. Congress has on several occasions ordered collection of biometrics from foreign nationals at the border, but it has never clearly authorized facial recognition for U.S. citizens.[112] CBP did not conduct notice-and-comment rulemaking prior to implementing its facial recognition program,[113] presumably invoking one of the APA exceptions. In 2011, however, the D.C. Circuit held that the Transportation Security Agency was required to go through the rulemaking process prior to employing body scanners at airports.[114] The D.C. Circuit found the agency's adoption of body scanners to be a legislative rule, requiring notice and comment. If the biometric program is analogous to the use of body scanners, it may similarly require rulemaking. The APA further exempts agencies from notice and comment when impractical or unnecessary[115] or for matters of military or foreign affairs.[116] The agency might claim this exemption, invoking Executive Order 13,780, which articulates the national security purpose and urgency driving the biometric Entry-Exit system.[117] But even then, it would likely need to incorporate a brief statement citing its rationale for excluding public comment into the rule itself.[118]

Finally, CBP must contend with serious political and public relations risks. Facial recognition and risk prediction remain some of the most contentious applications of AI/ML technology, in part because they implicate the aforementioned concerns regarding citizen consent, data privacy, equal protection, and searches and seizures. But growing agitation about facial recognition technology and risk prediction runs deeper than that, reflecting more existential concerns about the rise of the surveillance state. While the foregoing cannot possibly do justice to ongoing debate about the proper role of technology-enabled surveillance, it is a crucial debate to have.

* * * *

In sum, AI/ML tools may, as in CBP's case, significantly expand an agency's scope and reach and enable it to make agency operations more efficient and accurate. At the same time, such programs raise privacy and security risks and reveal basic tensions between the goals of law enforcement and agency transparency.

Formal Adjudication at the Social Security Administration

Federal agencies adjudicate more cases than all federal courts combined. Guaranteeing due process in the face of such high volumes remains one of the core challenges of administrative adjudication, with volumes of legal and scholarly analysis devoted to the topic.[1] When an immigrant appeals her asylum decision, when a veteran appeals the denial of disability benefits, or when an individual challenges the denial of Medicare coverage, how can we assure that decisions reached by adjudicators are accurate and consistent? And what potential does AI have to solve what some have called a "looming crisis in decisional quality"[2] in agency adjudication? In this chapter we cover the emerging uses of AI for formal adjudication.

KEY TAKEAWAYS

- AI tools are being developed to improve the accuracy and efficiency of formal adjudication.
- The most advanced tool relies on text parsing of draft decisions to flag potential errors.
- Entrepreneurs within the agency with both subject matter and technical expertise were critical in spurring early innovation to circumvent legacy systems and restrictions.
- As adjudicatory decision tools grow, they may raise questions about decisional authority and independence and the transparency of adjudicative decisions.

AI-based adjudication tools address two types of adjudication settings. First, the textbook APA category of formal adjudication centers on instances where an agency's enabling act requires adjudications to be made on the record.[3] Sometimes dubbed "Type A" adjudications,[4] such adjudications trigger procedural protections under the APA,[5] including the right for interested parties to submit evidence, a right to an exclusive record for the decision, the right to a presiding employee who is at arm's length from agency investigators or prosecutors (separation of functions), and the prohibition of ex parte communications.[6] Typically, the presiding employee is an administrative law judge (ALJ).[7] As of 2017, 27 agencies employed 1,931 ALJs. A small number of agencies employ the vast bulk of ALJs: 86% sit in the Social Security Administration, 5% in the Office of Medicare Hearings and Appeals, 2% in the Department of Labor, and 2% in the National Labor Relations Board.

Second, the tools discussed in this chapter also pertain to adjudications that require evidentiary hearings, but do not trigger the APA's default adjudicatory procedures. In such "Type B" adjudications, agencies employ administrative judges (AJs), whose adjudicatory work resembles that of ALJs, but without the same set of procedural protections as the APA.[8] Strictly speaking, such adjudications are "informal" under the APA, but the enabling act (and, potentially, procedural due process) can still trigger an administrative approximation of a civil trial. Asimow reports nearly 100 schemes where an office's primary function is "Type B" adjudication,[9] but, as with Type A adjudications, a small number of agencies, such as the Board of Veterans Appeals[10] and the Executive Office of Immigration Review,[11] employ the bulk of AJs. These types of adjudications typically also have an exclusive record requirement, the right to a neutral decision maker, separation of functions, and a prohibition on ex parte communications. While "wildly

diverse,"[12] these adjudicatory processes often also face case processing challenges due to significant caseloads. Tools developed at SSA to reduce error and improve processing times of disability adjudication may well translate, for instance, to the closely analogous BVA setting.

The tools we discuss in this chapter have somewhat less direct applicability to the vast residual category of informal adjudications that do not require an evidentiary hearing ("Type C" adjudications). We turn to that setting in the following chapter.

In what follows, we devote the bulk of the discussion to AI innovations in place at the Social Security Administration (SSA), both because the agency plays a central role when it comes to formal adjudication and because SSA has taken the largest strides in developing such tools. We close the chapter by spelling out implications for adjudication more broadly.

I. THE SOCIAL SECURITY ADMINISTRATION

A. Disability Claims System

The SSA was created in 1935 as part of the Social Security Act. Congress subsequently modified the Social Security Act of 1935 to cover people unable to work because of disabilities and created the Social Security and Disability Insurance ("SSDI") Program. The program currently covers all workers between 18 and 65 who have contributed to the program commensurate with their contributions.[13] In 1972, Congress created the Supplemental Security Income ("SSI") Program, which provides disability benefits on the basis of need regardless of past contributions.[14]

The Social Security Act defines "disability" as "inability to perform any substantial gainful activity by reason of a medically determinable impairment that is expected to last at least twelve months or result in death."[15] The statute and implementing regulations create a five-step process for determining whether an applicant meets the criteria and is entitled to benefits.[16] At step one, the agency denies the claim if the claimant is engaged in substantial gainful activity.[17] At step two, the agency considers whether the claimant has "a severe medically determinable physical or mental impairment" (or combination of impairments) of sufficient duration.[18] At step three, the agency determines whether the impairment(s) meets or equals a listed impairment,[19] in which case the claimant is entitled to benefits. If not, the agency determines at step four if the claimant is capable of performing her past relevant work, based on the claimant's "residual functional capacity" and past relevant work.[20] If not, the agency determines at step five whether the claimant is able to perform any other work, based on the claimant's residual functional capacity, age, education, and work experience. If not, the claimant is found disabled and entitled to benefits.[21]

The adjudication process is divided into four administrative levels. First, at the initial state level, the SSA receives an application and sends that application to the State Disability Determination Service ("DDS"). DDS staff work with medical or psychological consultants to initially determine whether the applicant is disabled under the rules. Second, a claimant dissatisfied with DDS's initial determination can request reconsideration. A different DDS examiner reviews all evidence submitted as part of the initial determination, and the claimant can submit further evidence. Third, a claimant dissatisfied with reconsideration results can request a hearing with an Administrative Law Judge ("ALJ") for de novo review of her claim, including review of additional evidence not available at the time of prior proceedings. The Appeals Council handles the final level of appeals within the agency. The Appeals Council is also in charge of performing quality review and policy interpretation.

B. Current Challenges

1. Caseload, Processing Time, and Differential Grant Rates

The SSA is likely "the largest adjudication agency in the western world."[22] Its Office of Disability Adjudication and Review ("ODAR") is in charge of scheduling disability hearings, handling the Appeals Council, and reviewing decisions made by ALJs. ODAR consists of over 160 offices and 10 regional offices. In 2016, SSA received more than 2.5 million disability claims, with almost 700,000 appealed to the hearings level.[23] Because of the caseload, ODAR offices experience a significant backlog of claims. Wait time for a hearing can range from a few months to more than two years.[24] Past efforts to reduce the hearings backlog and the average processing time for claims have had limited success.[25]

In addition, there are longstanding concerns about widely differential grant rates between ALJs, with some judges granting benefits over 90% of the time and others under 10% of the time.[26] Such disparities raise concerns about accuracy and fairness, but prior efforts to reduce disparities through

internal oversight have been held to violate principles of decisional independence under the APA.[27] These challenges of ensuring the accuracy and consistency of decision-making at SSA have persisted through decades of quality improvement efforts.[28]

2. Capacity Building

As part of its efforts to address these challenges, the SSA has undertaken a series of efforts to improve its data infrastructure and develop staff with technical competence to improve efficiency and accuracy in processing claims in data driven ways.

SSA has long sought to improve the quality of data available for data analysis to improve the efficiency and quality of adjudications. In contrast to other adjudicatory agencies, SSA adopted an electronic case management tool early on.[29] One of its current efforts involves obtaining medical records in electronic format and converting medical records, presently received as image files in most cases, into text files. SSA is currently in the process of digitizing its request-generation process and planning to obtain medical evidence in more accessible formats. In 2015, SSA also started to develop software to convert images of medical records to readable text. The SSA CARES initiative is planning to develop and pilot a software that uses artificial intelligence and NLP to automatically scan case files, identify duplicative medical evidence, and suggest those pieces of evidence for removal by SSA staff. Notwithstanding these efforts, significant challenges to the data infrastructure remain.

Agency leadership and entrepreneurship appears to have played a particularly important role for positioning SSA to be able to take advantage of data analytics. Gerald Ray, who spent most of his career at SSA, becoming an Administrative Appeals Judge and then deputy executive director of the Office of Appellate Operations (OAO), played an important role in creating the seedbed for prototyping AI tools. Described by one co-worker as the "Steve Jobs of the SSA," Ray realized early on that the formalization of adjudicatory policy, capturing data streams, and analytics could help address longstanding challenges of SSA adjudication. Because OAO was only authorized to employ attorneys, however, he strategically identified attorneys who also happened to have a knack for data analysis and software engineering, which became the core team building out SSA's early prototypes.[30]

II. AI USE CASES

We now describe three novel experiments in the use of AI/ML at the SSA in increasing order of sophistication, aimed to address the caseload, accuracy, and consistency challenges.

A. Clustering for Micro-Specialization

The Appeals Council has explored the use of clustering algorithms to improve case processing. The existing approach randomly assigned cases to adjudicators. SSA hypothesized that case clustering could help adjudicators accumulate expertise about and reduce research time into policies and regulations by examining similar claims together, potentially reducing case processing time and errors. One potential for case clustering would have been to develop specialty units, wherein adjudicators exclusively focused on distinct areas of law. Yet in response to concerns about differential workloads, the clustering algorithm was applied so that each adjudicator would (a) receive the same (randomly selected) set of cases as before, but (b) receive the suggested order in which to process them. Such "micro-specialization" might still enable adjudicators to develop expertise in one area and apply the relevant statutes and regulations to comparable cases.

The clustering tool uses meta-data available in SSA's case management system—including claimant's age, impairments, state of origin and other facts developed at the hearing level—to provide clusters of similar claims.[31] SSA reported 12% reduction in case processing time and 7.5% reduction in returns from administrative appeal judges to attorneys.[32] That said, the program is being used only on a voluntary basis and it is unclear how these effects were calculated. If more motivated adjudicators adopted micro-specialization or if the program was implemented along with other quality improvement efforts, the reported benefits may be overstated.

B. Accelerating Appeals with Predicted Likelihood of Success

In another effort to reduce case processing times, the SSA has developed two mechanisms for expediting claims likely to receive benefits.

To improve case processing at the initial application level, SSA promulgated new rules that included provision for automatically identifying claimants most likely to qualify for benefits for Quick Disability Determination (QDD).[33] The program identifies claims where benefits are likely to be awarded and where the information needed to make the

disability determination can be obtained quickly and easily. Each DDS unit establishes a QDD unit devoted to processing claims referred by the predictive model.[34] After referral, each QDD determination "involv[es] sign-off by both an examiner and a medical expert."[35] The expectation was that a small number of claims would qualify, but that more cases would be processed in this manner over time.[36] The proposed model was meant to identify claims using scores based on "such factors as medical history, treatment protocols, and medical signs and findings."[37] Features were selected based on prior DDS determinations.[38] In 2010, SSA revised its regulations to permit QDD examiners to grant claims without medical consultation,[39] while still requiring such consultation for denials.[40] Some scholars expressed enthusiasm about QDD: "The addition of QDDs at the initial decision level for selected types of claims—those where the information needed to decide disability can be obtained quickly and easily—is also a positive and practical approach to case management. Setting apart claims for which the evidentiary record may be compiled with little difficulty will allow SSA to direct much-needed resources to more difficult claims."[41]

SSA has also developed tools to expedite claims at the hearing level by predicting which claims were denied reconsideration but have a high likelihood of receiving benefits.[42] The predicted probability of a grant is "not used to adjudicate the cases, and the probabilities of allowance [are] not shared with adjudicators, but cases with higher probabilities of allowance [are] moved ahead of cases in the queue with lower probabilities of allowance under the notion that disabled claimants should receive their decisions as soon as possible."[43] Specialized teams then analyze the cases based on the estimates.[44] The model also seeks to identify claims dismissed for procedural reasons but that would otherwise be paid to ensure such claims receive proper review.[45] The supervised learning model (Naïve Bayes) uses outcomes at the hearing level (fully favorable, favorable, unfavorable, or dismissal),[46] age, and impairment as features. Officials at SSA reported that the model identified 10% of cases as likely to receive fully favorable as compared with the average fully favorable rate of 2.5-3% for all claims at the hearings level.[47]

C. Natural Language Processing for Quality Assurance

SSA has a long history of initiatives for quality assurance.[48] The most ambitious technological version consists of a suite of tools based on natural language processing (NLP). Known as the Insight program,[49] these tools were principally developed by Kurt Glaze, an SSA attorney-turned-programmer, primarily to improve the quality of decision writing. At the hearing level, Insight is used to identify weaknesses in draft opinions, ensuring that adjudicators have properly gone through the analysis required by regulations. At the Appeals Council level, it is used to identify inconsistencies in opinions appealed from ALJ decisions, either by claimants or on the Council's own motion. At the Appeals Council level, use is voluntary. At the hearing level, use is required for fully favorable decisions but voluntary for unfavorable decisions.[50] Since August 2017, the tool has been used 200,000 times at the Appeals Council level.[51] Since October 2018, the tool has been used approximately 70,000 times at the hearing level.[52] The SSA may be further expanding the program—Congressional testimony in the summer of 2018 indicated an intention to expand the use of NLP for quality review.[53]

Insight analyzes a draft decision and calculates a set of over 30 "quality flags" that are suggestive of policy noncompliance or internal inconsistencies in the decision.

Insight consists of several distinct parts. First, using information extraction from the case management system, Insight provides a case summary, including claim type, disposition, claimant information (date of birth, claimed onset dates, body mass index), and acquiescence rulings for the region. Second, Insight analyzes a draft decision and calculates a set of over 30 "quality flags" that are suggestive of policy noncompliance or internal inconsistencies in the decision. These flags can range from the simple (*e.g.,* the decision cited a provision in the C.F.R. that does not exist, the claimant's age and vocational grid rule are inconsistent) to the more complex (*e.g.,* the claimant's capacity is inconsistent with job requirements in the Dictionary of Occupational Titles). Critical in this analysis was the existence of SSA's policy decision tree that spells out the appropriate paths that an adjudicator could take, leading to approximately 2,000

possible case types within the five-step process described above.[54] At the Appeals Council level, reviewers answer a series of questions on the page "to verify that the hearing decision meets the requirements for a legally sufficient application of policy in deciding the disability claim."[55]

Because over 30 quality flags exist, there is no single algorithm underpinning the Insight system. The existence of the Findings Integrated Template, standard templates of decision writing for case types, facilitates information extraction from opinions and many quality flags are a hybrid of logical rules and supervised learning. For instance, a rule-based heuristic (regular expression) might identify specific language indicating the claimant's maximum capacity to perform certain functions, but a probabilistic NLP classification algorithm may also be developed based on (relatively) small training datasets as a fallback. The most advanced component uses part-of-speech tags and supervised classification to predict internal inconsistencies. Flags are included in the Insight system only if they have over 80% accuracy. Some of the most challenging scenarios where flags are inaccurate come from cascades of failures—for instance, if the optical character recognition engine improperly reads an impairment, multiple inaccurate flags may be thrown.

According to internal reports, the Insight system reduces processing time and the number of returns to staff making initial determinations at the Appeals Council level.[56] In an audit of the Insight system, SSA's Inspector General reported that 80% of survey respondents found the flags to be accurate while 20% found that the flags were not accurate. Only 30% of respondents found that Insight improved case processing time. Overall, the Inspector General recommended that the agency "develop metrics to determine whether Insight is achieving its goals."[57]

III. FUTURE TRAJECTORY OF AI AT SSA

SSA's prototype applications provide a blueprint for technology to address longstanding challenges of agency adjudication. The use of AI to enable adjudicators to process cases in a more coherent order (through micro-specialization), to triage cases in a data-driven way to fast-track those amenable to quick resolution, and to develop quality flags for written decisions each have potential applicability for other adjudicatory agencies, such as the Board of Veterans Appeals, the Office of Medicare Hearings and Appeals, and the Executive Office for Immigration Review.[58] The BVA, for instance, has recently experimented with a case specialty team, allowing for specialization across adjudicators, not only in the order of processing within adjudicators. SSA is by far the most advanced in its adoption of these tools and there is significant interest in tools that might enable other agencies to improve case processing. We now spell out some of the challenges that may limit the trajectory of AI applications in mass adjudication.

A. Improving Data Quality

The success of AI depends first and foremost on data. While SSA has rich raw data, there are significant impediments to being able to build out a more ambitious AI pipeline.

First, much data remains unstructured. The prior employment field, for instance, is a freeform field that applicants enter (sometimes in handwritten form), making it challenging to incorporate in a machine learning pipeline. Legal and regulatory issues are not used for clustering, as they are embedded in the text of decisions, leading some branch chiefs to be more hesitant about adoption. Medical records are available only in PDF format. Significant efforts are required to extract, standardize, and validate such information from raw records. As a result, SSA's current applications are limited in the amount of information they are able to incorporate. For its clustering algorithm, for instance, SSA relies on a limited number of basic demographic features (*e.g.,* age, impairment, and state of origin). The shift toward electronic health records (*e.g.,* Fast Healthcare Interoperability Resources) and intermediate tools to extract medical conditions (*e.g.,* Amazon's Comprehend Medical) may help to solve this bottleneck in the near future. If medical records were available in structured form, for instance, AI tools could be built out to aid in the time-consuming process of reviewing the claims folder.

Second, the structured data itself exhibits quality issues. One of the more important quality flags of the Insight system, for instance, compares maximum capacity to complete work-related physical functions with structured data on employment qualifications in the Dictionary of Occupational Titles, last updated in the late 1990s. Yet this structured data is increasingly becoming outdated and may not match the range of occupations in the current economy.[59] As put by one commentator, "There's widespread agreement

among the legal community and the SSA that the Dictionary of Occupational Titles is outdated, with inaccurate job descriptions, obsolete jobs, as well as missing jobs."[60] In addition, if this data is in fact updated, it may require significant resources to adapt to the rule-based system that SSA has developed.

Third, labeling data is expensive. This stands in contrast to many benchmark machine learning tasks (*e.g.,* sentiment analysis of movie reviews, object detection in images), where labels can be crowd-sourced by lay coders. Providing ground truth labels for legal decisions can be an expensive endeavor, requiring attorneys well-versed in the legal area to label decisions. For instance, whether an opinion applied the appropriate weight to a medical decision in the record requires an understanding of social security law. Generating large datasets of several hundred thousand labeled decisions that are the grist for deep learning can be challenging.

These data pipeline issues are not unique to SSA. OMHA is just now transitioning to an electronic case management system. In recent years, the U.S. Digital Service built out a new Caseflow for the BVA. Such systems promise to provide much richer structured information to facilitate prototyping of AI tools, but each of these agencies will likely face challenges comparable to SSA's. Here again, we see the utility of embedded expertise in helping an agency to identify the most appropriate data sources to build out important achievable solutions.

Glaze: "I developed the flags that I wanted to have available as an adjudicator."

B. Improving Methods and Evaluation

One of the strengths of SSA's AI strategy has been to rely on adjudicators (subject matter experts or SMEs) to determine what tools merit development. Providing acquiescence rulings, for instance, is straightforward based on ZIP-code matching, but requires substantive expertise as to what in fact slows down the decision writing process. As put by Glaze, "I developed the flags that I wanted to have available as an adjudicator." Consider a contrast with a search engine built out for the U.S. Patent and Trademark Office, as detailed in the next chapter. While technically sophisticated for its time, the only users reporting positive interactions with the tool were those with computer science backgrounds, leading the PTO to abandon the tool.

That said, the SME-driven approach constrains the sophistication of AI/ML deployed. The model for predicting grants (naïve Bayes), for instance, imposes a strong independence assumption. While SSA's model predicts that 10% of cases will receive fully favorable decisions, significantly fewer are granted upon review, suggesting that the model may require some more calibration. The model (logistic regression) used for a number of the Insight quality flags does not take advantage of the most important developments in natural language processing—*e.g.,* the deep learning revolution.

We hence spell out several areas where advances in AI may improve existing tools. First, advances in synthetic data generation and differential privacy may enable SSA to provide public data or query-based access to enlist the broader machine learning community in building out solutions.[61] Employing these methods may enable SSA to begin to release certain synthetic datasets to enlist the machine learning community to work on these important social problems.

Second, the lack of large datasets with ground truth (training corpora) may be overcome by deploying "fine-tuning" from (language) models pretrained on large corpora. Until recently, deep learning as applied to NLP required large volumes of ground truth data, making it infeasible to train such models on SSA data. As happened in computer vision years earlier, we have observed breakthroughs in NLP with pretrained models that facilitate fine-tuning. Google's BERT model, for instance, is a model trained on Wikipedia and the Books Corpus with 110 million parameters, learning the context of language better than prior models.[62] Most importantly, it is possible to retrain BERT to achieve substantial gains in benchmark NLP tasks with much smaller training datasets. By fine-tuning such language models, mass adjudicatory agencies may be able to deploy insights gained from much larger corpora to solve discrete adjudicatory tasks, with much less diversion of adjudicatory resources to hand-label case decisions.

Machine learning tools may be built to pinpoint the relevant records and passages in lengthy claims folders, potentially achieving for record review what electronic searchers have achieved for the discovery process.

Third, one of the most promising deployments of AI may lie in lowering the costs for claims folder processing and legal research. Conventionally, adjudication has been labor-intensive, requiring manual review of lengthy claims folders and research into applicable laws and regulations. The provision of acquiescence rulings illustrates how machine learning can create a more streamlined decision-making architecture, reducing the cost of locating circuit-specific policy. More generally, machine learning tools may be built to pinpoint the relevant records and passages in lengthy claims folders, potentially achieving for record review what electronic searchers have achieved for the discovery process. To be sure, before such tools can be built out, agencies need to digitize the claims folder and capture information on the relevant records from manual processing to be able to develop such models. And the task is challenging because any missed record (*e.g.,* doctor's note) could be the grounds for reversal. Nonetheless, these tools could potentially reduce the cost of decision writing by a wide margin. We explore more of these search and classification methods in the context of the PTO in the next chapter.

Fourth, generative language models may begin to automate parts of decision writing. SSA's Findings Integrated Template already provides most of the stock language for decisions, but adjudicators are still required to write portions of the decision. Generative language models have been used to predict words and sentences (*e.g.,* auto-complete). And such generative models could be developed to use structured data about the type of case and the relevant records to draft decision passages. Given the high volume and often formulaic nature of mass adjudications, such tools may ease the burden of drafting decisions.

Last, active learning methods could overcome static deficiencies of these tools. Conventional supervised learning trains a model based on a snapshot of data. As a result, such methods may not be able to account for dynamic changes over time. As mentioned in the enforcement context, past indicators of insider trading may not be future indicators of insider trading given strategic adaptation by regulated parties. In the benefits adjudication context, one acute concern is that machine learning methods may fail to adapt to dynamic changes in the economy with the same degree of flexibility as humans. Methods for active learning counteract this, by using the model to adaptively select units for labeling and updating the model-based outcomes. Such methods could take case triage, such as QDD, to its logical conclusion by deploying adjudicatory resources to where errors are the likeliest and updating the underlying model with each interaction. Adapting active learning principles to adjudication could hence respond to a longstanding critique of agency adjudication as failing to correct systematic sources of error due to arbitrary selection of cases for appeal. Instead, model-based methods could be used to deploy adjudicatory resources to discover systemic sources of error in a dynamic fashion.

We make one last comment on improving the methods for mass adjudication. As in the enforcement context, evaluation is lacking on two levels. First, little information exists to be able to assess the performance of AI tools based on conventional machine learning criteria (*e.g.,* accuracy, precision, and recall in a random test set). Second, little evidence is put forth to sustain claims about the causal effect of adopting these tools. While SSA touts how these tools have improved case outcomes, it provides no details on the method of evaluation. Without such evaluation, it is difficult to verify whether investment into an AI tool is worth the benefit, particularly because the data may not available for an external evaluation.[63] As the Inspector General concluded, we lack measurement "to determine whether Insight is achieving its goals."[64] The failure to implement such new systems without an evaluation plan makes it difficult to learn and generalize from these important interventions, stymieing cross-fertilization across agencies. Here, an important lesson comes from the SEC's openness toward experimentation: Identifying the right set of AI tools necessarily means evaluating them and allowing some to fail.

C. Improving Capacity

The SSA case study is particularly powerful in what it reveals about innovation within government. First, as we detail more fully in a later chapter, SSA illustrates how conventional boundaries (hiring for attorneys vs. hiring of information technology staff) can impede innovation. Gerald Ray was only authorized to hire attorneys, limiting the ability to prototype AI tools within the Appeals Council.

Second, the SSA experience illustrates the importance of blending subject-matter and technical expertise. Kurt Glaze was particularly successful in building out the Insight tool because he spent years deciding cases as an adjudicator, but also happened to have the requisite background and willingness to shift toward software engineering. Such a combination is particularly powerful, but rare. BVA's experience with building out the Caseflow provides a different model: James Ridgway, who helped oversee the project, felt strongly that members of the U.S. Digital Service needed to remain on site to observe in real time how the case management system would be used:

> If people come in for two weeks, do a bunch of surveys, go off site, spend two years building something, and then present it as a finished product, it's going to be a disaster. That happens all the time in the federal government. If you can't get the IT folks living with the people who are going to use the equipment, you should start looking for a new job now, because you want to get out of there before the dumpster fire is so bad that the new IT is leading mission failure.[65]

Third, SSA will need to give serious consideration to how to build out from the initial proofs-of-concept. While its techniques are the most advanced compared to any other adjudicatory agency we are aware of, as we spell out above, they do not yet take advantage of all the data available at SSA nor of state-of-the-art methods developed on machine learning. Finding, hiring, and training individuals with both technical capacity and institutional knowledge can be difficult and expensive. Without requisite technical capacity to calibrate models, AI tools may introduce inaccuracies into the adjudicatory process. Conversely, without the requisite substantive knowledge, an AI tool may improperly encode laws and regulations into computer code.[66] The establishment of the Analytics Center of Excellence (ACE), which aims to house technical and data science talent, is a positive step to institutionalizing AI innovation, though it remains to be seen how successful ACE proves to be in driving forward SSA's initiatives. Many individuals trained through ACE proceeded to help other agencies, limiting SSA's ability to capitalize on this investment.

IV. IMPLICATIONS: THE FUTURE OF MASS ADJUDICATION

At its most ambitious, AI could transform what it means to adjudicate a case. To be sure, current use cases are a far cry from full automation of adjudication, but the trajectory raises profound implications for the APA vision of adjudication. While we reserve a discussion of broader implications for later chapters, we briefly spell out several that are distinct to the adjudicative setting here.

First, the trajectory of AI tools in adjudication raises the normative question about the desired extent of discretion in adjudication. SSA moved early to formalize its policy as a decision tree, but few other adjudicatory agencies have formalized policy to that extent. Such formalization makes it easier to build out AI tools, yet it may be less clear as a normative matter whether such a shift is desirable. More rules-based adjudication may promote consistency, but may also undercut one of the rationales for adjudication: tailoring the application to individualized circumstances.[67]

Second, the development of AI tools raises questions about notice and transparency. Formal adjudication requires that a decision be based on the exclusive record, but AI tools involve a transfer of decision-making authority away from line-level adjudicators toward AI developers. Where the program fundamentally changes the way claims are adjudicated—akin to, say, the establishment of a vocational grid[68]—rulemaking may be required. At the state level, where benefits programs have effectively modified eligibility criteria through the use of algorithmic decision-making, some courts have found that the change violates notice and comment requirements and deprives claimants of due process.[69] The QDD program, for instance, appears to involve a criterion not explicitly envisioned in statute or regulation, screening applications based on the onset date of the disability and ruling out disabilities with earlier onset dates for fast-tracking. While

the use of the onset date might be protected under the APA's exemption for internal rules of agency organization,[70] the decision to privilege certain applications may be closer to occupational guidelines that structure adjudication.[71]

Ridgway: "If people come in for two weeks, do a bunch of surveys, go off site, spend two years building something, and then present it as a finished product, it's going to be a disaster."

Third, the adoption of AI tools could potentially erode the decisional independence of and de novo review by ALJs. An SSA ALJ has a "duty to fully and fairly develop the record and to assure that the claimant's interests are considered."[72] The duty is heightened where the claimant is not represented by counsel but exists also with represented claimants.[73] Clustering, for instance, might empirically narrow the scope of review or analysis to just one particular disability, resulting in insufficient consideration to other disabling conditions.[74] Allowing for voluntary adoption of these decision tools counteracts the potential political pushback by ALJs to perceived infringement on their decisional authority, but such pushback may weaken if adoption is seen to ease the work burden. Automation bias could mean that ALJ review of AI-generated content becomes increasingly perfunctory. Decision writers might increasingly rely on Insight to catch errors, for instance, ultimately ignoring errors that don't have existing flags in the Insight system. This dynamic of overreliance may be particularly acute given the high caseloads that adjudicators face.[75]

Last, as we have seen in many areas of machine learning, the adoption of such tools can heighten concerns of bias. For instance, if SSA is able to incorporate electronic health records for improving its model for expedited processing, the differential take-up rate of electronic health records across demographic groups could generate disparate impact. Understanding this potential for bias underscores the need for internal capacity to monitor and adjust methods.[76]

* * * *

Forecasting the trajectory of AI tools brings into relief longstanding debates about the core values of agency adjudication. At its best, AI may address longstanding problems of the accuracy and consistency of decisions. By increasingly automating core portions of the adjudicatory process, these tools may cut down on staggering agency caseloads without a sacrifice in the accuracy of decision-making. At the same time, this future of algorithmic adjudication may cause us to go back to the basic premises of procedural due process. Why do we hold hearings? Machine learning may enable agencies like the SSA, BVA, and OMHA to expedite decisions by skipping resource-intensive hearings. And while this may meet the goals of accuracy under due process, it may also cause us to revisit the lost constitutional rationale of dignity. The rationale for hearings may not solely be to promote accuracy, but also to explain the law, to engage claimants, and to make them feel heard. This, then, is the challenge of the push for AI solutions in mass adjudication: Agencies seek out these solutions to accelerate case processing, but that same pressure may cause agencies to crowd out the dignitary values of an adjudicative hearing.

Informal Adjudication at the United States Patent and Trademark Office

Informal adjudication is a large residual category under the APA[1] that spans a wide range of decision-making contexts, from government grants by the National Science Foundation to campsite permits by the National Park Service, from gaming licenses by the National Indian Gaming Commission to wastewater treatment plan permits under the Clean Water Act, and from an audit of government contracts by the Department of Defense's Defense Contract Audit Agency to farm inspections by the Food and Drug Administration. Even enforcement decisions covered earlier are, for APA purposes, classified as forms of informal adjudication.

KEY TAKEAWAYS

- The PTO has prototyped AI/ML tools for improved patent and trademark classification and search.
- In patent examination, the reduction of search costs with deep learning models appears particularly promising.
- In trademark examination, the computer vision model to detect visually similar trademarks is one of the more advanced forms of AI/ML.
- Challenges in the adoption of such tools include employee / union resistance, maintaining internal due process, adversarial learning, and potential contractor conflicts.

We specifically focus here on the kinds of adjudicatory proceedings that do not require an evidentiary hearing either under the APA or the enabling act ("Type C" adjudications)—which may comprise as many as 90% of all federal agency adjudication.[2] Per Michael Asimow, "The term evidentiary hearing means one in which both parties have the opportunity to offer testimony and rebut the testimony and arguments made by the opposition and to which the *exclusive record principle applies*."[3] To illustrate the distinction between Type B and C adjudications, consider different patent proceedings. The Patent and Trademark Office's Patent Trial and Appeal Board, amongst its duties, hears appeals of denials of patent applications, with a formal hearing and an exclusive record, and is hence classified as a Type B adjudication. On the other hand, patent examination in the first instance has no such closed record requirement. Because patent examiners are tasked with searching all available databases for relevant prior art, patent examinations are classified as Type C adjudications.

We further focus on the United States Patent and Trademark Office (USPTO) as an illustration of how AI tools have begun to transform informal adjudication. The common challenge running through Type C adjudications is that the open record (*e.g.,* scientific scholarship for the USPTO) can make information management challenging. The USPTO case study illustrates how AI/ML can potentially reduce the cost of such information management.

I. THE PATENT AND TRADEMARK OFFICE

A. Patent and Trademark System

The USPTO is a federal agency within the Department of Commerce responsible for granting and issuing patents and registering trademarks.[4] Customers use the agency to protect their intellectual property by filing applications for these patents or trademarks. When a person seeks a patent from the USPTO, she submits an application to the agency, and an examiner determines whether to grant a patent for her application.[5] Similarly, when a person seeks to register a trademark, she submits an application to the agency, and an examining attorney assesses the application to determine whether the trademark should be registered.[6] At the end

of 2018, the USPTO employed 8,185 patent examiners and 579 trademark examining attorneys, and applicants filed 643,349 patent filings (down 1.1% from 2017), and 638,847 trademark filings (up 7.5% from 2017).[7]

The examination process comprises three steps. First, the USPTO classifies a mark into available codes selected from more than 4,000 possible design codes,[8] or classifies the subject matter of a patent application into one or more codes selected from over 250,000 possible classification codes.[9] For trademark examination, the USPTO identifies design search codes so that attorneys and other applicants are able to "thoroughly and efficiently search the USPTO database" for similar marks.[10] Trademarks consist of text and/or design elements, and each mark is assigned one or more design search codes.[11] For example, one registered trademark for the "PUMA" sportswear brand contains the depiction of a puma and the textual name of the brand. The mark is assigned the design code that falls into the "Animals" category, "Cats, dogs, wolves, foxes, bears, lions, tigers" division, and the "Tigers and other large cats (such as leopards or jaguars)" section.[12]

For patent classification, the USPTO assigns each patent application one or more Cooperative Patent Classification (CPC) codes indicating the relevant subject areas for the claimed invention.[13] The CPC classification scheme is "the result of a partnership between the European Patent Office and the USPTO in their joint effort to develop a common, internationally compatible classification system for technical documents." [14] Its code structure is hierarchical. It contains nine sections—such as "Operations and Transport", "Textiles", "Physics," etc.—at the highest level, and each of these is further subdivided into classes, then into subclasses, then groups, and then main groups.[15] Each additional level in the hierarchy refines the level of specialization for inventions. For example, the section "Physics" contains the class "Measuring Instruments", which further contains subclasses such as "measuring length", "measuring distance", and "measuring volume." CPC code(s) are used to route an application to the appropriate technology centers[16] and determine the art unit, and hence scope of the prior art search.

Second, in part based on that classification, examiners or examining attorneys conduct extensive searches of trademark registrations or prior art (patents, non-patent literature) that would legally disqualify the applicant from obtaining a trademark or a patent. The USPTO currently provides at least two search tools to support prior art search for patents: Examiner's Automated Search Tools (EAST) and Web Examiner Search Tool (WEST).[17] These search tools access published U.S. patent applications, U.S. patents, and some foreign patent documents, and allow search through Boolean Retrieval.[18] Boolean search provides control and transparency in searches due to exact match constraints. For trademarks, examining attorneys similarly perform searches for conflicting marks and review the written application to determine eligibility.[19] The two search systems (X-Search and Trademark Electronic Search System (TESS)[20]) allow examining attorneys and the broader public to conduct searches for text and images in pending applications, abandoned applications, and registered marks.[21] TESS is very similar to the Boolean search system used for patents. It requires examiners to use keywords for textual marks and to manually look up design search codes for designs.[22]

Third, patent examiners determine whether to issue a patent or to reject the patent based on patentability requirements such as novelty or non-obviousness. In the statement of rejection, examiners must include citations to material prior art and "properly communicate the basis for a rejection so that the issues can be identified early and the applicant can be given fair opportunity to reply."[23] Similarly, trademark examining attorneys determine, based on searches for conflicting marks (based on a likelihood of confusion assessment), whether the application is eligible for registration.[24] If not, the attorney issues an action including the grounds for refusal.[25] In both cases, examination concludes either when the application is approved or the applicant abandons the application. The full examination process can hence entail rounds of interactions between the examiner and the applicant.

B. Current Challenges

As in formal adjudication, the USPTO faces quantity and quality challenges. First, the agency has a considerable backlog. In 2018, the average amount of time between a patent application filing and a first action by the USPTO (*e.g.,* a rejection or a notice of allowance) was 15.8 months, 0.4 months greater than the target.[26] The USPTO "receives hundreds of thousands of patent applications every year, and the examiners who process the applications operate under severe time pressure."[27] The backlog is less severe for trademarks, with the average first action pendency of 3.4 months.[28] Second, the USPTO has engaged in a range of quality improvement initiatives to reduce, for instance,

patents that are granted but invalidated or patents that are wrongly denied.[29] Patent examiners spend 19 hours on average per application.[30] They "operate under time and other resource constraints that make it difficult to guarantee the adequacy of the cited prior art for analyzing patentability."[31]

The USPTO's strategic plan at the end of 2018 included goals to improve both patent and trademark examination timeliness and quality.[32] The specific objectives to achieve these goals include increasing international cooperation and work sharing,[33] increasing efficiencies during examination,[34] and leveraging machine learning and AI techniques "to benefit every operational level of the USPTO."[35]

II. AI USE CASES

We now provide an overview of the use cases the USPTO has for patent classification, patent prior art search, trademark classification, and prior trademark search.

A. Patent Classification

The USPTO classifies new patent applications into CPC codes using a third-party contractor.[36] The contractor appears to use a human-in-the-loop approach that combines a machine learning model with human expertise. The supervised machine learning classifier uses the specification, claims, and drawings from the application as inputs, and trains on labels generated by the human experts to learn a mapping from a patent application to the set of output CPC codes, allowing them to "streamline the classification decision process and enhance classification quality."[37] CPC schemes and definitions, however, can change,[38] resulting in a need for the model's training data to be re-annotated. In this case, the human experts continue to process and classify the applications, both old and new, to provide new labeled data to train the model.

B. Patent Prior Art Search

Prior art considered by the examiner can include art the applicant submitted, art received in counterpart applications in foreign jurisdictions, patent and patent-related literature found by the examiner, prior public use, and non-patent literature found through an online search.[39] As the examiner performs the search, she generally records her search history and tracks her search strategies.[40]

Existing search methods rely heavily on matching keywords in the query.[41] In one pilot, the USPTO designed an alternate in-house search tool called "Sigma," which used a more complex document annotation pipeline and a more sophisticated search engine using term frequency inverse document frequency (TF-IDF) scores for retrieving and ranking documents.[42] Yet USPTO never deployed Sigma because it was found to improve efficiency only for examiners with a computer science background.[43]

USPTO never deployed Sigma because it was found to improve efficiency only for examiners with a computer science background.

The USPTO has considered other ways of incorporating machine learning tools into the patent search process. The agency has discussed plans to build "an AI-based search platform" that would use content-based recommendation engines to suggest prior art for a given application.[44] In addition, the agency announced plans to use neural word embeddings (akin to synonyms) to expand the search queries "to promote consistency in searching and to more quickly surface prior art that may be located in any of several disparate databases."[45] Such models use neural networks to learn dense vectors for words in a large collection of documents, such that words appearing in similar contexts have similar vectors. Synonyms are then generated by searching for similar words to those in the patent application, by matching word vectors. Overall, the ultimate goal of incorporating these machine learning models would be to provide cost-effective and time-efficient means for providing the examiner with relevant prior art.[46]

C. Mark Classification

The USPTO has experimented with AI/ML tools to automate mark classification. Historical practice has been exclusively manual. After receiving a trademark registration application, "specially trained Federal employees in the Pre-Examination section of the USPTO review the mark drawing and assign" relevant design codes.[47] The application includes a written characterization of design elements in the mark "to assist the USPTO in making accurate and comprehensive design-coding determinations."[48] After the Pre-Examination section codes the design elements of the application, an examining attorney "reviews the mark, the design codes, and the mark description and may determine whether codes should be

added or deleted."[49] The applicant and the general public then have the opportunity to suggest changes to design codes for the application.[50] At each of the steps, guidelines specify how design codes should be selected.[51]

The experimental AI system aims to suggest trademark design codes. This supervised classification task uses images of trademarks as inputs, with an output of the set of applicable design codes for each image. The deep learning image classifier was implemented in Google's TensorFlow framework,[52] consisting of a convolutional neural network that applies several transformations to an input image, generating a dense image vector or embedding. These image embeddings are meant to represent features most useful for identifying appropriate design codes for the input trademark. The USPTO has also experimented with a model pre-trained on a large image database (ImageNet) to use transfer learning to improve model performance.[53]

D. Prior Mark Search

Because search is a critically important part of the trademark examination process, the USPTO has also prototyped deep learning models that could make retrieval more accurate and efficient. Robust trademark search systems that can achieve high recall or coverage over existing marks can allow the examiner to divert efforts from the time-consuming task of manually searching through tens of thousands of potentially related marks to substantively determining whether an application should be allowed. The deep learning prototype takes as input an image (*e.g.,* the applied-for mark) and outputs a list of the most visually similar images from an existing database.[54] The prototype appears to use an unsupervised approach in which the top matches are presented as a ranked list of similar trademarks.[55] Such a model is likely pre-trained on data consisting of millions of images.[56]

The adoption of deep learning models into classification and prior art search holds great promise for improving the accuracy and efficiency of patent examination.

III. FUTURE TRAJECTORY FOR USPTO

A. Improving Patent Examination

The adoption of deep learning models into classification and prior art search holds great promise for improving the accuracy and efficiency of patent examination. Examiners grapple with information overflow and evolving or non-standard terminology,[57] and legacy systems have limited capacity to cover nonpatent literature and foreign language literature.[58]

First, deep learning tools could potentially improve patent classification. In a survey conducted by the U.S. Government Accountability Office in 2016, 75% of patent examiners claimed to have encountered misclassified applications.[59] Deep learning models can be used to classify each claim separately and then tag the application using the most confidently identified codes.

Second, neural networks could improve efficiency and quality of search. Neural networks are used to learn dense vectors for words appearing in a large collection of documents, such that vectors of words that appear in similar contexts are located close together. Using a system that computes similarity in this vector or embeddings space could allow examiners to search for claims that are relevant to claims in the application being examined, regardless of whether they share terms.[60] This could improve recall of the search process, thus ensuring that highly relevant documents are retrieved. Neural networks could also allow the use of a single tool to search over all text from patents, nonpatent literature, and foreign patent literature. This tool could enable examiners to conduct prior art searches by simply using entire claims from the application as search queries, something that patent examiners expressly desire.[61]

Third, deep learning tools could also be trained to precisely highlight the passages that make a retrieved document relevant, and map them to elements of the current application's claims.[62] Such features—which are far more sophisticated than the current USPTO search system's feature of simply highlighting searched keywords—could drastically reduce the time spent in trying to determine why a document is relevant to a particular claim element in the application. Deep learning tools could also allow examiners to quickly expand queries with retrieved claims written in a different style or using different terminology and language. These tools could also be helpful in decoding relevance of prior art cited by applicants who often supply excessive references that tend

to slow down examiners.[63] Viewing highlighted claims might also allow examiners to piece together evidence showing the obviousness of the proposed invention more quickly.

Fourth, AI/ML tools may aid in searching foreign-language literature. Neural machine translation has improved significantly in the last decade and could be directly incorporated within the prior art search tool.[64] This could serve as an important step in streamlining the patent examiners' search process and alleviating time pressures which make it harder to conduct thorough prior art searches.

Fifth, AI/ML may improve work-sharing and office action drafting. For a particular patent or trademark application, algorithmic systems could leverage rejections rendered in other jurisdictions and provide these rejections to the examiner to determine whether the rejections would be applicable under U.S. law. Such systems could use style transfer tools and sequence transduction models to map specific reasoning in office actions issued in counterpart foreign applications to specific rejections under U.S. patent law. In addition, ML methods could determine which rejections from previous office actions the applicant has properly addressed and which rejections the applicant has not properly addressed, and then populate a draft office action template for the examiner.

Last, improvements may be seen in dynamically updating models for improved generalizability. While the classification and search systems may have efficiency advantages in most cases, they may fall short in other cases. For example, for patent applications on newer subject matters where inventors are just developing new patentable technologies, an examiner using the prior art search system may be unable to find relevant prior art. Indeed, "[t]o the extent that the AI-assisted search used by the Patent Office does not account for potentially rapid change in the average skill of practitioners itself spurred by AI, it will fall short."[65] As is the case with the SSA, active (or online) learning methods may improve generalizability.

B. Trademark Examination

Deep learning models for image classification and prior trademark search may also significantly improve the trademarking process. Yet performance of the tools piloted so far has been suboptimal due to several problems, including class imbalance, duplicate images, and text identification.[66]

First, AI/ML may aid in determining the specific goods and services classification for prior uses. The scope of a trademark depends on the specific goods and services that it is used to sell, and prior use of similar marks used to sell similar goods and services would also lead to a rejection.[67] For searches of prior use, the USPTO could use ML to additionally search for similarity in the goods and services space.

Second, because trademarks can often consist of more than one element, a mark may not be easily sorted into a single category. Design code classification accounts for this by assigning multiple codes to a single mark, with a code assigned to each design element of the mark. The image classification model could be augmented to first identify each design element and then classify each element into a design code class. Using object detection as a first step may also aid in identifying text in an image. Text elements are not assigned design codes but can still prove to be useful during both classification and search processes. Once all the objects have been identified, a classifier can determine which of the objects contain text before recognizing and generating the specific characters and words contained in the image.[68] This text can serve as useful metadata for the trademark, particularly during the search for similar marks.

IV. IMPLICATIONS: THE CHALLENGES FOR AI IN INFORMAL ADJUDICATION

We now highlight some legal and policy implications presented by the USPTO case study.

First, as the USPTO increasingly incorporates AI into its examination process, its results and decisions may be harder to decipher, potentially putting the agency in conflict with administrative law's demand for explainability.[69] To be sure, the effect on applicants' procedural rights is likely minimal. Because a human examiner ultimately reviews the factual record and prepares the reasoning for a decision on an application, AI systems at the USPTO likely would not violate legal due process rights of applicants or their rights under the Administrative Procedure Act (APA). When the USPTO chooses to reject a patent application or a trademark application, under section 555(e) of the APA, the USPTO generally must give a brief statement of the grounds for denial.[70] But with respect to classification and search, the APA does not necessarily impose a specific requirement on the reasoning that the USPTO must provide for the specific classification of an application or for the specific search results.

Even so, explainability remains a normative goal.[71] Internal due process that supports the explainability of examination ensures both quality and efficient examination. The USPTO has set forth guidelines affirming the importance of explainability in patent and trademark examination. For prior art searches, the USPTO has specific guidelines for recording search data "to provide a complete, accurate, and uniform record of what has been searched and considered by the examiner for each application,"[72] explaining that the record "is of importance to anyone evaluating the strength and validity of a patent, particularly if the patent is involved in litigation."[73] Specifically, USPTO guidelines require that an examiner provide search results as well as notes indicative of the nature of the search conducted.[74] Similarly, for trademark examination, the file wrapper includes search histories that specify the key word terms used by the examining attorney in her searches for prior trademark registration applications. The USPTO explicitly refers to this information as being "helpful for internal review,"[75] facilitating supervision and work-sharing.

As more of the classification and search functions migrate to AI-assisted systems, the USPTO will need to consider how to maintain existing forms of internal due process.

AI-based systems may undercut such internal process, as search notes would become increasingly less useful. A supervisory patent examiner would be able to extract little information on the efficacy of a junior patent examiner from simply looking at search notes. In addition, search notes for an application would have less generalizable value for both U.S. and non-U.S. examiners conducting searches on related applications. Examiners for continuation or divisional applications, applications with similar inventorship, or counterpart foreign applications would not be able to easily revise their search strategies based on the search notes. These examiners, at best, could use the search results. Furthermore, the applicant would not be able to easily discern the specific types of prior art that the search system scanned to find the output prior art results. As more of the classification and search functions migrate to AI-assisted systems, the USPTO will need to consider how to maintain existing forms of internal due process.

Second, piloting AI use cases may trigger resistance by examiners and their union representatives. The Patent Office Professional Association represents both examiners and classifiers,[76] and the National Treasury Employees Union, Chapter 245, represents trademark examining attorneys.[77] And, at least on the patent side, the union is "relatively powerful."[78] While unions "should not, in principle, necessarily oppose a tool that would allow more effective search within the same number of hours,"[79] unions may bristle at the prospect of any reduction in hours for examination,[80] potential employment effects, and tools that lack an intuitive and accessible user interface.[81] Thus, the USPTO must ensure that AI tools consider the needs of end-users and articulate a clear vision for AI-assisted examination.

Third, applicants may seek to game AI-based methods to improve their chances of obtaining an allowable patent or trademark registration application. For example, with respect to search, applicants could draft patent or trademark applications such that the search systems do not capture relevant prior art or relevant registered marks, respectively. In the context of classification for patents, artificial intelligence-assisted CPC classification could encourage applicants to draft their applications in a way to achieve a certain classification such that the USPTO directs the application to an art unit having more permissive allowance rates.[82] Such gaming behavior could be interpreted to implicate duties and obligations that practicing patent practitioners and trademark attorneys have to the USPTO. The agency imposes a duty of disclosure, candor, and good faith on individuals associated with filing and prosecuting a patent application, requiring that such individuals "disclose to the Office all information known to that individual to be material to patentability."[83] A violation of this duty can raise inequitable conduct issues during litigation of the patent that could end up invalidating the entire patent.[84] On the trademark side, before the USPTO begins trademark registration examination, the applicant must submit a statement under oath that to the best of her knowledge and belief, no other person has the right to use an identical or similar mark in commerce that would, when used in connection with goods of this other person, likely cause confusion, cause mistake, or deceive.[85] Such guidelines may not yet contemplate knowledge of strategic conduct to avoid

a rejection at the USPTO. To curb such conduct, the USPTO could promulgate rules clarifying these duties and obligations. For example, the duties and obligations of applicants to the USPTO could be clarified to cover strategic conduct like adversarial learning or 35 U.S.C. § 112 could be used to reject applications that strategically use terms to fool machine learning-assisted search algorithms.[86]

Last, because the USPTO relies on contractors to build out some tools, the USPTO must carefully manage potential conflicts of interest. The USPTO "follows the [Federal Acquisition Regulations] as guidance in [their] acquisition decisions whenever it is appropriate to do so."[87] These regulations require contracting officers to "[i]dentify and evaluate potential organizational conflicts of interest as early in the acquisition process as possible; and. . . [a]void, neutralize, or mitigate significant potential conflicts before contract award."[88] The regulations further specify that "contracting officers should obtain the advice of counsel and the assistance of appropriate technical specialists in evaluating potential conflicts and in developing any necessary solicitation provisions and contract clauses."[89] As we spell out in more detail in the Part III of this report, it is unclear how well such conflicts are managed.

* * * *

As illustrated by the USPTO example, the potential benefits of AI/ML in supporting informal adjudication are substantial. AI-supported tools may empower agency officials to divert their scarce time and expertise to other important parts of the informal adjudication process. With respect to trademark registration and patent examination at the USPTO, trademark examining attorneys and patent examiners, not their tools, ensure quality adjudication of trademark and patent rights. Improved search results can provide the factual basis for a rejection, but only trademark examining attorneys and patent examiners have the technical and legal expertise to determine whether applicants are entitled to intellectual property rights. AI-assisted systems ensure that examiners and attorneys can focus their time and efforts on the analysis necessary to provide reasoned decisions for applicants. Artificial intelligence systems thus can be valuable tools for informal adjudication.

Regulatory Analysis at the Food and Drug Administration

Virtually all federal agencies issue statements of general applicability explaining how they expect the public to behave within the agency's regulatory domain. Agencies often make such statements by engaging in rulemaking to establish legally binding regulations pursuant to their Congressionally-delegated authority. Rulemaking, alongside closely related administrative outputs such as standard-setting and "guidance" documents, are at the heart of the regulatory work many federal agencies perform. Indeed, the Congressional Research Service estimates that federal agencies publish between 2,500-4,000 final rules each year.[1]

KEY TAKEAWAYS

- FDA has piloted NLP-based engines for postmarket surveillance of drugs and medical devices based on adverse event reports that contain substantial freeform text.
- Such tools have played a role in facilitating a shift from premarket approval to postmarket surveillance.
- AI/ML-based tools for adverse events can help to prioritize which reports should receive attention, but have been less successful when they verge on attempting to make causal inferences based on unrepresentative data.
- Agencies should consider collecting "structured data" in the first instance, rather than building out NLP-based tools to extract structured data from unstructured text.
- FDA has invested significant resources to develop AI capacity, with benefits extending beyond regulatory analysis.

While it is commonly accepted that rulemaking involves a mix of policy considerations and prudential assessments, rulemaking also routinely involves complicated technical judgments of a predictive or contingent nature. Federal agencies have for many years used statistical decision-making techniques to help make those judgments, but the deployment of AI and machine learning technologies represents a new level of sophistication. Many agencies have begun to incorporate AI/ML into their analytic processes, and their use of such techniques will likely grow more important going forward.

This chapter examines the Food and Drug Administration's (FDA) piloting of AI/ML techniques to identify emerging safety concerns in reports made to its Federal Adverse Event Reporting System (FAERS). Because preapproval studies cannot identify all possible side effects or problems with a drug or therapeutic biological product, the FDA maintains a system of postmarket surveillance and risk assessment centered on analysis of a growing pool of data about adverse events and medication error reports.[2] The agency uses the results of these analyses to update rulemaking and guidance, and, on rare occasions, to reevaluate an approval decision.[3] The FDA has publicly discussed the FAERS pilot projects since at least 2017.[4] This case study is informed by our review of publicly available documents, as supplemented by interviews with key FDA officials.

We present the FAERS pilot projects as illustrative of federal agencies' growing interest in the use of AI/ML to analyze data relevant to rulemaking, standard-setting, and guidance. The FAERS projects likewise highlight the critical need—shared across nearly all federal agencies—to develop internal technical capacity, a topic explored in more detail in Part III. This technical capacity is not only important so that agencies can leverage data in crafting and promulgating rules and regulations, but is also necessary as agencies increasingly regulate the use of AI-powered products and services using agencies' conventional, non-AI-based regulatory instruments.

That said, the FAERS pilots also underscore the challenges agencies face in leveraging growing streams of data in performing regulatory analysis. AI can address some—but not all—of these challenges. To be sure, the FDA is in the vanguard among agencies in its experimentation with advanced AI/ML techniques, including "deep learning" approaches, to meet those challenges.[5] And the agency's FAERS work, while plainly less sophisticated than other FDA efforts, is part of that work and, in addition, sits at the center of a significant policy challenge as the FDA contemplates shifting its regulatory focus from premarket approval to postmarket surveillance efforts. But focusing on FAERS also makes sense because it highlights a key lesson about the possibilities and limits of algorithmic governance tools used in regulatory analysis: While useful, predictive analytics cannot substitute for conventional principles of causal inference.[6]

I. THE FOOD AND DRUG ADMINISTRATION

The FDA oversees products that represent over $2.5 trillion in annual consumption, or about 20% of household spending in the United States.[7] This vast regulatory scope means that even limited use of AI/ML tools by the FDA have a substantial impact on public welfare.

The primary statutory authority governing the FDA is the 1938 Food, Drug, and Cosmetic Act ("FDCA") and its amendments.[8] Under the FDCA, the FDA is tasked with ensuring the safety of food, drugs, medical devices, and cosmetics. While the FDA has expansive rulemaking authority,[9] it has employed guidance documents as its primary means of policymaking for the last several decades.[10] In addition to guidance documents, the FDA uses "warning letters" to communicate directly with firms and to make criminal referrals.[11] The FDA has several additional enforcement mechanisms at its disposal, including recalls, license suspensions, and product seizures.[12] Together, these varied regulatory options assist the FDA in fulfilling its statutory mandate despite being a relatively resource-constrained agency.[13]

Beyond rulemaking and guidance, the FDA also uses a mix of premarket approval and postmarket surveillance methods to ensure the safety of drugs and medical devices. The core of the premarket approval process for brand-name prescription drugs is the New Drug Application ("NDA") process.[14] The NDA process requires that drug manufacturers submit evidence from clinical trials that is sufficient to demonstrate that a drug is safe and effective for its intended use.[15] For generic prescription drugs, the drug manufacturer may submit a streamlined Abbreviated New Drug Application ("ANDA"), which, in lieu of clinical trial evidence, presents evidence that is sufficient to show that the generic drug is "bioequivalent" to an already approved brand-name drug.[16] For medical devices, the FDCA mandates that the FDA classify all medical devices by risk and administer the Premarket Approval ("PMA") and premarket notification ("510(k)") processes.[17] It also authorizes the FDA to ban devices if necessary.[18] To a certain degree, this device approval process mirrors the drug approval one—with the stringent PMA pathway mirroring the NDA pathway and the streamlined 510(k) pathway mirroring the ANDA pathway.

The FDA also conducts extensive postmarket surveillance, collecting and monitoring millions of adverse event reports.[19] The Food and Drug Administration Amendments Act of 2007 ("FDAAA") expanded the FDA's postmarket (post-approval) authority. Passed in the wake of several reports excoriating the FDA's lackluster postmarket surveillance efforts,[20] the FDAAA empowers the FDA to "require a drug sponsor to conduct post-approval studies or new clinical trials at any time after approval of a new drug application if FDA becomes aware of new safety information, . . . to require labeling changes to disclose new safety information, . . . and to require 'risk evaluation and management strategies'[.]"[21] The expansion of the FDA's regulatory authority and resources into the post-approval regulation and surveillance realm reflects a significant agency shift in emphasis fueled by technological innovation and big data.[22]

For several years, the FDA has been investing in human capital in the AI space. Starting in 2017, Dr. Bakul Patel, the Associate Director for Digital Health at FDA's Center for Devices and Radiological Health, hired "13 engineers—software developers, AI experts, cloud computing whizzes—to prepare his agency to regulate a future in which healthcare is increasingly mediated by machines."[23] The FDA also announced the creation of an Entrepreneur-in-Residence program in 2017 as part of its Digital Health Innovation Action Plan.[24] In the FDA's 2019 budget proposal, former Commissioner Scott Gottlieb requested roughly $70 million from Congress to fund a center with significant implications for AI.[25] Gottlieb explained that "the agency would create a Center of Excellence on Digital Health to establish more efficient regulatory paradigms, build new capacity to evaluate and recognize third-party certifiers, and support a cybersecurity unit to complement the advances in software-based devices."[26] According to Gottlieb, AI "holds enormous promise for the future of medicine."[27] Under his leadership, the FDA began work in "the field of radiogenomics, where AI algorithms can be taught to correlate features on a PET or

MRI scan with the genomic features of tumors."[28] Moreover, the FDA "is exploring the use of a neutral third party [to] collect large annotated imaging data sets for purposes of understanding the performance of a novel AI algorithm for a proposed indication."[29]

II. AI USE CASE

Beginning in 2016, the FDA has experimented with AI/ML techniques to analyze data to assist the agency in detecting and addressing adverse drug events brought to its attention through its postmarket surveillance regime. In particular, the agency has sought to develop innovative, AI-based methods to parse the millions of text-based reports of adverse events that flow into the agency's FAERS database.

A. The FAERS Database

The FAERS database is one of several databases the FDA maintains to assist in its postmarket surveillance activities.[30] FAERS contains "adverse event reports, medication error reports and product quality complaints resulting in adverse events that were submitted to FDA."[31] Information in the FAERS database comes from two sources: Patients, caregivers, and healthcare professionals voluntarily submit information to "FDA MedWatch" (5% of all reports), and manufacturers are required to submit information to the FDA (95% of all reports).[32]

FAERS is useful for the FDA's postmarket surveillance regime given the breadth of the information it captures. Submission does not require demonstrating causation (*i.e.,* there is no need for the submitter to show that the adverse event was, in fact, caused by the drug), so the database casts a wide net.[33] Likewise, FAERS reports include information that is unlikely to be captured via clinical trials (*e.g.,* off-label uses, co-morbidities, or long-durational use).[34] The database contains a considerable quantity of information: According to the FDA, over 1.81 million reports were submitted in 2017 alone.[35] FAERS is not, however, without its disadvantages. The database's utility is somewhat limited by duplicative reports, reports of variable quality and completeness, and unverified data.[36] Given the sheer volume of data contained in FAERS, and the varied types of data—both structured and unstructured—the FDA has sought more efficient ways to extract and use this information.

B. NLP for Adverse Event Detection

Details of two pilot FDA efforts demonstrate the potential of using a combination of text mining and NLP to parse adverse event reports and identify emerging safety concerns.

One of the FDA's pilot efforts experimented with NLP techniques to convert the large amounts of unstructured data flowing into FAERS into structured data and then to model relationships between drugs and a single medical condition, hepatic (*i.e.,* liver) failure.[37] FDA analysts first retrieved data of FAERS hepatic failure reports from November 1997 to March 2018.[38] Each report included structured information regarding the patient's age group, the report type (direct/expedited/non-expedited), the seriousness of the condition (serious/non-serious), and one or more reported outcomes (such as death, disabled, and required intervention).[39] Analysts then used a range of techniques to identify important textual cues associated with adverse drug events.[40] First, the project applied text mining[41] and topic modeling[42] to identify important information contained within the materials and to map associations between terms or topics.[43] For instance, the models found a strong association between the terms "hepatic failure" and "death."[44] Second, analysts experimented with text-based rules, decision trees using text clustering inputs, and a simple neural network to predict serious drug-related adverse outcomes. The decision tree performed best in predicting a serious outcome in FAERS cases, with a true positive rate of 91% (*i.e.,* correctly predicted adverse events) and a false positive rate of 4.9% (*i.e.,* non-adverse events incorrectly predicted as adverse).[45]

A second pilot effort mounted by FDA scientists and researchers from Stanford University used similar techniques but adopted a subtly different analytic tack.[46] The team employed three FDA safety evaluators to label a sample of reports on modified World Health Organization–Uppsala Monitoring Centre (WHO-UMC) criteria for drug causality assessment. They then used structured features and expert-derived terms from unstructured text (*e.g.* "drug interaction") to predict these ground truth labels. They trained (regularized) logistic regression, random forest, and support vector machine models and then constructed a rank-ordering of reports based on their probability of containing policy-relevant information about safety concerns. They showed that such a ranking could help prioritize review by FDA evaluators, although there was still considerable predictive uncertainty.[47] Much like the SEC's enforcement tools or the SSA's case-clustering tool profiled above, the tool can be thought of as performing a kind of triage to better target scarce agency resource rather than displacing human assessments.

III. FUTURE TRAJECTORY OF AI AT THE FDA

The FDA's FAERS efforts have been successful, to an extent. The liver-focused project uncovered some previously undetected relationships between hepatic failure and drugs, particularly drug combinations.[48] It also showed some promise in distinguishing between predictors of the degree of hepatic failure from serious and less serious events.[49] The second of the two efforts profiled above also demonstrated potential. The approach identified six data features that can actionably guide the FDA's analysis of reports going forward.[50] The tool, as the FDA and Stanford data scientists noted, can serve as "the foundation" of a system that better economizes on scarce agency resources in identifying emerging postmarket safety concerns.

But the pilots also reveal numerous challenges. The main one was the difficulty of uncovering causal relationships between drugs and hepatic failure using predictive analytics given the available data. By definition, most or all of the FAERS data consists of adverse events, meaning the models are selecting on a negative outcome. Without knowing baseline drug usage within the population, it remains challenging to infer which drugs cause hepatic failure.

These causal inference challenges are exacerbated by the use of NLP methods on unstructured data. NLP algorithms are continually improving. However, as NLP technology currently stands, without some expert input, it may not be well-suited to critical tasks in which inaccurate predictions could have severe life-or-death consequences. NLP's shortcomings are almost certain to be magnified in an environment where the text is both unstructured and highly technical and where precision and expertise are at an absolute premium.[51] While the pilot project served as a useful "proof of concept,"[52] FDA officials conceded that it was not fully successful as it did not generate outputs accurate enough for deployment.[53] The second pilot might hence be a more promising deployment of ML with FAERS data to prioritize how reports are processed.[54]

The future trajectory of these projects remains uncertain, and FDA officials continue to discuss the current and future role for NLP and alternatives at the agency.[55] The FDA may be at a crossroads with respect to whether it continues to use NLP to handle unstructured data, or whether it instead restructures its data collection. FDA officials at the Center for Drug Evaluation and Research (CDER) maintain that there is substantial value in using NLP to understand the FDA's large volume of existing structured and unstructured data. Some advocate exporting NLP applications to other domains, including the vaccine adverse event reporting system (VAERS), which contains "information on unverified reports of adverse events (illnesses, health problems and/or symptoms) following immunization with U.S.-licensed vaccines."[56]

A different path would instead have the FDA focus on requiring regulated entities to submit fit-for-purpose, structured data to the FDA in the first instance, thus obviating the need for NLP techniques to mine unstructured data.[57] NLP and other machine learning tools could then be used down the road to analyze patterns and generate usable predictions. Given that the FDA collects enormous amounts of data annually, it may make sense to require industry to alter its data submissions since highly functional NLP capable of analyzing largely unstructured data may not be developed for some time. The FDA will need to consider how best to harness the data the agency receives, as that could substantially further its ability to employ AI/ML both now and in the future.

In the FDA's case, uptake of AI/ML tools may herald a broader shift away from premarket approval and toward postmarket surveillance efforts.

IV. IMPLICATIONS: THE FUTURE OF AI-DRIVEN REGULATORY ANALYSIS

The FDA's FAERS pilot projects provide valuable lessons for other administrative agencies, especially regarding (i) the potential power of new algorithmic tools to shift regulatory paradigms (in the FDA's case, potentially facilitating a shift from premarket approval to postmarket surveillance); and (ii) the multiple ways in which internal technical capacity can advance the missions of safety-focused agencies like the FDA, particularly by augmenting their ability to regulate AI-based products and services using conventional regulatory tools. We discuss each of these in turn.

A. AI and Shifting Regulatory Paradigms: From Premarket Approval to Postmarket Surveillance

The FAERS NLP pilot projects illustrate how increased use of AI/ML tools may precipitate or accelerate shifts in agencies' regulatory paradigms. In the FDA's case, uptake of AI/ML tools may herald a broader shift away from premarket approval and toward postmarket surveillance efforts.

Although the FDA historically de-prioritized its postmarket surveillance activities, the FDA's leadership—following congressional mandate—has made it a higher priority over the past decade.[58] This shift toward postmarket surveillance is critical given that clinical trials (which are the basis for ex-ante premarket approval of drugs) can mask, or fail to reveal, serious issues attendant to a drug. Clinical trials are limited by a relatively small *n*, relatively brief durations, and selection bias among the patients selected for study.[59] Thus, the FDA must monitor drugs (and devices) after they have been approved to ensure safety and efficacy for all patients.[60]

According to the FDA, its increased attention focused on postmarket surveillance is beginning to bear fruit.[61] The AI revolution could facilitate and dramatically improve such postmarket surveillance. AI/ML tools will make it easier for the agency to capture and mine large quantities of postmarket data. Proliferating use of AI may also render postmarket surveillance practically necessary—for instance, where certain new devices incorporate AI/ML tools that change over time.[62] The use of AI/ML tools can potentially be beneficial beyond ensuring the safety and efficacy of a particular drug or device. As former Commissioner Gottlieb noted: "Traditional postmarket studies typically require years to design and complete and cost millions of dollars." By encouraging collection and analysis of "real world data" and "real world evidence" and developing the analytic tools and techniques necessary to use it, the FDA "may be able to provide patients and providers with important answers much sooner by potentially identifying a broader range of safety signals more quickly."[63]

B. Building Multi-Purpose Technical Capacity

The FDA's FAERS pilot projects also demonstrate how developing internal AI-based technical capacity will serve multiple agency missions, particularly at safety-related agencies. At the FDA, internal technical capacity appears to be paying dividends beyond the use of AI/ML for regulatory analysis.

Developing internal AI-based technical capacity will serve multiple agency missions.

First, internal technical capacity will be increasingly important as the agency applies its conventional regulatory tools—for instance, approval decisions, as well as issuance of rules and guidance—to the growing set of AI-based products and services offered by regulated parties. Concrete examples abound. The FDA recently gave marketing clearance to several medical devices that incorporate AI: Viz.AI[64] detects strokes, OsteoDetect[65] recognizes bone fractures, and IDx-DR[66] identifies diabetic retinopathy. These devices went through the FDA's de novo review process, an alternative pathway for "novel devices of low to moderate risk" to gain approval to "be marketed and used as predicates for future 510(k) submissions."[67] In its market authorization of Viz.AI, the FDA specified that it is "creating a regulatory framework for [clinical decision support] products that encourages developers to create, adapt and expand the functionalities of their software to aid providers in diagnosing and treating diseases and conditions."[68] The FDA cleared each of these devices for marketing on the basis of similar criteria: performing better than the existing baseline, often a human medical professional.[69]

The FDA's market authorization of these AI-devices and its recent release of a discussion paper on its plans to regulate AI/ML-based software as a medical device[70] suggest that the FDA aims to move quickly on the AI/ML front. The FDA also recently provided additional guidance as part of its Digital Health Innovation Action Plan to give "more clarity on [its] risk-based approach to digital health products."[71] However, the AI-devices that the FDA has cleared for marketing with de novo review represent a small swath of the future of potential AI-devices. To date, the reviewed devices have been "locked," in that they "don't continually adapt and are dependent on updates from the manufacturer, which can include training the algorithms with new data to improve their performance."[72] But, as former FDA Commissioner Scott Gottlieb noted, "[T]here's a great deal of promise beyond locked algorithms that's ripe for application in the health care space."[73] Regulating AI-devices will grow increasingly complicated as medical devices incorporate AI that dynamically updates in the future.

Second, internal technical capacity empowers agencies like the FDA to make smart investments in data infrastructure as building blocks to make effective use of AI/ML tools going forward. The FDA's Information Exchange and Data Transformation (INFORMED) illustrates how internal technical capacity can shape data infrastructure.[74] INFORMED is an oncology data science initiative, motivated in large part by the perception that the FDA was not capitalizing on the data it collects.[75] Although ostensibly oncology-focused, INFORMED's goals appear to be broader. INFORMED acts as "a sandbox," wherein the FDA pairs "new talent such as entrepreneurs-in-residence, engineers, and data scientists with subject matter experts such as oncologists at the FDA,"[76] in order to expand the organization's capacity for big data analytics.[77] INFORMED emphasizes "data sharing and the creation of new data assets,"[78] as well as "opportunities for machine learning and artificial intelligence to improve existing practices."[79] The group is currently working on several projects, including some with other public and private entities.[80]

In addition, internal technical capacity appears to be helping the FDA collect real-world data. For example, the National Evaluation System for health Technology ("NEST") was designed to "help improve the quality of real-world evidence that FDA can use to detect emerging safety signals quickly and take appropriate actions."[81] The MyStudies App[82] was built to "foster the collection of real world evidence via patients' mobile devices" in a way that is both useful to manufacturers and "compliant with the FDA's regulations and guidance for data authenticity, integrity and confidentiality."[83]

In sum, while the specific FAERS NLP applications may for the moment have limited utility, FDA's AI investments have generated a range of innovative efforts within the agency.

* * * *

While many challenges lie ahead, NLP-based regulatory analysis promises to transform the work of the FDA and other agencies in the years ahead. For NLP tools to be successful, it is imperative that the FDA and other agencies cultivate internal technical capacity—both to leverage a dizzying array of data and to better regulate new AI products and services. Moreover, as agencies like the FDA become more reliant on AI, they will likely have more time and tools to devote to tasks that were once cumbersome and costly, shifting agencies' regulatory paradigms in the long run.

Public Engagement at the Federal Communications Commission and Consumer Financial Protection Bureau

Administrative agencies increasingly use AI/ML tools to engage with and provide services directly to citizens. These engagements include "customer service" interactions, such as applying for a passport, a license, or benefits.[1] They also include interactions facilitated by "civic tech" applications,[2] such as open data portals and chatbots.[3] Where successful, such tools can streamline and improve the quality of diverse interactions between the public and government. This chapter explores the use of AI/ML to streamline two related forms of citizen engagement: notice and comment rulemaking and complaint review.

KEY TAKEAWAYS

- With the availability of online portals, many agencies have grappled with the sharp rise in volume of complaints or comments submitted in notice-and-comment rulemaking.
- Sentiment analysis, topic modeling, and information retrieval can be useful tools for agencies to process comments and complaints submitted by citizens.
- As with Regulations.gov, such tools may be useful across a wide range of agencies, raising the question of how to coordinate interagency efforts.

We begin with a brief overview of the growth of notice and comment in the digital age. We then examine how AI/ML was used to analyze comments during the Federal Communications Commission's (FCC) Net Neutrality rulemaking, as well as how the Consumer Financial Protection Bureau (CFPB) has used NLP to process consumer complaints. We conclude by discussing the broader legal implications of deploying AI/ML in these contexts.

I. THE GROWTH OF COMMENTS IN RULEMAKING

Federal agencies publish between 2,500-4,000 final rules each year.[4] Most use the interagency website Regulations.gov to coordinate the notice and comment process. The back-end system undergirding Regulations.gov allows agencies to track, review, and publicly re-post comments to Regulations.gov so that other interested parties are able to view what has been submitted.[5] Although the system can sort and group comments based on some basic criteria, it does not deploy ML.

Online platforms have lowered the cost of participation and agencies across the federal government are experiencing an overall increase in public comments to proposed rules.[6] Under the APA, agencies must give "interested persons" an opportunity to comment "through submission of written data, views, or arguments."[7] Most comment periods typically range from thirty to sixty days.[8] Executive Order 12,866 requires that interested persons have at least sixty days to comment on "significant rules,"[9] after which agencies must disclose any information they have relied on in drafting a final rule.[10] These requirements—a "safeguard against arbitrary decision-making"— are critical to insulating rules from legal challenge.[11] Failure to comply with APA requirements can spark litigation, especially in the case of major or controversial rules.[12]

The internet has also led to *mega-participation*, in which regulators have begun to receive an unprecedented quantity of comments from a wider array of stakeholders.[13] Where a proposed rule garners widespread attention because of media coverage or organized efforts to mobilize the public, agencies often receive

comments that––however relevant in principle––may be poorly suited for analysis by the coterie of agency staff who normally tend to the notice and comment process. Even when the majority of comments come from lawyers and sophisticated technical experts, the detailed information they contain can challenge and sometimes overwhelm agency capacity. AI/ML tools might help federal agencies respond to this high volume of information.[14] For instance, AI/ML tools can help to identify duplicates and form letters, summarize overall comment sentiment, and identify relevant comments that could save significant resources and enhance the quality of the rulemaking process.

Recent years have seen tremendous advances in NLP tools that could streamline agency processing and analysis of public comments.

II. AI USE CASES

At a high level, the core technical task agencies face in the notice-and-comment process is large-scale textual analysis. Recent years have seen tremendous advances in NLP tools that could streamline agency processing and analysis of public comments. Agencies and private-sector actors have become more interested in using technical tools to assist in the notice and comment process, but many potential solutions remain at the prototype stage.[15] While some agencies are working with third-party organizations to develop scalable tools, these tools have yet to be integrated into internal agency processes.[16] This chapter considers AI/ML comment analysis by reference to one of the first—and most public—use cases: the Net Neutrality rulemaking at the FCC. We then consider analogous use cases by the Consumer Financial Protection Bureau for complaint processing.

A. FCC's Net Neutrality Proceeding

In June 2014, late-night comedian John Oliver released a viral monologue critiquing the FCC's proposed net neutrality regulation and encouraging viewers comment on the proposed rule.[17] In response, nearly 3.7 million comments flooded the Federal Communications Commission (FCC) website.[18] In May 2017, in response to another round of net neutrality rulemaking, Oliver yet again urged his viewers to comment on the FCC's proposed rollback of the prior regulation. He directed his viewers to a link titled "GoFCCYourself" that in turn redirected to the official FCC rulemaking webpage.[19] The agency's site was again deluged with comments that overwhelmed the agency's servers.[20] Over twenty million comments were ultimately submitted on the proposed rollback.[21] A significant number of these comments (1) included false or misleading personal information, (2) were part of an organized grassroots or "astroturfing" campaign, or (3) were submitted at exactly the same time.[22] An independent study found that as many as two million of the comments were "fake."[23]

Broadband for America engaged the consulting firm Emprata to analyze these comments using a range of basic NLP tools.[24] As we do not have access to a government-commissioned report, the Emprata report provides one illustration of how NLP might be used to analyze a large volume of comments. A key component of Emprata's work was a sentiment analysis of submitted comments in order to extract their most salient features. Sentiment analysis is an especially challenging problem for NLP. Human language constructs often rely on contrasting sentences (*e.g.,* "I hate this agency but this rule is not bad!") or sarcasm (*e.g.,* "Rescinding net neutrality is a great idea. Let's return to the Stone Age, too").[25] Heuristic approaches (*e.g.,* assigning a sentiment score to each word in a sentence) and simple NLP models (*e.g.,* so-called "bag-of-words" approaches that disregard word ordering and grammar[26]) tend to be less effective at such extraction,[27] although neural networks are more promising.[28]

When it analyzed the comments the FCC received, Emprata found that only 8% of the more than 21 million comments were unique, and "[t]he top 10 and 100 most prevalent comments accounted for 66% and 89% of the total comments, respectively."[29] To measure overall comment sentiment, Emprata "employed a hybrid text mining approach consisting of (1) manual sentiment assignment, (2) keyword/phrase matching, and (3) natural language processing (NLP)."[30] This tiered approach enabled Emprata to first manually assign the 500 most prevalent comments into groups based on whether they supported or opposed the proposed rule.[31] The firm then used simple keyword and phrase matching rules to assign the majority of the remaining comments to a group, validating the results by manually verifying a random sample.[32] Emprata

applied machine learning only to the final 1% of comments, using NLP classification.[33] "After multiple iterations, the NLP model reached an overall accuracy of 95.2%,"[34] although it was more successful at classifying comments against repeal rather than for repeal.[35]

Pure sentiment analysis may prove insufficient, however, when dealing with the kinds of mass-generated comments that begin with a common template and replace sentences with semantically identical but reworded content. This might occur when an advocacy organization sends a form letter to its members and requests that they edit it before submission. Bots can use this strategy at scale to generate large amounts of technically unique yet semantically similar comments.[36] Previously-uncovered bot comments, such as comments on the FCC's net neutrality rule, appear to stem from a simple generative process that starts from a template comment and randomly replaces words with synonyms or entire sentences with semantically identical phrasings.[37] This process, when applied at scale, is detectable using comment de-duplication and clustering techniques.[38] But as bot-generated comments become increasingly similar to human-written comments,[39] agencies must exercise caution. Even humans often use similar generative processes as part of common mass-commenting efforts coordinated by advocacy organizations.[40] The Electronic Frontier Foundation, for example, published a "DearFCC" automated comment-generator for use by concerned citizens.[41] Such tools may generate outputs that resemble bot-generated comments.

The FCC uses an online comment submission system similar to Regulations.gov, which allows users to enter demographic data that the agency does not verify at any stage in the process.[42] In its overall sentiment analysis, Emprata "did not eliminate or discount comments that seemed artificially generated, duplicated, or submitted by actors who may have intended to influence the final sentiment tally."[43] Nor did it remove multiple comments from the same submitter.[44] Because it could not verify any comments, Emprata found it "very difficult to draw any definitive conclusions" without assuming some subset of the data was "real."[45] It did, however, find that controlling for the data discrepancies it discovered would have supported the opposite conclusion: The majority of unique, American comments *favored* repealing net neutrality protections.[46]

To analyze comment authenticity, Emprata first conducted a rudimentary data scrubbing, marking indecipherable comments (such as those that contained random characters) and comments submitted with incomplete or indecipherable demographic information. Emprata also validated the addresses of 65% of the comments using a geocoding service.[47]Although the FCC had received "significantly more unique/non-form letter comments" *against* repealing net neutrality,[48] a large percentage of those comments came from "email domains associated with FakeMailGenerator.com" or other unverifiable email sources.[49] Comments against repeal also included "[e]ssentially all international comments," [50] and "[t]he majority of duplicative comments" submitted using the same address or email.[51] Controlling for these discrepancies would have thus led Emprata to the opposite conclusion in its sentiment analysis. For these reasons, Emprata found that its pure sentiment analysis was perhaps telling the wrong story by overcounting duplicate, inauthentic, or bot-generated comments.

B. CFPB's Reliance on NLP

We can develop a further sense of the potential use of NLP in citizen engagement by considering how the CFPB has deployed such tools in processing consumer complaints.

To balance the unprecedented scale of consumer complaints relative to the CFPB's resources and personnel capacity, the agency deploys NLP to automatically analyze text to categorize narratives, identify trends, and predict consumer harm.

As a consumer financial protection agency, the CFPB has received more than 1.5 million consumer complaints since it was established in July 2011.[52] Processing, prioritizing, and responding to the thousands of consumer complaints submitted weekly is the CFPB's main regulatory and administrative challenge. To balance the unprecedented scale of consumer complaints relative to the CFPB's resources and personnel capacity, the agency deploys NLP to automatically analyze text to categorize narratives, identify trends, and

predict consumer harm.[53] The CFPB's NLP tool augments its platform for processing complaints (the "Consumer Response System"), and the agency's strategic plan and budget report in March 2014 reveals that the CFPB is investing annually ($10.7M in 2014, $8.1M in 2015) to develop this system.[54]

The Bureau makes publicly available all complaints in its Consumer Complaint Database, which removes personally identifiable information (PII) then shares the allegations to encourage consumer awareness of potential and repeat violators.[55] This public database discloses the company name, financial product and sub-product, details of the issue, and whether there was a timely response by the company.[56] Consumer narratives are coded by the nature of the issue and corresponding financial product, such as "incorrect information on [a] credit report," which is regulated in part by the CFPB through the Fair Credit Reporting Act. The CFPB maintains the Consumer Complaint Database with the caveat that the agency does not verify all the facts alleged in the complaints. However, the tool provides an opportunity for the companies to publicly respond to complaints and publish their resolution, which minimizes the enforcement tasks of CFPB.

Although the CFPB does not publish full technical details, it appears to use NLP for two primary purposes. First, the Bureau uses software to scrub PII from the Consumer Complaint Database. The first step of the scrubbing process is done by automated software. Two further steps include separate reviews by a trained human reviewer and a quality assurance specialist to further ensure that the data has been effectively de-identified.[57] Given that this is a three-stage process—with two of the steps being controlled by human readers—the scrubbing process appears to be very thorough. The computer-generated scrubbing makes use of open-source part-of-speech tagging software to determine which parts of the speech are pronouns, verbs, etc.[58] Such de-identification may be a way to ensure authentication of submissions while protecting the privacy of submitters in notice-and-comment.

Second, the CFPB is deploying contextual NLP tools to categorize complaints via topic modeling. The agency uses an off-the-shelf Structured Topic Model[59] (STM) that builds on Latent Dirichlet Allocation (LDA).[60] Such topic modeling may enable agencies to quickly build typologies of the types of comments submitted. Coupled with anomaly detection, such NLP techniques could facilitate the retrieval of relevant comments more efficiently.

III. FUTURE TRAJECTORY OF AI TOOLS FOR COMMENT AND COMPLAINT ANALYSIS

Given the resource-intensive nature of the comment and complaint analysis process, agencies have a strong incentive to build AI/ML comment analysis tools. Some agencies, such as the Department of Health and Human Services, have database management systems that can run rudimentary comment de-duplication or grouping, while others outsource this work to contractors.[61] The Wage and Hour division of the Department of Labor, for example, has relied on a private contractor to help sort through comments and identify those that may warrant further attention from the agency.[62] Still other agencies, however, rely on manual labor to sort, process, and respond to comments.[63] Given these expensive investments in comment analysis, the long-term savings of an AI/ML notice-and-comment tool are likely to outweigh development and adoption costs, particularly for agencies such as the EPA and FCC, which have continued to see a rise in comments.[64]

Despite the fact that agencies face common challenges with complaint submissions and notice and comment, developing an efficient, widely applicable interagency AI/ML tools requires overcoming a collective action problem. The history of Regulations.gov illustrates this concern. Regulations.gov was launched in 2003 by the EPA and later shared with other agencies.[65] While these latter agencies benefitted from the site's user interface and comment storage functions, agencies had little incentive to contribute when the EPA, which still maintains Regulations.gov, hired consultants to upgrade the site's capabilities.[66] Meanwhile, some agencies, such as the FCC and FEC, maintain their own independent notice-and-comment sites.[67] Hence, while it may be most efficient for agencies to pool their technical and financial resources to develop an effective set of tools built upon a broad dataset, it is more likely that an agency facing an especially burdensome notice and comment process will have to take on the burden and then share the tool over time, as the EPA did with Regulations.gov.

IV. IMPLICATIONS

AI/ML comment analysis sits at the center of growing debate about the legitimacy of the rulemaking process. For the moment, Regulations.gov—the platform used for most federal notice and comment processes—merely requires that users complete an optional web form with basic identifying information. Even users who indicate they are submitting a

comment on behalf of a third party can, but are not required to, disclose the name of the third-party. Given these minimal requirements, some proposed rules are flooded with bot-generated comments as well as comments that use fake names or that impersonate prominent public figures (possibly written by humans or bots).[68] The problem of bot-generated and inauthentic comments may well worsen with time.

Automated comment analysis can help mitigate these problems, but its use also raises a host of further concerns. First, bots can generate mass comments intended to persuade an agency of a particular approach to a proposed rule, but they can also be deployed to intentionally overwhelm an agency's capacity to respond to genuine, human-submitted comments. Where bots aim only to overwhelm an agency, use of automated comment analysis to identify and filter out perceived bot comments should be uncontroversial. Identifying and filtering out bot comments ensures that human voices are not drowned out. However, where bots are used in an effort to shape an agency's decision-making, the rationale for using automated comment analysis is more complicated. One reason is that bots can blur the distinction between grassroots campaigns (*i.e.,* political mobilizations reflecting authentic mass concern and consensus on an issue) and astroturfing (*i.e.,* elite-funded campaigns designed to create an appearance of mass concern where there is none). Bot-generated comments of the latter sort can over-represent the level of genuine concern because they can be deployed by a single actor. That said, the line between grassroots mobilizations is a blurry one.[69] Moreover, it is at least possible that bots can provide useful information to regulators (*e.g.,* using a knowledge base to automate the submission of relevant information). If the purpose of rulemaking is to sharpen an agency's analysis, one might argue that even bot-generated comments should be welcomed.[70] Finally, even where there is consensus on the need to filter out bot-generated comments, calibration can be challenging. An overinclusive filtering tool might reduce the risk of removing human-generated comments but increase the risk of allowing undetected bot-generated comments to color the debate. An underinclusive tool, on the other hand, might reduce the risk of allowing too many bots to drown out human-generated comments at the risk of erroneously marking some human commenters as bots.

Second, in addition to ensuring that relevant and valuable comments do not slip through the cracks, agencies must be especially attuned to any systematic reasons *why* a tool might be overlooking particular comments or information. To be effective, any such tool must first be trained using representative data. This might be especially hard in the context of comments because materiality may vary across language patterns. An NLP tool might fare much worse on comments that use colloquial language or sarcasm, or that contain certain grammatical or spelling mistakes. This may disparately impact certain members of the general public, whose substantive comments could be ignored because they use language the tool mischaracterizes or fails to recognize. An agency which fails to consider a material comment because an AI/ML tool mistakenly categorizes it as immaterial would likely run afoul of the APA's requirement to "consider . . . relevant matter"[71] and address such comments in the statement of basis and purpose. An agency that *systematically* fails to consider comments from specific groups might further raise equal protection concerns. Such a concern is especially important in the case of discrimination on the basis of race or gender where government actions must meet a higher standard of scrutiny.[72] Systematically ignoring comments based on race, gender, or even education level also risks undermining public confidence in AI/ML tools.[73]

Third, contractor relationships may pose unique conflicts of interest in the notice and comment space. Agencies often hire third-party entities, which also advise private-sector clients on commenting campaigns, to build better comment analysis tools. Potential cross-pollination between the private lobbying and government consulting arms of these organizations raises concerns about the integrity of the process. For example, in 2017, consulting firm CQ Roll Call worked directly with the FCC to upload its clients' bulk comments into the FCC comment processing system.[74] More generally, FiscalNote works with government agencies while simultaneously providing "Issues Management" tools to make it easier for private organizations to track regulatory changes and submit comments on proposed rules.[75] Consultants might be able to monetize their access to the inner workings of agency notice and comment systems by advising clients on how to carefully draft comments in order to achieve the desired agency classification and avoid being filtered out by an algorithm.

They may also charge a premium for this "insider" expertise, disadvantaging stakeholders who do not hire their services.

* * * *

While still in their early stages, AI/ML tools may assist agencies with notice and comment analysis and complaint processing—a key function of democratic decision-making—even in the age of mega-participation. Doing so while maintaining fidelity to legal requirements may require thinking through difficult normative questions. Agencies willing to do so may be able to transparently design an effective yet inclusive tool that enhances public participation in the rulemaking process.

Autonomous Vehicles for Mail Delivery at the United States Postal Service

The prior chapter examined how AI can change individuals' interactions with government agencies. The two examples in that chapter — comment and complaint submissions — both concerned the relatively straightforward application of AI to massive text databases to improve efficiency of processing, responsiveness, and analysis. In this chapter, we examine a less intuitive AI application that will also alter interactions with the government: autonomous vehicles currently being developed by the U.S. Postal Service for mail and parcel delivery.

KEY TAKEAWAYS

- The Postal Service has prototyped autonomous vehicles for rural delivery routes, enabling drivers to sort mail while the vehicle is driving along the route, and long-haul trucking.
- The Postal Service anticipates autonomous vehicles will improve productivity and save money on overtime, fuel, and costs associated with collisions.
- Government agencies considering adopting autonomous vehicles will face an uncertain regulatory future on such issues as tort liability and data privacy.

In many ways, autonomous vehicle navigation is unique among AI applications examined in this report. Like many agencies, the Postal Service proposes to use AI to improve and streamline its services. And like individuals who submit complaints to the CFPB or comments to the FCC, individual Postal Service customers will engage with AI technology to varying degrees in the course of their interactions with the agency. But anyone who shares the road with autonomous delivery vehicles or walks alongside will also interact with the technology and be affected by its data collection and analysis. As a physical, mobile manifestation of AI, autonomous vehicle technology thus holds distinct legal implications for the Postal Service and other agencies that adopt it in their vehicle fleets.

I. UNITED STATES POSTAL SERVICE

The Postal Service handles close to half of the world's total mail volume: In 2018 it delivered 146 billion pieces of mail to 159 million delivery addresses.[1] For much of the rural United States, the Postal Service is the only postal delivery option. And, unlike many federal agencies, the Postal Service enjoys favorable public opinion, with three out of four Gallup poll respondents indicating the agency does an excellent or good job.[2]

By statute, the Postal Service has a broad role and mandate, often called the "universal service obligation" (USO):

> The Postal Service shall have as its basic function the obligation to provide postal services to bind the Nation together through the personal, educational, literary, and business correspondence of the people. It shall provide prompt, reliable, and efficient services to patrons in all areas and shall render postal services to all communities.[3]

The agency must also provide "a maximum degree of effective and regular postal services" to urban and rural communities alike.[4] Since the Postal Service is self-financed, Congress granted it monopolies over both letter delivery and mailbox delivery as a means of fulfilling its broad mandate.[5]

Despite its popularity, the Postal Service has struggled with declining revenues and climbing operating costs. It has been in a budget shortfall for more than a decade, recording net losses of $3.9 billion in FY2018[6] and more than $69 billion in cumulative losses since FY2007.[7] Relevant contributing factors include the shift to digital correspondence; related shifts in consumer demand from less labor-intensive services like mail delivery to more labor-intensive services like package shipping; statutory restrictions on pricing for many services and products; and competition from private entities like UPS and FedEx.[8]

As revenues have declined, the Postal Service's labor and operating costs have continued to rise. Personnel costs, which account for about 76% of Postal Service expenses,[9] are projected to increase along with demand for labor-intensive package deliveries.[10] A growing shortage of long-haul truck drivers, in particular, has driven up labor costs.[11] Much of the agency's aging delivery vehicle fleet was acquired as far back as 1987, such that the Postal Service incurs high maintenance costs—about $4,500 per vehicle each year[12] across a fleet of more than 232,000 vehicles.[13] The Postal Service also has to budget around rising (and often unstable) fuel costs[14] and the cost of traffic accidents. Postal vehicles were involved in nearly 30,000 accidents in FY2018,[15] and 11 postal workers were killed in roadway accidents in 2017.[16] The Postal Service paid about $67 million in FY2016 in repair and tort costs related to motor vehicle accidents.[17]

In light of its uncertain financial status, the Postal Service has been designated a "high-risk" agency by the GAO since 2009.[18] A presidential Task Force recently recommended several dramatic changes, including ending collective bargaining for Postal Service workers and lifting certain price caps.[19]

II. AI USE CASE

To combat some of these challenges, the Postal Service has begun developing autonomous delivery and hauling of mail and parcels.

The Postal Service has been testing AI applications for transporting mail and parcels since 2014.

In general, autonomous vehicles navigate via an onboard computer, which uses AI to interpret data from mounted sensors—typically a combination of cameras, radar, and lidar.[20] The computer's AI is trained on driving techniques using data captured from human drivers responding to audiovisual cues on the road. The vehicle's computer combines sensor data with detailed digital maps that indicate road layouts, speed limits, the location of traffic signs, and other information relevant to navigation. Some complex autonomous vehicle systems exchange data with nearby cars to coordinate driving patterns. Depending on their level of sophistication, autonomous vehicles may request or require a human driver to remain on standby to intervene in circumstances.[21]

The Postal Service has been testing AI applications for transporting mail and parcels since 2014.[22] It currently has two pilot-phase projects: autonomous delivery vehicles and autonomous long-haul trucks. The Postal Service also recently issued a Request for Proposal for a third potential application, unmanned aerial vehicles (UAVs),[23] but this chapter focuses on road vehicles since these pilot projects are closer to operational deployment.[24]

The first Postal Service application of autonomous vehicles is for "last mile" delivery—*i.e.,* delivery of mail and parcel to endpoint addresses. In 2017, the agency issued a grant to a robotics laboratory at the University of Michigan to develop a prototype autonomous vehicle for rural delivery routes, the Autonomous Rural Delivery Vehicle (ARDV).[25] The Postal Service determined rural routes were appropriate for testing autonomous technology because they are less congested and have fewer sensor inputs than urban or suburban roads. Under its agreement with the University of Michigan, researchers engineered the ARDV to carry a human postal carrier, rather than conduct the entire delivery process autonomously: The ARDV drives along the route, and the carrier sorts mail between stops and delivers through the window. As designed, the carrier also drives the ARDV manually from the post office to the beginning of the route and manually crosses intersections. The study, which completed in 2018, simulated real world delivery environments on a facility test track but did not test the ARDV in the field.[26]

In February 2019, the Postal Service issued a Request for Information along similar lines to the ARDV for an

"autonomous delivery vehicle that will allow delivery of mail and parcels to curbside mailboxes from the right side of the vehicle" and that "enable[s] the operator to sort and organize mail while the vehicle autonomously drives between delivery points/mailboxes."[27] As of this writing, the Postal Service is reviewing both the University of Michigan project test results and the RFI responses and indicates it will identify next steps based on the results and responses received.[28]

The Postal Service's second AI application is autonomous long-haul trucking. In May 2019, the agency partnered with TuSimple, a self-driving truck company, for a pilot program that hauled USPS trailers between facilities in Phoenix, Arizona and Dallas, Texas,[29] a trip of more than 1,000 miles each way.[30] The pilot study involved five roundtrips, each of which took approximately 22 hours, less than half the typical time of 48 hours required for human drivers, including stops for rest.[31] As with the ARDV, human drivers remained on board for the duration of the pilot. According to TuSimple, all deliveries during the two-week pilot were made ahead of schedule and without any traffic incidents.[32] As with the delivery vehicle projects, the Postal Service indicated it is reviewing the long-haul pilot results and will identify next steps in the near future.[33]

As the Postal Service pursued the above applications, it also conducted opinion research to understand the public's perception of self-driving postal vehicles.[34] In an April 2017 nationwide online survey, the agency's OIG sought to "gauge public perception of driverless technology" for both of the above applications, including "the overall appeal of the technology, the believability of claims about its potential benefits, the public's expected timeframe for implementation, and many of their potential concerns." At a high level, the survey results indicated that the public, while highly aware of self-driving cars generally, has only a "shallow awareness" of their potential application for postal delivery.[35] Once informed of the concept, a large majority indicated they believed autonomous delivery vehicles would be deployed in the near future, but many expressed skepticism about both the potential benefits and safety of autonomous delivery vehicles. Young and urban respondents were more amenable than older and rural respondents, as were those who were already aware of the concept of self-driving postal vehicles.[36]

Less favorably, the survey indicated the public may not trust the Postal Service to lead on autonomous delivery. Asked to rank which of four organizations — the Postal Service, Amazon, FedEx, and UPS — they trusted most to successfully deploy autonomous vehicle technology, respondents ranked the Postal Service last.[37] The OIG cautioned that the Postal Service must monitor public opinion as part of its deployment strategy, and not just "the usual feasibility assessments."[38]

The Postal Service will need to coordinate with the relevant postal worker labor unions as it develops and deploys autonomous delivery vehicles. The status of such coordination is unclear. In late 2018, the president of the National Association of Letter Carriers (NALC), which represents city delivery letter carriers, said the union needed to keep careful watch on autonomous delivery vehicles and other technologies.[39] As of 2018, the Postal Service planned to bring in union representatives to review the ARDV prototype "to discuss some of the human/machine interaction issues and craft employee training guidelines," [40] but the agency did not say whether this consultation took place.[41] The National Rural Letter Carriers' Association (NRLCA), which represents rural letter carriers, indicated the Postal Service did not reach out to the NRLCA about the ARDV project.[42] The Postal Service did, however, notify the NALC of its February 2019 RFI.[43] Following the TuSimple long-haul pilot, the American Postal Workers Union called on members to submit concerns to the Federal Motor Carrier Safety Administration and called for testing "at least as stringent as the requirements for a professional driver operating commercial vehicles on the streets with our families and the public at risk."[44]

III. FUTURE TRAJECTORY OF AUTONOMOUS VEHICLES IN THE ADMINISTRATIVE STATE

Manufacturers are racing to bring automated vehicle navigation to public roadways for a variety of purposes, from personal transportation to shipping and delivery. And, based on its investment to date in the technology, the Postal Service seems highly committed to integrating autonomous technology into its delivery model.[45] It remains to be seen, however, what form such technologies will take and on what timeline they will be rolled out.

Various levels of automated navigation are already on the road. Certain personal vehicles have an "autopilot" or similar feature.[46] Current systems require human drivers to remain ready to intervene, but manufacturers are actively pursuing full self-driving functionality.[47] Shipping vehicle manufacturers and delivery companies themselves are also aggressively

pursuing automated navigation. Amazon and FedEx have both invested in autonomous vehicle navigation startups.[48] UPS has partnered with TuSimple, the Postal Service's partner in the autonomous hauling study, on daily tests of automated long-haul trucks on Arizona highways.[49] TuSimple aims to eliminate the need for "fail-safe" human drivers by 2021.[50]

The precise trajectory of autonomous vehicle navigation will depend on how the fragmented regulatory regime evolves. At the federal level, the National Highway Traffic Safety Administration has endorsed the technology's potential to decrease traffic-related injuries and deaths.[51] But many states have not yet enacted laws about automated vehicles, and those that have vary widely.[52]

IV. IMPLICATIONS: LEGAL AND POLICY LESSONS FOR AUTONOMOUS DRIVING TECHNOLOGY

Even in this early stage, the Postal Service's exploration of autonomous driving technology contains lessons for successful development of government AI applications. Many of these implications flow from the distinctive nature of autonomous vehicles as a physical, public-facing manifestation of AI.

There is not yet a comprehensive federal regulatory framework for autonomous vehicles, so the full legal implications of the technology remain unclear. Two bills stalled in the Senate last session,[53] but legislators are currently working on a new bill — the Senate Commerce and House Energy and Commerce committees circulated a joint discussion draft in October 2019.[54] The Senate Commerce Committee held a hearing on autonomous vehicle regulation in November 2019.[55] State legislatures have also been active on the topic.[56] So far, the primary focus of Congress and state legislatures has been how to regulate the manufacture of autonomous vehicles, rather than their use.

The Postal Service, which operates one of the largest vehicle fleets in the country and delivers nationwide, must engage with this debate over reshaping tort liability regimes.

The most immediate legal implication relates to the Postal Service's tort liability for vehicle accidents. Although automated vehicles are predicted to reduce accident frequency, they will not eliminate collisions. There is debate about how liability should be determined for collisions involving automated vehicles, particularly in the early stages of deployment.[57] Although the discussion draft of the current federal bill does not address liability,[58] its predecessor bills both contained explicit provisions that left common law liability undisturbed.[59] Some commentators advocate shifting liability from the individual operator to manufacturers or suppliers as an extension of products liability law. Relatedly, some have advocated tying tort liability to compliance with federal regulatory standards instead of state law.[60] The Postal Service, which operates one of the largest vehicle fleets in the country and delivers nationwide, must engage with this debate over reshaping tort liability regimes, including perhaps to advocate for adoption of federal standards.

The second potential legal implication concerns data ownership and privacy.[61] Autonomous vehicles collect and analyze enormous amounts of data about the surrounding environment, including about nearby vehicles and even pedestrians. As with liability, the Postal Service's legal obligations with respect to data collected by its autonomous vehicles will vary depending on how regulations take shape. The current discussion draft of the federal bill does not address data privacy,[62] but its predecessor bills would have ordered studies into how data collected by autonomous vehicles should be processed and safeguarded[63] and would have required manufacturers to disclose both the types of information their systems collect and how that information is used.[64] Some have called on Congress to include a provision granting exclusive ownership over any data collected by an autonomous vehicle to the vehicle's owner.[65] At least one state legislature has considered a similar data ownership provision.[66]

In addition to any obligations deriving from autonomous vehicle regulations, the Postal Service may have obligations under the Privacy Act of 1974[67] for data collected by its autonomous vehicles.[68] Since the Privacy Act applies broadly to "information about any *individual*" that is maintained collected, used, or disseminated by a government agency,[69] its application depends on the extent to which Postal Service autonomous vehicles collect data that is *individually*

identifiable, particularly individualized location information. This, in turn, depends on the deployed technology: Systems that use lidar, for example, or other sensors that capture the relative proximity of a vehicle, pedestrian, etc., but do not capture any identifying information (such as faces or license plates) are less likely to trigger the Privacy Act.[70] By contrast, systems that use cameras or other sensors to collect information about individual vehicles or people are more likely to qualify as collecting "individual" information subject to Privacy Act obligations.[71] Autonomous vehicle systems that communicate directly with adjacent vehicles may also trigger Privacy Act obligations.[72]

In addition to legal implications, there are a number of policy and practical implications to deploying autonomous vehicles. The agency has explicitly addressed the potential benefits and costs of deploying autonomous vehicle technology in a comprehensive OIG report.[73] On the benefits side, the report highlights improved safety and lower accident rates projected to accompany autonomous technology generally,[74] which, in turn, would save the Postal Service money on both tort liability and vehicle repairs. The agency anticipates cost savings on fuel as well, given projections that autonomous vehicles will generally decrease traffic congestion (and thus improve fuel economy).[75] The Postal Service also anticipates autonomous driving applications will increase delivery carriers' labor productivity by freeing up carriers to sort mail and do other tasks while the vehicle drives between addresses, and thus decrease expensive overtime.[76] Similarly, using autonomous vehicles for long-haul trucking would likely ease contract expenses that come with the truck driver shortage, since autonomous shipping would decrease (or potentially eliminate) the need for such drivers.[77] Finally, the OIG report suggests autonomous vehicles would be good for the agency's brand and likelihood of being "viewed as an innovative company."[78]

The potential benefits of autonomous vehicles are closely tied to their potential downsides. First, as the Postal Service recognized in the OIG report, are the long-term labor implications. Although the agency's current plan is to use AI to assist human mail carriers, it is not difficult to imagine the next step of eliminating mail carriers entirely, particularly if technology becomes sophisticated enough to eliminate the need for human drivers. This possibility poses both political and practical difficulties for deployment. Employees and their labor unions may be hesitant to accept the technology even at an early stage, and customers may not like the possibility of fully autonomous postal services. As the OIG report summarized, "a machine does not easily replace the institutional knowledge, judgement, and human contact that carriers can provide." The negative labor impact on contracted truck drivers is likely to be more immediate.

* * * *

In sum, autonomous vehicles will remake not only American roads, but also the day-to-day work of the administrative state. Done well, the advent of autonomous vehicles will make the work of agencies like the U.S. Postal Service both safer and more efficient. But, if rolled out poorly, this new technology threatens to displace the labor force, exacerbate ongoing data privacy concerns, and collide with existing legal regimes. Careful attention to these risks is necessary as federal agencies and the private sector plow ahead on developing and deploying autonomous vehicle technology.

Part III. Implications and Recommendations

Part I offered an overview of how 142 of the most significant federal agencies are currently using AI/ML. Part II dove deeper into a select set of use cases, highlighting the many complexities that attend AI adoption by the administrative state. This Part steps back and, cutting across the full set of use cases, focuses on the policy and legal issues raised by agency use of AI tools. More specifically, we describe the challenges that lie ahead for agencies seeking to develop and deploy AI/ML tools, and where possible, recommend how to mitigate them. We focus on six major implications: (1) the challenges of building AI capacity in the public sector, including data infrastructure, human capital, and regulatory barriers; (2) the difficulties inherent in promoting transparency and accountability; (3) the potential for unwanted bias and disparate impact; (4) potential risks to hearing rights and due process; (5) risks and responses associated with gaming and adversarial learning; and (6) the role of contracting for supplementing agency technical expertise and capacity.[1]

Building Internal Capacity

No agency can effectively deliver on its mission without access to the people, infrastructure, and organizational resources necessary to understand and respond to its environment. As policymakers and civil servants increasingly seek to rely on AI and algorithmic governance, a core challenge for agencies is how to generate the necessary technical capacity—the ability to identify, develop, responsibly use, and maintain complex technical solutions. Our report suggests that building internal capacity, rather than simply embracing a default practice of contracting out for technical capacity, will be crucial to realizing algorithmic governance's promise and avoiding its perils.

The literature on government capacity building generally boils down to the "make-or-buy" decision.[2] An agency can *make* the goods and services needed to perform governance tasks by hiring personnel and building its own infrastructure, or it can *buy* them through the procurement process.[3] In theory, the private sector has greater expertise and can produce at lower cost.[4] In practice, however, procurement has downsides. For "hard" or "commodity" goods and services, where quality is easily measured and tasks involve little discretion, government can fully capture private sector expertise and efficiencies. By contrast, "soft" or "custom" ones—where monitoring quality is more difficult and tasks involve more discretion—invite corner-cutting by contractors that degrade quality.[5] Contracting out makes more sense for police cars than for police officers.

Realizing the full potential of algorithmic governance tools will thus often require internal capacity.

These generalizations are coarse but useful. Certain components of algorithmic governance tools appear suitable for procurement. Upgrading computer systems and consolidating databases, for instance, are more likely to be standard services.[6] In other ways, however, AI poses heightened capacity-building challenges for agencies. For example, a private sector contractor's software engineers often will not have a nuanced understanding of the problems a given algorithm tool is aimed to address or the legal, regulatory, and organizational environment within which the tool will operate. Realizing the full potential of algorithmic governance tools will thus often require internal capacity. We describe the potential pitfalls of relying on external contractors, and explicitly compare the pros and cons of internal and external sourcing of AI tools, in a separate section below on "External Sourcing."

Focusing on internal capacity building, we here address four main considerations. First, agencies will need to invest in their technical and data infrastructure. In most cases this will require not only hardware and software upgrades but also collecting, standardizing, and securing the data required to deploy AI tools. Second, agencies will need to cultivate in-house human capital to produce AI tools that are not only usable at the technical level but also compliant at the legal and policy levels. Third, agencies will need to invest in comprehensive and flexible AI strategies that allow agencies to learn strategically from failures and evolve. For agencies developing their own AI tools, this means creating iterative development and evaluation processes with clear success metrics. For agencies that regulate private sector AI, these strategies may include regulatory "sandboxes" to develop and enforce standards not just for present applications but also for future ones. Finally, in-house design and deployment can enhance public accountability and transparency.

I. BUILDING TECHNICAL INFRASTRUCTURE AND DATA CAPACITY

Because AI tools require complex software packages and computing power to process large datasets, agencies may have to upgrade legacy systems or integrate new systems

with old ones.[7] This is a challenge for agencies that excel, as one agency official facetiously put it, at "having the latest technology of the last decade."[8] By one estimate, SSA has over 14 petabytes of data, but data is stored in roughly 200 separate databases. Linking, cleaning, and merging such data remains an ongoing process. Most of SSA's supporting applications remain written in outdated (COBOL) programming language, stemming from initial development some 30 years ago. SSA is in a process of updating these applications into more modern languages, but such modernization is resource intensive, requiring, for instance, personnel trained in different generations of languages.[9]

Since all AI tools—whether supervised or unsupervised—are data-hungry, agencies must also invest in the necessary input data. Investing in data capacity requires addressing the interrelated challenges of data *collection*, data *standardization*, and data *security*.

A. Data collection

Deploying AI tools requires collecting the right data and enough of it. But before collecting data at scale, agencies may need to clarify their statutory and regulatory authority. A far-flung statutory fabric, including constitutional provisions, federal, state, and local laws, defines government duties and obligations around data and includes transparency statutes such as the federal Freedom of Information Act and state law equivalents. At the federal level, the Privacy Act and amendments provide the closest to a comprehensive scheme for information practices.[10] Among other things, agencies must, where possible, obtain data from individuals and may not use data for secondary purposes without consent.[11] The law also significantly constrains the government's ability to knit together datasets across agencies.[12] Other pillars of the federal regime include the Paperwork Reduction Act, which constrains an agency's ability to collect new data from the public,[13] and the Information Quality Act, which constrains agencies' ability to open-source data holdings to achieve transparency.[14]

Agencies may also face data collection limitations due to lack of specific authority. For example, NHTSA's enforcement and vehicle safety research divisions seek to use AI/ML to model historical crash data for simulated testing of automated vehicles. But NHTSA may currently lack authority to compel manufacturers to produce crash data.[15] The agency's voluntary data collection mechanism[16] captures only a fraction of the vast data that manufacturers generate.[17]

Data collection poses related logistical challenges for agencies that rely on third-party data. Third-party data may be hard to obtain, incomplete, or unrepresentative due to selective or inaccurate reporting.[18] For example, pharmaceutical companies may not want to share the most comprehensive clinical trial data on which the FDA could train its AI/ML.[19] At present, agencies like NHTSA and the FDA are encouraging third parties to voluntarily provide data.

B. Data standardization

To be of any use, data must be in an appropriate format. Different types of AI tools require different levels of data standardization, but standardization can pose significant barriers to virtually any AI deployment. As detailed in Part II, an alternative path at the FDA is to defer the agency's NLP projects until it can obtain standardized, fit-for-purpose data. Some standardization issues arise from the data storage or submission medium. The IRS, for example, continues to process paper-filed tax returns that often contain missing information.[20] Even digital data may not be standardized. The SSA, as another example, processes unstructured digital text, such as paragraphs describing disability circumstances and non-uniform medical records maintained in PDF files. The SEC, too, struggles to compare companies in its centralized CIRA system because companies can use varying semantic tags or use incorrect tags.[21] Many agencies face a trade-off between data depth and uniformity.[22] In addressing data standardization challenges, agencies must consider which data they are willing to standardize and at what stage: at the collection phase—by "outsourcing" standardization to regulated entities—or at the processing phase by developing advanced tools that can standardize unstructured data.

C. Data security

Data often comes with security requirements. The Federal Information Security Management Act, for example, requires agencies to develop data security programs, breach notification policies, and disposal routines,[23] and then subjects them to civil suits for failures.[24] Building data capacity requires agencies to address these requirements, typically by developing strict internal guidelines for the use and sharing of data that contains personal information. Agencies should also leverage technology to reconcile data sharing needs with data privacy concerns. Researchers at the Department of Veterans Affairs, for example, used cryptographic hashes to obscure lab results and other sensitive data in its partnership with Alphabet's DeepMind unit.[25] An official at the IRS similarly

proposed employing Generative Adversarial Networks (GANs),[26] and the CFTC proposed anonymizing data to enable collaboration with market participants.[27]

II. BUILDING INTERNAL STAFF CAPACITY

An agency's AI tools must be both usable and compliant. As to usability, optimal design and deployment will often depend on a deep understanding of the problem an algorithmic tool seeks to solve, an ability to convince skeptical agency staff to utilize the tool, and a user-friendly interface that eases that pitch. And as to compliance, algorithmic tools themselves encode legal and policy choices, some of which will be subject to judicial review.[28] Software engineers, especially those outside the agency, may lack the insights or training necessary to faithfully translate law into code. While in-house production may strain project budgets and introduce recruitment challenges, building internal staff capacity may yield tools that are better tailored to the relevant task and legal requirements.

Several agencies have already demonstrated the value of embedded expertise. As detailed in Part II, the SSA developed NLP tools to identify potential errors in draft disability determinations as a result of a multi-year strategy to hire and then repurpose lawyers with technical skillsets.[29] This strategy helped facilitate an iterative design process in which system architects could readily work back and forth between code choices and legal, policy, and organizational considerations. The Internal Revenue Service's (IRS) development of algorithmic enforcement tools similarly illustrate the value of in-house, embedded expertise in automating tasks that are inherently dynamic. As Part II's case study of the SEC noted, enforcement agencies must engage in continuous, iterative updating of their AI tools as enforcers unearth new modes of wrongdoing.[30]

Federal agencies seeking to build internal technical capacity must grapple with budgetary and other human resource constraints. In addition to overall budget caps, civil service laws capping allowable salaries can price government agencies out of the technical labor markets. Agencies can offer job stability and work-life balance, whereas technology companies incentivize talent by offering competitive salaries or stock options. The Competitive Service hiring process also constrains recruitment,[31] although the Office of Personnel Management (OPM) has taken steps to ease hiring burdens for technical positions, including by establishing a "data scientist" classification.[32] OPM also recently established a government-wide "direct hire" appointing authority for a variety of STEM positions and for all IT positions for agencies that can demonstrate "the existence of a severe shortage of candidates or critical hiring need."[33]

III. INVESTING IN AI STRATEGY AND "SANDBOXES"

Deploying AI technology requires agencies to invest in comprehensive strategies to test, evaluate, update, and retire AI tools. An important part of this strategic process is articulating metrics for measuring the success of innovations that align with the agency's risk-profile and level of comfort with failure. Agencies should also develop testing "sandboxes" that allow for failure and iterative evaluation of new governance tools and, for agencies that regulate private-sector AI, a testing infrastructure that can help guide regulated entities.

A. Evaluation metrics

In advance of deployment, agencies should develop metrics for measuring the success of AI tools. These evaluation metrics should be tied to the agency's broader mission, rather than focused purely on efficiency or return on investment.[34] Correspondingly, agencies should establish a process for "returning to the drawing board" when tools fail to satisfy these metrics. Given the dynamic nature of AI/ML models, these metrics should also guide subsequent evaluations and decisions about when to refine or retire a given tool.[35] Front-line enforcers may provide ongoing feedback on models.[36]

B. "Sandbox" testing and regulatory infrastructures

To build technical capacity, agencies will likely have to develop a comfort level with technological failure—and this may be easier for some agencies than for others.[37] Agencies like the FDA must maintain a relatively low risk tolerance: Failing to detect adverse postmarket effects of a pharmaceutical can have critical public health consequences. By contrast, the IRS has continued to experiment with technology despite low accuracy rates.[38]

Risk-taking is crucial to developing successful tools, and agencies seeking to employ AI must be willing to fail.[39] As many agencies found, initial efforts and failures create a "supersized sandbox"—a playground for developing future AI applications and learning important lessons. Agencies should structure projects to allow some margin of error and treat failures not as losses but as opportunities to share lessons across the agency.[40] Although it began over twenty

Building internal expertise and technical capacity may also be essential to accountability and building trust.

years ago, the IRS's Compliance Data Warehouse established a foundation that is enabling the agency to consider more complex AI applications moving forward.[41]

Similarly, agencies that regulate AI deployments in the private sector should also build regulatory "sandboxes." For example, at the FDA, "INFORMED has created a unique sandbox for networking, ideation and sharing of technical and organizational resources, empowering project teams with the tools needed to succeed in developing novel data science solutions."[42] These sandboxes, moreover, can signal minimum standards for AI and help regulated entities "de-risk" their development decisions. The proliferation of guidance and reports can also serve this goal. In the context of cybersecurity, the FDA has provided some guidance on what the agency expects to see in premarket submissions, including certain specific design features and cybersecurity design controls.[43] Recent approval of several AI-included devices along with the release of a discussion paper on its plans to regulate AI/ML-based software as a medical device,[44] serve to provide additional guidance to manufacturers. Further, the FDA, in conjunction with MITRE, released a report entitled Medical Device Cybersecurity: Regional Incident Preparedness and Response Playbook.[45] The FDA explained that the report can serve "as a customizable tool for health care delivery organizations to aid in their preparedness and response activities for medical device cyber incidents."[46]

IV. LINKING CAPACITY TO ACCOUNTABILITY

Building internal expertise and technical capacity may also be essential to accountability and building trust.[47] The scholarly literature may be moving away from individual, privately enforced rights as the best way to achieve accountability in favor of "accountability by design."[48] Kroll et al. offer a catalog of tools that engineers can incorporate into algorithmic systems to facilitate evaluation and testing.[49] This "accountability by design" trend links to longstanding calls among administrative law scholars for agencies to develop an "internal law of administration" distinct from—and often more effective than—externally imposed accountability.[50] However, some agencies are more likely than others to incorporate accountability and transparency by design—with agencies such as the FDA and NHTSA being more incentivized, given that both are subject to a high potential of judicial review and public scrutiny.

Transparency and Accountability

Administrative law—the mix of constitutional and statutory law that governs how agencies do their work—is premised on transparency, accountability, and reason-giving.[51] When government takes action that affects rights, it must explain why. Yet many of the algorithmic tools that federal agencies use to make and support public decisions are not, by their structure, fully explainable.[52] The challenge is how to craft concrete legal and regulatory mechanisms for algorithmic tools that meaningfully fulfill transparency values and ensure fidelity to the agency's legislative mandate and other legal commitments (*e.g.,* non-arbitrariness, non-discrimination, privacy).

I. BRIDGING TRANSPARENCY AND ACCOUNTABILITY

Subjecting algorithmic decision systems to meaningful accountability poses two main challenges: achieving transparency into a tool's workings, and then selecting the best regulatory mechanism for translating that information into desired compliance.

The gold standard of transparency in any decision-making context is a full account of a decision's "provenance," including its inputs, outputs, and the main factors that drove it.[53] The problem, as Part II noted, is that machine learning models are often inscrutable. Even a system's engineers may not understand how it arrived at a particular result or be able to isolate the data features that drove the model's prediction. Algorithmic outputs are also often nonintuitive in that the data relationships they surface may not map to any common-sense understanding of how the world works. Even full disclosure of a system's source code and data and an opportunity to observe its operation "in the wild" will not necessarily facilitate either insight or accountability.[54]

Two approaches to transparency have begun to emerge in response to these concerns. One camp focuses on how to mix modes of explanation to achieve desired transparency. For instance, an incomplete accounting of a particular decision can be supplemented by a "system-level" accounting of the tool that made it,[55] including data descriptions,[56] modeling choices,[57] and general descriptions of factors that drive the model's predictions.[58] A second camp advocates simplification of models to make them more parseable.[59] These measures might take the form of a ceiling on the number of data features used or outright bans on particular tools (*e.g.,* facial recognition) or particular models, such as powerful "deep learning" techniques that generate more accurate predictions but are often less interpretable.[60]

Even where AI systems can be made transparent, there remains the challenge of choosing regulatory mechanisms that can translate that transparency into meaningful accountability. Here regulatory architects have numerous options. They can choose mechanisms that promote legal accountability (*e.g.,* judicial review of agency action) or political accountability (*e.g.,* public ventilation through notice and comment or mandatory agency-conducted "impact assessments"[61]). They can also opt for "hard" rules (*e.g.,* prohibitions on certain models, a licensing or certification requirement prior to use akin to FDA drug approvals, or liability rules that allow injured parties to recover damages), "soft" rules (*e.g.,* impact assessments designed to ventilate concerns about algorithmic tools but confer no substantive rights),[62] or something in between (*e.g.,* notice, consent, correction, and erasure rights like those given data subjects in the European Union's General Data Protection Regulation[63] or the U.S. Fair Credit Reporting Act[64]). If hard rules are chosen, regulatory designers can choose to delegate enforcement authority to public enforcers, including, as some advocate, an "FDA for AI,"[65] or to private enforcers deputized to sue in court or incentivized via whistleblower bounty schemes.[66] Finally, regulatory architects can opt for *ex ante* regulation before a model runs—think once again of an FDA-style pre-certification scheme or prohibitions on uses or model types—or *ex post* regulation of results, as with lawsuits seeking damages.[67]

II. DESIGN PRINCIPLES

For the moment, no single best solution from this menu of options has emerged. However, Part II's in-depth case studies, by showcasing a wide range of AI-based governance tools, help establish some working premises that can frame the possibilities and limits of competing approaches.

First, consideration of actual use cases reveals hard trade-offs between accountability and efficacy. Imposing constraints on model choices—by, for example, limiting the number of data features or prohibiting more sophisticated modeling approaches—trades off interpretability against a tool's analytic power and, thus, its usefulness.[68] As just one example, requiring the SEC to deploy a less sophisticated but more interpretable algorithmic tool in making enforcement decisions may make it easier for regulated parties or agency overseers to evaluate the tool's workings but may also bring substantial costs, subjecting regulated parties to undue prosecutions and wasting scarce agency resources in the process. Here and elsewhere, interpretability may come only at the cost of efficacy.[69]

Second, the pros and cons of transparency will often vary by governance task and the rights and interests at issue. In the enforcement context, public disclosure of the "internals" of an algorithmic enforcement tool can impair or defeat the tool's utility by facilitating evasion and gaming by regulated parties—an issue we explore in more detail later in Part III's section on "Adversarial Learning." In certain mass adjudicatory contexts, by contrast, full open-sourcing of algorithmic tools might make sense as an accountability measure. One might conclude, for instance, that disability or veterans' benefits determinations are too important to risk erroneous determinations and, in any event, present a lower risk of gaming by beneficiaries.

Third, efforts to build effective accountability systems will have to reckon with the existing structure of administrative law. To date, much academic debate has focused, at a high level of abstraction, on procedural due process under the Fifth and Fourteenth Amendments to the United States Constitution.[70] Far less work explores the more fine-grained statutory requirements of administrative law and, even then, offers mostly a surface-level tour of potentially applicable doctrines.[71] This is problematic because the doctrine of constitutional avoidance—which holds that courts should avoid ruling on constitutional issues in favor of other grounds—means that much, or even most, of the hard work of regulating algorithmic governance tools will come not in the constitutional clouds but rather in the streets of administrative law.[72]

So-called "reviewability" doctrines in administrative law offer a compelling example. Current administrative law, as exemplified by the Supreme Court's *Heckler v. Chaney* decision, insulates agency enforcement decisions from judicial review except where Congress has clearly specified a standard for the agency's exercise of discretion or where an agency has wholly "abdicated" its enforcement duties.[73] The reasons are many, but the main anxiety is about judges' ability to reconstruct or evaluate specific enforcement decisions, which often rest on subtle judgments about how best to allocate scarce agency resources.

Interestingly, algorithmic enforcement tools may make these reviewability concerns worse or better. On one hand, the black box nature of machine learning tools may further obscure agency enforcement decisions, strengthening the rationale for hiving off those decisions from judicial review. Something very near the opposite, however, may also result. By allowing agencies to formalize and make explicit organizational priorities, algorithmic tools have the potential to render enforcement decision-making somewhat more tractable than the dispersed human judgments of enforcement staff. For instance, if appropriately balanced with the need for a degree of confidentiality of agency enforcement goals, code may help provide the missing "focal point" for judicial evaluation of agency enforcement decisions and rebut the current doctrine's presumption against reviewability. Moreover, because algorithms encode legal principles and agency priorities, they perform regulatory work and so may qualify as "rules" under administrative law, thus requiring mandatory ventilation via the notice-and-comment process or exposing them to pre-enforcement judicial review. The counter-intuitive result is that continued proliferation of algorithmic enforcement tools may, on net, yield an enforcement apparatus that is more transparent and less opaque than the current system.[74]

Reviewability only scratches the surface of ways that the administrative law will modulate federal agency use of AI tools. Administrative law may also need to adapt in determining whether agency decisions made or supported by an algorithmic tool are "arbitrary and capricious." Courts will

thus grapple once more with whether such review is a matter of light-touch review[75] or deeper "hard look" review.[76] And, as we explore in more detail elsewhere in Part III, agency use of AI-based tools to support adjudication raises distinct legal questions relating to hearing rights and due process.

Fourth, looking across concrete use cases underscores administrative law's potential limits in achieving algorithmic accountability. Meaningful accountability must be built upon *actionable* transparency. It does little good to give judges transparency into an algorithmic system's "internals" if they lack the technical understanding necessary to make sense of it. The same is true of ordinary citizens who are the objects of algorithmic decisions. If engineers cannot understand a system's outputs, then there is little reason to believe that less technically trained actors can do any better.

Actionable transparency can also falter when data and algorithms change dynamically. For instance, the SEC's supervised learning model for Form ADV disclosures is trained on past referrals to the SEC's enforcement arm, but the pool of referrals grows over time, with different human input for each referral. This means that each model may be distinct. A model reviewed at one stage (during the notice-and-comment process) may already be substantively different upon deployment. Conversely, problematic predictions at one point (a specific enforcement decision) might vanish as the model is updated. By their nature, the notice-and-comment process and APA-type judicial proceedings are static and may not generate the information required to understand an algorithm in action.

Finally, administrative law works in tandem with an array of data and disclosure laws that, at least in their current form, can sharply limit transparency. In the SSA context, individual data is protected under the Privacy Act of 1974.[77] Similarly, the raw disclosures that serve as inputs for the SEC's enforcement tools are publicly available, but data from prior investigations—that is, the filings that triggered elevated review—are likely protected under the Freedom of Information Act's exemption for law enforcement purposes.[78] Finally, a contractor-provided algorithmic tool's technical guts may be protected by patent, copyright, or trade secrecy laws,[79] and government use of the tool provides no further right of disclosure.[80]

III. CONCRETE REFORM IDEAS

Given these challenges, judges, agency administrators, and legislators will face difficult questions about whether to retrofit existing accountability frameworks or mint new ones.

A minimalist option would retrofit or, to the extent feasible, reinterpret the APA to enable prudent *ex ante* review of algorithmic tools through the notice-and-comment process and/or judicious *ex post* review by courts. On the latter, *ex post* side, Congress or courts may wish to relax the presumption against reviewability of enforcement decisions under *Heckler v. Chaney*.[81] On the *ex ante* side, an amended APA could set new triggers for when an algorithmic tool is subject to notice and comment. One could peg notice and comment to whether staff use of the tool is mandatory or voluntary as a crude proxy for how much the tool displaces human discretion. A more technical approach would key notice and comment to the numerical threshold the tool establishes. For example, an enforcement tool that flags potential violators as "high risk" necessarily sets a probability threshold from 0 to 1. The higher the threshold, the greater the risk that human discretion is displaced.[82] The chosen threshold also fixes the relative number of false negatives and false positives to be expected. As a result, the choice of threshold cannot be made without weighing the social costs of each type of error—precisely where public participation via notice and comment may be most useful.

Given the limitations of *ex ante* and *ex post* review under the APA, a more comprehensive institutional solution would be to create an AI oversight board, either within each agency or as a freestanding agency with oversight over all other agencies. Staffed with technologists, lawyers, and agency representatives, an oversight board could be tasked with monitoring, investigating, and recommending adjustments to agency adoption and use of AI.[83]

A third possibility would be to require agencies to engage in prospective "benchmarking"[84]—that is, to create random hold-out sets to compare AI-assisted outcomes and human (status quo) decision-making. In the SSA context, for instance, the Insight system could be deactivated for a random hold-out set and each case adjudicated in analog fashion. In the SEC context, investigators could be required to investigate a subset of cases without the aid of risk scores. Benchmarking would provide a practical test of a tool's facial validity, smoking out bias and arbitrariness and enabling agencies, courts, and the public to meaningfully assess the impact of AI use cases. Benchmarking would also generate new training data and provide a check on procurement-provided tools.

IV. LINKING ACCOUNTABILITY TO CAPACITY

While these are promising reforms, it is worth noting that formal accountability frameworks are not the only way to ensure responsible agency use of algorithmic governance tools. Internal agency supervision and embedded expertise can also be a powerful source of accountability. As noted in the previous section on capacity building, embedded expertise facilitates "accountability by design" in which agency technologists proactively design and maintain systems that are more transparent and auditable and less arbitrary and biased not as a response to legal or other external threats, but as a matter of good government, good engineering, and professional ethics.

Bias, Disparate Treatment, and Disparate Impact

The administrative state's growing adoption of AI tools risks compounding biases against vulnerable groups. If biases go unchecked, agency tools will only deepen existing inequities and also likely run afoul of antidiscrimination law. Yet, many proposed solutions to combat bias would themselves violate other core legal principles, such as equal protection. In short, agencies can find themselves in a bind. Given these challenges, it is critical that agency administrators, legislators, judges, and academics devote more attention to developing agency-level mechanisms to detect, monitor, and correct for bias, as well as appropriate legal regimes to govern them.

I. EMERGING EVIDENCE OF BIAS

It is well-documented that AI tools have the potential to exacerbate bias against vulnerable groups. Three lessons have emerged from a rapidly developing literature on fairness and machine learning. First, the potential for machine learning to encode bias is significant.[85] Criminal risk assessment scores, for instance, may exhibit higher "false positive rates" (wrongly classifying individuals as "high risk") for African-American than white individuals.[86] An NLP-based engine for job applicants may score applicants who graduated from women's colleges more poorly, because of the existing demographic composition of the work force.[87] Second, while many potential approaches to "fair machine learning" have been proposed, a basic challenge is that divergent notions of fairness can be mutually incompatible.[88] In the presence of underlying differences between demographic groups, for instance, it is not possible to simultaneously equalize false positive rates, false negative rates, and predictive parity across groups. Third, critical questions remain as to how AI-assisted decisions fare compared to human decisions, given that human decisions are themselves often the origin of bias.[89]

II. THE POTENTIAL FOR BIAS IN USE CASES BY THE ADMINISTRATIVE STATE

Our case studies corroborate this risk across the administrative state. The sources of such bias can be varied. First, training data may be unrepresentative of the population of interest. Facial recognition technology that has been trained disproportionately on lighter skin tones, for instance, may be significantly less accurate for darker skinned individuals,[90] potentially introducing bias into CBP's reliance on facial recognition. Second, a number of use cases rely on linking different administrative datasets together, and coverage may skew toward certain demographic groups.

Formal blindness can be functional discrimination.

The Internal Revenue Service, for instance, developed a Return Revenue Program (RRP) to detect fraudulent refunds. This RRP program uses a wide range of sources, including data from the Federal Bureau of Prisons and prison systems in all states.[91] Such record linkage poses a risk of disparate impact on subgroups, although it remains hard to assess in the abstract. To illustrate, consider a similar setting that has been subject to more examination. The Allegheny Family Screening Tool for child welfare relies extensively on record linkage of administrative data from means-tested programs. Eubanks argues that the system hence relies on data for the poor that it does not observe for the wealthy (*e.g.,* private drug treatment, mental health counseling). The effect is that it disproportionately rates the poor as "high risk" of child welfare placements.[92] Third, some systems may simply replicate existing bias in human decisions. If agencies used a predictive model for which comments are likely relevant, for instance, such decisions may simply encode existing agency tendencies

to rely on lengthier documents, written in non-vernacular, submitted by legal counsel.[93] In the PTO's prior art search, a machine learning model trained on historical labeled data may replicate the tendency by patent examiners to neglect non-patent literature.[94]

The rise of AI decision tools will increasingly challenge conventional principles of antidiscrimination law.

III. THE WAY FORWARD

Grappling with such forms of bias will be a significant undertaking for federal agencies adopting AI/ML. First, the emerging consensus within machine learning is that "blinding" algorithms to protected characteristics is unlikely to be effective. As the feature set (*i.e.,* the number of variables in the model) grows, protected characteristics, such as race and gender, can be inferred with extremely high probability.[95] Formal blindness can be functional discrimination. Researchers have hence argued that "fairness-through-awareness," not blinding, will be a more promising approach to ensure fairness.[96] Yet because there are no consensus measures for fairness, government agencies will have to increasingly engage with evolving standards and methods for assessing the potential for bias in machine learning and such judgments may be highly domain-specific.

Second, the rise of AI decision tools will increasingly challenge conventional principles of antidiscrimination law. As noted, protected characteristics can be inferred with high likelihood as the feature set (of unprotected characteristics) grows. This challenges the anticlassification principle, which posits that the law should not classify individuals based on protected attributes (*e.g.,* gender and race). Similarly, the rise of AI/ML tools will test doctrinal frameworks of narrow tailoring and "individualized" consideration under the Equal Protection Clause. The Supreme Court has not clarified the operation of those principles specifically in the context of machine learning, but its affirmative action cases illustrate the tension. In the affirmative action context, the Supreme Court held that the University of Michigan law school's consideration of race in "individualized" admissions was constitutional,[97] but held that the practice of awarding 20 points on a 150-point scale for underrepresented minorities in undergraduate admissions violated equal protection.[98] "[I]ndividualized consideration," the Court noted, "demands that race be used in a flexible, nonmechanical way."[99]

Machine learning, however, challenges this doctrinal distinction. Is an algorithm that uses 1,000 features, including a protected attribute, "individualized" or is it "mechanical"? Is the mere use of the point scale problematic, or is it about the relative weight of protected characteristics? In *L.A. Water & Power v. Manhart*, the Supreme Court found that the use of gender in calculating pension plan contributions violated equal protection, despite the actuarial gender difference in longevity.[100] In *State v. Loomis*, the Wisconsin Supreme Court did not find a due process violation when gender was used in a criminal risk assessment score, finding that the "use of gender promotes accuracy that ultimately inures to the benefit of the justice system."[101] Due to the doctrinal uncertainty, states and localities using criminal risk assessment scores remain split in whether they rely on gender.[102] To the extent that the machine learning literature calls for awareness of protected attributes to promote fairness, it is on a collision course with equal protection doctrine.

Even if an algorithm passes constitutional muster, it is unclear how administrative law will grapple with claims of disparate impact. Litigants may claim that the adoption of an algorithmic decision tool causes disparate impact across demographic groups and that the failure to address and explain such consequences is arbitrary and capricious. Yet whether courts will entertain such claims and how courts weigh the fairness-accuracy trade-off remains an open question. The D.C. Circuit, for instance, has held that disparate impact arguments may not be brought under the APA when Title VI of the Civil Rights Act—then assumed to provide a private right of action—provides an alternative adequate remedy.[103] Since that decision, the Supreme Court held that there was no private right of action under Title VI, but no court has explicitly considered whether that opens the door to disparate impact claims under the APA. Mounting evidence of the potential for disparate impact with AI decision tools will put pressure on courts to grapple with this gap.[104]

Third, no agency examined in this report has established systematic protocols for assessing the potential for an AI tool to encode bias. While some application areas (*e.g.,* facial recognition) present obvious risks, the need for such

protocols may be even greater for use cases where bias is less obvious. Might FDA's adverse event reporting system be driven by reporting bias along demographic groups, say due to differences in access to health care?[105] If SSA builds out its expedited review program using electronic health records, does that advantage certain types of applicants who are more likely to have access to health care providers with interoperable electronic health record systems? Would neural network models deployed by the PTO actually fail to capture temporal drift and, as a result, disadvantage pathbreaking research by smaller entrepeneurs?[106] The upshot here, as earlier, is that developing internal capacity to rigorously evaluate, monitor, and assess the potential for disparate impact will be critical for trustworthy deployment of AI in federal administrative agencies.[107]

In sum, the rise of algorithmic decision-making raises novel and important questions about disparate impact. Fortunately, administrators, technologists, legislators, and judges can draw from the rapidly emerging literature on bias in machine learning to proactively assess the potential for bias. Efforts will need to be focused on developing the appropriate institutional mechanisms for detecting, monitoring, and correcting for bias in AI decision tools.

Hearing Rights and Algorithmic Governance

Much of the decision-making of modern administrative government comes after a "hearing." Such hearings provide affected parties the opportunity to submit evidence in-person or on paper to a decision-maker, often an administrative judge or other agency employee. An array of laws dictates the form these hearings take, among them the default requirements of the APA, agency enabling acts, agency regulations, and the Constitution's Due Process Clause. In some contexts, the procedural bundle is meager: An agency need only rise above the floor set by due process by providing advance notice of the decision and a brief opportunity to be heard.[108] In others, constitutional and statutory mandates require an administrative approximation of a full-dress trial, with rules dictating who can participate, the types of evidence that can be considered, record-keeping requirements, appeal rights, and restrictions on ex parte contacts.[109] How will the rise of AI decision tools alter the form and function of these hearings and how should administrative law adapt in response? The role of hearing rights cuts across adjudicatory contexts, from formal adjudication at the SSA to more informal patent decisions at the PTO and enforcement decisions at the SEC, IRS, or CMS.

This section makes three points about the future of hearing rights in the face of the AI revolution. First, while the most optimistic version of AI tools may improve accuracy and efficiency of adjudicatory decisions, such tools may also expose trade-offs in normative values underpinning hearing rights. Second, we articulate how procedural due process and statutory hearing rights may need to adapt if AI tools proliferate. A core challenge in the near-term will be crafting legal and institutional vehicles to detect and address systemic sources of error in light on the current structure of individualized decision-making. Third, the rise of AI tools in adjudication potentially raises longer-term, foundational questions: Do due process and statutory hearing rights imply a right to a human decision-maker? And what role is left for hearing rights in a world in which legal and regulatory mandates are crafted, adjudicated, and enforced with increasingly limited human involvement?

The promise of AI is that it may cut the Gordian knot of this accuracy/efficiency trade-off by making possible efficiency gains without reductions in accuracy.

I. THE PROMISE AND PERIL OF ALGORITHMIC ADJUDICATION

Conventional wisdom holds that due process poses an accuracy-efficiency trade-off. Adding procedures can improve a decision's accuracy by ensuring close consideration of a wider range of evidence and subjecting arguments and evidence to more robust and often adversarial testing. But process is also costly. Importantly, procedure's costs are both social (*e.g.,* the resources required to operate the system) and individual (*e.g.,* the costs to parties in real resources

and delay). On the latter, one need look no further than the significant backlogs at the SSA and PTO, which can delay desperately needed disability benefits and innovation-spurring intellectual property protections.

The promise of AI is that it may cut the Gordian knot of this accuracy/efficiency trade-off by making possible efficiency gains without reductions in accuracy and vice versa. Some AI-based decision tools may even yield simultaneous improvements in both, yielding better decisions *and* at lower cost. The SSA's tool for clustering like cases, for instance, potentially enables adjudicators to work through cases more quickly *and* more equitably, improving the consistency of decision making. Similarly, the SSA's "easy grant" identification tool routes easy cases to staff-level decision-makers for rapid resolution so that administrative judges can focus their energies on more difficult and demanding cases. These and other AI-based tools profiled in Part II's case studies might finally crack the code of mass adjudication, improving accuracy while shrinking the inter-judge decision disparities and backlogs that have long plagued a wider range of agencies.

If AI tools are indeed able to solve the quantity-quality trade-off, they may also make room for other adjudicatory values. As noted in the SSA chapter, AI tools might help reclaim a part of constitutional due process that has been in part sidelined in modern jurisprudence: the dignity interests of the parties. By eliminating rote and repetitive tasks, AI might free adjudicators to focus on procedural fairness: to engage parties more extensively, to issue tentative orders, and to explain the complex legal provisions to affected parties. A long line of research establishes that individuals may perceive a process as more legitimate if afforded a voice.[110] Dignitary interests may have value independent of accuracy.[111]

Despite such optimistic glosses, AI-based tools also raise significant concerns. First, AI tools displace adjudicator discretion and independence, potentially draining the system of its deliberative and adaptive capacities.[112] Importantly, displacement of discretion can occur even where manual review nominally remains. One reason is automation bias—*i.e.,* the over-reliance of decision-makers on automated predictions, even when such deference is unreasonable and mistaken.[113] Faced with rigid quotas, patent examiners may be unwilling to expend additional effort to second-guess AI-prioritized search results. Adjudicators at the SSA may review cases solely to pass Insight quality flags, progressively ignoring errors that evade automated detection. Machine predictions might allow an administrative judge to readily compare her inclination to that of others, threatening notions of decisional independence. And algorithmic search tools may diminish engagement with the record, functionally undermining *de novo* review. All of these dynamics can stifle the emergence of exceptions and the dynamic, iterative effort to conform legal mandates to changing circumstances.

A key challenge then is to build decision tools that complement, rather than substitute for, human decision-making— i.e., human-centered AI.

Second, adjudicators may simply ignore AI tools. Such aversion to algorithms erodes the accuracy and efficiency gains of automation, even where human decision-making may be demonstrably inferior.[114] Under-reliance on algorithmic tools may be particularly likely when decision-makers are field experts, as is the case with administrative judges.[115] And because administrative judges may vary in their receptiveness to AI tools and in their willingness to review machine outputs or deviate from recommended results, inter-judge decision disparities and high reversal rates may persist. A key challenge then is to build decision tools that complement, rather than substitute for, human decision-making—*i.e.,* human-centered AI.

Last, algorithmic systems may simply get things wrong, eroding decision quality under a false veneer of efficiency gains. Statutory interpretation and implementation are open-ended and difficult tasks. Algorithmic outputs might deviate from the statutory mandate or prove non-policy-compliant. And, as we describe next, current systems are ill-suited to detecting such sources of systemic error.

II. GETTING HEARING RIGHTS RIGHT

As AI-based decision tools proliferate, how can hearing rights adapt to harness AI's positive potential while mitigating its costs? Administrative hearings come in myriad policy contexts and, as already noted, the procedures that apply in each take

many forms. The optimal mix of procedural rights may vary significantly across contexts. For now, we make several points that can help guide judges, administrators, and legislators in adapting the current system.

First, the current system of hearing rights fits awkwardly with the most pressing challenge raised by algorithmic decision tools: identifying and remedying *systemic* sources of error. Part of the challenge is inherent to the structure of individualized hearing rights. A single judicial challenge to agency decision-making may correct a specific error, but such challenges are unlikely to surface and remediate entrenched pathologies within the system.[116] Specific to the algorithmic context, Danielle Citron argues that there is an additional doctrinal challenge: The Supreme Court's longstanding test for procedural due process, which requires courts to focus on only the case at hand and weigh the private interest, the government interest, and the likely value of additional process, may neglect the fact that algorithmic tools are designed to operate at scale.[117] 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.[118]

Second, AI-based decision tools may progressively scramble the foundational distinction between rulemaking and adjudication under the APA and Constitution. For adjudication, procedural due process and applicable statutes safeguard the interests of a single person or a small group of affected people.[119] For rulemaking, the Constitution requires little and the APA requires only a general level of public participatory engagement when a rule is addressed to a large class of people with common circumstances.[120] In *Heckler v. Campbell*, the Supreme Court affirmed the statutory authority of the SSA to decide common issues in adjudications via rulemaking.[121] Despite the fact that the Social Security Act requires "individualized determinations based on evidence adduced at a hearing," the Court held that the act "does not bar the Secretary from relying on rulemaking to resolve certain classes of issues."[122] As AI systems become more sophisticated, a key question will be when they function as "legislative rules" that have "binding effects" on the agency and regulated parties, triggering notice-and-comment rulemaking. Was SSA required to undergo notice-and-comment for its QDD system? And if so, should it have been required to disclose more of the underlying feature set and model? Answers to these questions are easier for top-down expert-based AI systems (if-then rules). But modern machine learning systems are "bottom-up" in that they construct rules based on learned associations from prior decisions. Whether the system has a binding effect hence depends empirically on (a) the level of adherence to the rule, and (b) the extent to which models prospectively adapt. Such adaptation also makes it more challenging *ex ante* to disclose the nature of the decision system in contrast to a decision tree from an expert-based system.

Third, as technology advances, parties may petition agencies to adopt such systems. Forms of pure internal agency management are typically seen to escape notice and comment and judicial review, but as AI systems become increasingly powerful, parties might challenge the failure to adopt an AI-based system as arbitrary and capricious or as violating due process.

Going forward, judges, administrators, and legislators will need to think about more appropriate legal and institutional vehicles to challenge the accuracy not only of individual decisions, but also of algorithms. The APA's interpretation of a binding rule may need to be pegged to the degree to which human discretion is displaced or, alternatively put, the degree to which a human remains "in the loop." Agencies will need to experiment with the best ways to surface, investigate, and debug potential errors when adjudicators and affected parties suspect such errors. Such mechanisms appear to be lacking currently. In its audit of the Insight system, for instance, the SSA's Office of Inspector General surveyed adjudicators and 20% indicated that the flagged errors were inaccurate and 35% reported that there was no method for submitting feedback where improvement was necessary.[123] The broader scholarship suggests that appeal rights alone are unlikely to provide a full solution, so other institutional and managerial solutions, such as quality control programs, audits, oversight, and external review, are well worth piloting and evaluating.

III. HEARING RIGHTS INTO THE FUTURE

The academic literature is replete with references to "robo-judges"[124] and even an eventual state of "legal singularity,"[125] when machines can perfectly predict the outcomes of cases before they are filed. Only slightly less futuristic are predictions that the law will steadily transform into a "vast catalogue of precisely tailored laws," or "microdirectives,"[126] that adjust in real-time—for instance, an individualized speed limit for a given driver with a given amount of experience operating in specific driving conditions—and are enforced via automatic penalties.[127]

These possibilities may seem far-fetched in the current moment, but the more limited tools profiled in this report do gesture toward the longer technological horizon. The SSA Insight system's ability to spot errors in draft decisions and arm administrative judges with raw materials, including agency non-acquiescence decisions, point to a world in which decision-making becomes more fully automated. While administrative judges and other adjudicators may balk at full automation, some interview subjects seek tools that can build a "decisional shell" around a case by gathering factual and legal materials to which an adjudicator can then more efficiently apply her human discretion. Yet even decisional shells will displace human discretion based on editorial judgments about which legal issues, and which materials, are and are not relevant. These tools may be different in degree, not in kind.

While it may be far away, fully automated decision-making raises rich, and existential, questions to the American legal system, built around participatory rights and adversarialism. Does the notion of due process imply the right to a human decision-maker?[128] Full automation promises "a fast and refined prediction of the relevant legal effect"[129] and thus achieves one of the highest purposes of law, but may drain the law's capacity to adapt and to ventilate legal rules though dialogue and debate in fully public interpretive exercises.[130] Something may be lost when the process of enforcing collective value judgments about right conduct plays out in server farms rather than as part of a prolonged and often messy deliberative and adjudicatory process, even where the machine-driven version proves perfectly accurate.[131] These debates are well beyond the scope of this report. But the tools profiled herein suggest it is not too early to start them.

Gaming and Adversarial Learning

A challenge that cuts across growing agency reliance on algorithmic governance tools is the risk of adversarial learning and gaming by regulated parties.

I. THE RISK OF GAMING AND ADVERSARIAL LEARNING

Whenever the government brings greater transparency to previously discretionary decisions, those decisions become more gameable, with parties adjusting their behavior to maximize their chances at a favorable outcome. Algorithmic governance is no exception. Where algorithms are known to rely on particular variables or cutoffs, regulated parties can manipulate those variables and the values they take in order to secure a desirable result from the system. "Adversarial machine learning," or the use of machine learning to fool algorithmic models, only exacerbates this inherent risk.[132] With simpler forms of adversarial machine learning, adversaries can, for instance, exploit algorithmic tools to obtain favorable determinations, without changing the underlying characteristic the algorithm is designed to measure.[133] At the extreme, regulatory targets can even gain access to the tool itself and feed it new data to corrupt its outputs.[134]

As a concrete illustration, consider how adversaries might exploit the PTO's tools to adjudicate applications, as described in Part II. These tools help classify patent and trademark applications according to the PTO's taxonomy, as well as search for "prior art" and visually similar trademarks. Patent applicants have long tweaked their applications to try to obtain a desired classification, pushing their application to a unit with higher grant rates. Machine learning magnifies these opportunities for gaming. For example, adversaries could manipulate images in their patent applications to include random noise, which has been shown to dupe leading machine learning models into mis-classifying images.[135] Adversaries could thereby divert their applications to units more likely to rule in their favor, undermining the fairness and accountability of the underlying algorithm.

II. THE EFFECT ON AGENCIES AND ALGORITHMIC SYSTEMS

Gaming and adversarial learning have profound implications for the efficacy of algorithmic governance tools as well as political support for their use.

Gaming poses profound distributive concerns.

At the outset, it is worth noting that gaming can sometimes be salutary. While gamers are often self-serving—that is, seeking to maximize their take or minimize their loss within an algorithmic system—they need not be.[136] Gamers, depending on one's perspective on automation, range "from parasitic, to benign, to downright noble."[137] Gaming opportunities can also be deliberately built into a system to avoid unduly regressive policies, promote redistribution, or otherwise blunt the force of rigid regulatory regimes. Some have suggested that lax tax enforcement of the cash economy is one such example, where gaming in fact serves potentially desirable redistributive ends.[138]

That said, gaming often reduces the accuracy and efficacy of algorithmic systems. Consider, for example, the SEC's Form ADV Fraud Predictor, which aims to identify bad apple investment brokers and subject them to greater regulatory scrutiny. Regulated parties with knowledge of that tool's inner workings can adversarially craft their disclosures; they can include or omit key language in order to foil the system's classifier and keep their personnel off the SEC's radar.

This type of gaming poses profound distributive concerns. Better-heeled and more sophisticated regulated individuals and entities may have the time, resources, or know-how to navigate or even reverse-engineer algorithmic systems and

then take the evasive actions necessary to yield positive determinations and avoid adverse ones. As noted in Part II's profile of the SEC's enforcement tools, larger and better-resourced firms with a deeper bench of computer scientists and quantitative analysts may prove better able than smaller ones to reverse-engineer algorithmic enforcement tools and avoid regulatory action.

These distributive concerns can be amplified by contractor conflicts. Government contractors may seek to monetize or exploit their relationship to algorithmic tools for financial gain in other business relationships. Given that contractors are responsible for roughly 30% of AI/ML use cases, these concerns are grave. To take but one example, the company that provides the PTO's classification tool also sells services to patent applicants, advertising its PTO experience as one of its major assets.[139] Yet because not all parties can afford such services, better-resourced companies and individuals will be better able to game the system, whether to obtain government benefits or avoid regulatory scrutiny.

These distributive challenges may politicize the use of AI/ML tools over time. While regulatory "haves" may welcome government uptake of algorithmic tools if they believe they are better-equipped to game them or that the new tools will yield enforcement against a more diverse set of regulatory targets,[140] the "have nots," including the poor but also more middling segments of society, may not support a more efficient and effective algorithm-wielding government if they believe they will disproportionately shoulder its burdens. Indeed, some initial research suggests that citizens tend to rate algorithmic decision-making negatively compared to the status quo.[141] Support for government innovation can evaporate quickly if it is perceived as unfairly wielded.

III. THE WAY FORWARD

Given these challenges, administrative agencies need to be mindful in developing and deploying algorithmic tools. Architects of algorithmic models must consider whether and how to design their models to minimize opportunities for gaming and adversarialism.

For example, programmers can increase model complexity, reconfigure models periodically, and/or add randomness, all of which will make models harder to game.[142] They can also build models that rely on immutable traits, which regulated parties cannot readily change.[143] And they can use generative adversarial networks, training new tools against hostile adversaries that seek to fool them, which will make the algorithm less susceptible to attack in the long run.[144] But these measures, while making AI/ML tools harder to game, also come at a cost. They risk making models less interpretable to regulated parties (let alone the average citizen), reducing transparency and accountability.

Another reform would have agencies impose sanctions to encourage compliance with underlying regulatory procedures. For example, the PTO sanctions parties who breach duties of disclosure, candor, and good faith. To be effective, however, this approach requires strong mechanisms to detect regulatory violations, which can prove expensive and difficult to implement.

Ultimately, none of these reforms is a panacea. As administrative agencies develop algorithmic tools, they must balance the risk of gaming against other public values, including transparency, efficacy, and distributive concerns. Sometimes, agencies must tolerate gaming and adversarialism in service of a more transparent, more effective algorithmic system. In other cases, the right answer may be to create no algorithm at all, especially if it would lead to an expensive arms race of machine learning tools, without ultimately improving efficacy or citizen confidence in the system.

The External Sourcing Challenge: Contractors and Competitions

Part I's canvass of federal agency use of AI identified 157 use cases. While more than half of these (53%) were developed in-house by agency technologists, nearly as many came from external sources, with one-third (33%) coming from private commercial sources via the procurement process and a further, non-trivial proportion (14%) resulting from non-commercial collaborations, including agency-hosted competitions and government-academic partnerships. This roughly even split between internal and external sourcing suggests that each approach has significant advantages and disadvantages that agency personnel must weigh when developing AI-based tools. This section focuses on the benefits and costs of external sourcing over internal sourcing and fleshes out some of the trade-offs agencies face when choosing between them.

I. THE "MAKE" DECISION REVISITED

Sourcing decisions, as noted in the earlier section on capacity-building challenges, reflect the basic make-or-buy choice that agencies often face when performing governance tasks.[145] An agency can either hire and train personnel and assemble the raw materials needed to perform government tasks, or it can contract through the procurement process to buy them.[146]

As described previously, internal agency production of AI tools requires substantial agency technical capacity but can also yield a range of benefits. Advantages of internal sourcing include tools that are better-tailored, more policy compliant, and more accountable. An apt illustration, as described in detail previously, is the Insight tool internally developed at the Social Security Administration by Kurt Glaze, the attorney-turned-programmer. In designing that system, Glaze specifically designed the error flags that can be raised in draft decisions based upon "the flags that [he would have] wanted to have available as an adjudicator."[147] Importantly, co-location of policymakers and technologists can matter even where an agency opts to make its own tools. James Ridgway, who helped oversee the Board of Veteran's Appeals Caseflow project, ensured that the staff of the U.S. Digital Service would remain on site to avoid "deliver[ing] a system two years later that no one [would] use."[148]

Embedded technical expertise may also be necessary to automate tasks that are dynamic and changeable. For example, algorithmic enforcement tools like those deployed by the SEC use classifiers trained on past enforcement actions to "shrink the haystack" of current violators and direct the attention of line-level enforcement staff. But as noted in Part II, the misconduct those tools target is rarely static. Embedded expertise facilitates the continuous, iterative updating of algorithmic enforcement tools necessary to incorporate new modes of wrongdoing unearthed by agency staff and avoid an undue focus on past forms of misconduct.

Finally, internal agency development of AI tools limits leakage of information about a tool's technical and operational details that can undermine its utility. Here again, the enforcement tools under development and in use at the SEC, IRS, CMS, and EPA provide a compelling illustration because of their potential vulnerability to being reverse-engineered and evaded through adversarial learning. Another example is DHS's facial recognition system, which attackers might, with access to technical details about the system, be able to trick into incorrectly matching an innocent face with the no-fly list or permitting an individual on the no-fly list to escape detection. Even relatively simple attacks, as noted previously, can defeat the most advanced algorithmic systems.[149] Leakage of a tool's technical and operational details facilitates those attacks.

II. THE "BUY" DECISION: PROS AND CONS

While internal sourcing has many virtues, the benefits of external production are also significant. First, external sourcing may yield more technically sophisticated tools. One reason is that the private sector is not burdened by the compensation and hiring limitations that restrict the pool of talent that government agencies can tap. Budget constraints, civil service laws capping allowable salaries, and political sensitivities mean that government agencies may be priced out of labor markets for employees with advanced technical skillsets. Further, agency leadership may not prioritize technological innovation. Gerald Ray, a longtime Administrative Appeals Judge at the SSA who eventually became deputy executive director of the Office of Appellate Operations (OAO), worked around limitations on hiring technologists by identifying attorneys skilled in data analysis and computer science in order to develop the agency's AI toolkit. While the Office of Personnel Management (OPM) has since established a "data scientist" classification, thus easing the hiring burdens for technical positions, compensation caps and other limitations remain.[150]

Differences across the public and private sector can also make externally sourced governance tools cheaper than internally sourced ones. A long academic literature concludes that the private sector will often produce goods and services at lower cost because of a better-incentivized workforce and tighter managerial control.[151] Government-side employment constraints again loom large, including limits on hiring and firing and an inability to offer incentive-based compensation.

While external sourcing has numerous benefits, its drawbacks are also significant. Some of these are merely the flip-side of internal sourcing's advantage in generating well-tailored, policy compliant, and accountable tools. Algorithmic enforcement tools, as just noted, may require frequent updating to maximize efficacy in ways that generic commercialized AI systems, and the often protracted back-and-forth of the procurement process, are ill-suited to provide.[152] Because contractors typically operate at a remove from agency operations, external sourcing can also impose heavy monitoring and transaction costs. Where monitoring costs are low, as with well-specified services like garbage collection, the government can gain from the efficiency and expertise of the private sector. Where monitoring costs are high, however, and the governance tasks at issue involve significant discretion, private contractors may have incentives to engage in strategic corner-cutting, thus systematically degrading quality.[153] Profit-motivated contractors may also be less likely to ground key design and implementation decisions in public values like transparency and non-discrimination than civil servants as a matter of professional identity.[154] In the AI context, technically complex but standardized tasks, such as consolidating databases and upgrading computer infrastructure,[155] may prove more amenable to external contracting than the design and maintenance of enforcement tools, where the need for tailoring and updating is greater and consideration of public values are thought to be more salient.

Usability may militate in favor of internal capacity building.

Finally, external sourcing of algorithmic governance tools raises significant conflict-of-interest concerns. In the patent and trademark context, the same contractor that produced the PTO's classification tool advertises its experience supporting the agency in order to sell its services to patent applicants.[156] This raises the potential for conflicts of interest and deliberate leakage of information about governance tools. At the same time, because contractors seek to maximize commercial gain, they also face incentives to cloak the technical and operational details of AI tools by invoking intellectual property and trade-secret protections. As just one example, the DHS reported that it could not explain the failure rates of iris scanning technology due to the "proprietary technology being used."[157] This example underscores both the potential accountability costs of procurement-generated AI tools and also the importance of developing and maintaining a baseline level of internal technical capacity even when an agency chooses to buy AI tools.

In sum, usability may militate in favor of internal capacity building. Privately produced, procurement-generated tools may boast the most cutting-edge analytics, but may also be less tailored to the task at hand, be less attuned to legal requirements and an agency's bureaucratic realities, and do not necessarily come with ongoing and regular engagement between technologists and agency enforcement staff. In contrast, in-house production may strain agency budgets, but will yield governance tools that are, on average, better tailored to subtle governance tasks, more law- and policy-compliant, more attuned to complex organizational dynamics, and less subject to information leakage and conflicts of interest that can reduce a tool's efficacy and raise significant distributive concerns.[158]

III. A THIRD WAY: NON-COMMERCIAL COLLABORATIONS

While the make-or-buy choice clearly entails significant trade-offs, a third approach may be gaining momentum: non-commercial collaborations and competitions.[159] Questions of scalability remain, but this third approach highlights the potential for government to realize the benefits of make *and* buy while avoiding some of the costs of each.

Collaborations with professional associations, academe, and NGOs allow the government to leverage mutually beneficial relationships and gain access to external talent and expertise while maintaining control and monitoring quality. Examples of successful non-commercial collaborations are growing and include: the FDA's partnerships to address cybersecurity risk;[160] NHTSA's use of IMB's Watson to process and respond to safety complaints;[161] Stanford's partnership with the EPA;[162] the VA's partnership with Google's DeepMind to protect personal information and thus permit data-sharing;[163] and the FDA's "regulatory science" with Johns Hopkins, MIT, Stanford, and Harvard.[164] The FDA is also "exploring the use of a neutral third party [to] collect large annotated imaging data sets for purposes of understanding the performance of a novel AI algorithm."[165]

Government-sponsored competitions, which leverage the public's ideas and talent around declared government priorities, often with prize money attached, are a potentially valuable source of innovation and an increasingly prevalent part of the capacity-building landscape.[166] Through the use of prize money, public recognition, and even follow-up contracting work, government can leverage the public's talent to generate and prototype ideas.[167] While there is relatively little empirical or theoretical work on the subject, the benefits seem clear: incentivizing innovation while maximizing return by only rewarding success.[168]

At the same time, of the 28 competitions documented in Part I, half showed no public evidence of government adoption or intended adoption of technology created in the competition, raising doubts about their usefulness. Moreover, while competitions have grown exponentially from $247,000 in prize money awarded in FY2011 to over $30 million in FY2016, this amount remains small in comparison to the trillions in annual government outlays, raising questions about whether competitions can sufficiently scale to meet agency needs.[169] Finally, competition-generated tools are sometimes criticized as interstitial and small-bore. They may not substitute for a comprehensive automation strategy and, as with tools generated through the traditional procurement process, be insufficiently attuned to the complexities of tasks or organizational environments.

Finally, agencies can collaborate with each other, pooling scarce resources to tackle parallel technical challenges. Preliminary examples of such partnerships include the FDA and DHS' announced cybersecurity "memorandum of agreement … for greater coordination and cooperation … for addressing cybersecurity in medical devices."[170] The FDA's analysis of drug adverse event reports is remarkably similar in objectives to NHTSA's identification of trends in consumer complaints, raising the prospect of cross-agency technical collaboration through a central team building shared AI infrastructure.[171] Agencies like the FDA and SSA are consolidating technical expertise and self-assessing their technical infrastructure in order to improve technical performance, and a new proposed bill has been introduced to promote innovation and develop AI governance government-wide.[172]

Collaborations with professional associations, academe, and NGOs allow the government to leverage mutually beneficial relationships and gain access to external talent and expertise while maintaining control and monitoring quality.

Conclusion

Across the federal government, we are beginning to observe the dawn of a new chapter—perhaps even a digital revolution—in how government does its work. Half of surveyed agencies have experimented with AI/ML. AI-based governance tools already touch virtually every aspect of government, from enforcement to adjudication and from regulatory analysis and monitoring to citizen services. And though the sophistication of many of these tools lags behind the private sector, the pace of AI/ML development in government seems to be accelerating.

Few, however, have recognized, much less captured in any substantial detail, the breadth and depth of this transformation or the extent to which it is already underway. Until now, the state of knowledge about algorithmic governance has been marked above all else by its generality. The resulting high-abstraction mappings of concepts and core trade-offs have laid a valuable foundation. But further progress in thinking about the optimal regulation of the new AI governance tools is unlikely to take the form of a unified field theory. Instead, it will require a relentlessly interdisciplinary approach that engages with, rather than abstracting away from, the technical and operational details of the government's new algorithmic toolkit. This report has provided the first comprehensive effort to provide such an analysis by examining in detail what agencies are actually doing and then offering concrete recommendations for how agency officials, judges, and legislators should respond.

Further progress in thinking about the optimal regulation of the new AI governance tools is unlikely to take the form of a unified field theory. Instead, it will require a relentlessly interdisciplinary approach that engages with, rather than abstracting away from, the technical and operational details of the government's new algorithmic toolkit.

In providing a synoptic accounting of government use of AI, this report confirms both the promise and the peril of the current algorithmic moment. The prospect that AI can transform government has long excited some commentators and worried others, and with good reason. As we have detailed, the proliferation of AI throughout the federal administrative state is already raising urgent questions about how to resolve core trade-offs around accountability, efficacy, capacity building, and adversarialism. How much transparency is necessary to judge a tool's fidelity to law, and to what extent should we be willing to ease commitments to accountability and reason-giving in the service of governance tools that promise more effective and more equitable deployments of government power? To what extent can existing legal oversight tools, particularly administrative law, achieve meaningful accountability, and to what extent will accountability require newly minted interventions? How to weigh the relative merits of internal agency production of new governance tools and capacity building as against agency reliance on the ranks of private contractors with AI-based solutions at the ready?

Although the answers to these sorts of questions will in some cases bring welcome progress in society's efforts to make government more efficient and responsive to public needs, no doubt those answers will also deliver their share of surprises and continuing disagreement. What is clear at this point is that federal agencies are capable of significant innovation involving AI. They have begun making systematic use of AI for a vast range of functions, and have made, and will surely continue to make, widely varying assumptions about the

proper use of this technology. Agencies also have differential capacity to anticipate challenges and to make the most of the opportunities for innovation in this area as they are actively exploring new ways of using AI to advance their missions. Above all else, we hope the terrain we have covered through the descriptions and ideas in this report will help agency officials to talk to one another and ensure that they are asking the right questions.

Endnotes

Endnotes to *Introduction*

1 AI in Government Act of 2019, H.R. 2575, 116th Cong. (2019).

2 Press Release, Portman, Schatz Reintroduce Legislation to Improve Federal Government's Use of Artificial Intelligence, May 8, 2019, https://www.portman.senate.gov/newsroom/portman-schatz-reintroduce-legislation-improve-federal-governments-use-artificialhttps://www.portman.senate.gov/newsroom/portman-schatz-reintroduce-legislation-improve-federal-governments-use-artificial.

3 The Commercial Facial Recognition Act of 2019 proposes to regulate facial recognition at the federal level. S. 847, 116th Cong. (2019). At least one state is considering statewide regulation. Christian M. Wade, *Massachusetts Considers Bill to Limit Facial Recognition*, Gov't Tech. (Feb. 11, 2019), https://www.govtech.com/policy/Massachusetts-Considers-Bill-to-Limit-Facial-Recognition.html. Cities in California and Massachusetts have banned facial recognition outright. See Kate Conger et al., *San Francisco Bans Facial Recognition Technology*, N.Y. Times (May 14, 2019), https://www.nytimes.com/2019/05/14/us/facial-recognition-ban-san-francisco.html; Nik DeCosta-Klipa, B*rookline becomes 2nd Massachusetts Community to Ban Facial Recognition*, Boston.com (Dec. 12, 2019), https://www.boston.com/news/local-news/2019/12/12/brookline-facial-recognition; Sarah Ravani, *Oakland Bans Use of Facial Recognition Technology, Citing Bias Concerns*, S.F. Chron. (July 17, 2019), https://www.sfchronicle.com/bayarea/article/Oakland-bans-use-of-facial-recognition-14101253.php. *See also* Lily Hay Newman, *The Window to Rein in Facial Recognition is Closing*, Wired (Jul. 10, 2019).

4 *See* Tal Zarsky, *Governmental Data-Mining and Its Alternatives*, 116 Pa. St. L. Rev. 285 (2011); Fred H. Cate, *Government Data Mining: The Need for a Legal Framework*, 43 Harv. C.R.-C.L. L. Rev. 435, 438 (2008); Daniel J. Steinbock, Data Matching, Data Mining, and Due Process, 40 Ga. L. Rev. 1 (2005). For an overview of the last round of government automation in the 2000s, see William D. Eggers, Government 2.0: Using Technology to Improve Education, Cut Red Tape, Reduce Gridlock, and Enhance Democracy (2005).

5 *See* David E. Osborne & Ted Gaebler, Reinventing Government: How the Entrepreneurial Spirit Is Transforming the Public Sector 142 (1992). For an overview of the forces that birthed the "automated administrative state," including the "reinventing government" movement, *see* Danielle Keats Citron, *Technological Due Process*, 85 Wash. L. Rev. 1249, 1259 (2008).

6 *See*, *e.g.,* Herbert A. Simon, Administrative Behavior (4th ed. 1997). "Expert systems" denote symbolic logic-based AI systems in which a programmer or designer uses the knowledge of a subject-matter expert, such as a pathologist or an organic chemist, to specify a set of rules, generally based on logical inference, that can be used to reach a sensible decision.

7 *See, e.g.,* Jenna Burrell, *How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms*, 3 Big Data & Soc'y 1 (2016); Andrew Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 Fordham L. Rev. 1085, 1094-96 (2018).

8 Selbst & Barocas, *supra* note 7, at 1096-99. They cite Paul Ohm's example of predicting a shoe purchase on the basis of what kind of fruit one eats for breakfast as paradigmatically nonintuitive. Paul Ohm, *The Fourth Amendment in a World Without Privacy*, 81 Miss. L.J. 1309, 1318 (2012).

9 *Id*. at 1096-97.

10 Aaron Rieke, Miranda Bogen, & David G. Robinson, Upturn & Omidyar Network, Public Scrutiny of Automated Decisions: Early Lessons and Emerging Methods 19 (2018).

11 Deven R. Desai & Joshua A. Kroll, *Trust But Verify: A Guide to Algorithms and the Law*, 31 Harv. J.L. & Tech. 1, 5 (2017) ("[F]undamental limitations on the analysis of software meaningfully limit the interpretability of even full disclosures of software source code."); Joshua A. Kroll et al., *Accountable Algorithms,* 165 U. Pa. L. Rev. 633, 661 (2017). For a more general version of the point, *see* Mike Ananny & Kate Crawford, *Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability*, 20 New Media & Soc'y 973, 980 (2018).

12 For recent reviews of this active research area, see Ashraf Abdul et al., Trends and Trajectories for Explainable, Accountable, and Intelligible Systems, *An HCI Research Agenda*, CHI Conf. Hum. Factors Computing Syss. Procs. (2018); Finale Doshi-Velez & Been Kim, *Towards a Rigorous Science of Interpretable Machine Learning,* Cornell Univ. (2017), https://arxiv.org/abs/1702.08608. For visualization techniques and machine-based textual justifications, see L. A. Hendricks et al., *Generating Visual Explanations,* Eur. Conf. on Computer Vision (Springer, 2016), at 3-19; Chris Olah et al., *The Building Blocks of Interpretability*, Distill (Mar. 6, 2018), https://distill.pub/2018/building-blocks; Marco Tulio Ribeiro et al., *"Why Should I Trust You?" Explaining the Predictions of Any Classifier*, KDD '16 Procs. 22nd ACM SIGKDD Int'l Conf. Knowledge Discovery Data Mining (2016). That said, input-output analysis need not be technical. Some advocate interactive "tinker" interfaces that allow data subjects to manually enter and change data and observe results, yielding a "partial functional feel for the logic of the system." Selbst & Barocas, *supra* note 7, at 38. For a rough accounting of the relative opacity of different machine learning approaches, see Desai & Kroll, *supra* note 11, at 52. On the problem of dynamic algorithms, which "may (desirably) change between decisions," *see id.*, 41-43.

13 Wojciech Samek, Thomas Wiegand & Klaus-Robert Müller, *Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models* (2017), https://arxiv.org/abs/1708.08296.

14 For instance, advances in computer vision can reduce tasks that would comprise years of manual remote sensing to several days. See Cassandra Handan-Nader, Daniel E. Ho & Larry Y. Liu, *Deep Learning with Satellite Imagery to Enhance Environmental Enforcement,* in Data-Driven Insights and Decisions: A Sustainability Perspective (Prasanna Balaprakash & Jennifer B. Dunn eds., forthcoming 2020); Daniel E. Ho & Cassandra Handan-Nader, *Deep Learning to Map Concentrated Animal Feeding Operations*, 2 Nature Sustainability 298 (2019).

15 See Frank Pasquale, *Restoring Transparency to Automated Authority*, 9 J. Telecomm. & High Tech. L. 235, 235-36 (2011).

16 *See* Citron, *supra* note 5, at 1252, 1263-64; Cary Coglianese & David Lehr, *Transparency and Algorithmic Governance*, 71 Admin. L. Rev. 1, 4 (2019).

17 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).

18 *See* Ali Alkhatib & Michael Bernstein, *Street-Level Algorithms: A Theory at the Gaps between Policy and Decisions*, CHI Conf. Hum. Factors Computing Syss. Procs. (2019); Mark Bovens & Stavros Zouridis, *From Street-Level Bureaucracies to System-Level Bureaucracies: How Information and Communication Technology Is Transforming Administrative Discretion and Constitutional Control*, 62 Pub. Admin. Rev. 174 (2002). *See generally* Michael M. Lipsky, Street-Level Bureaucracy: The Dilemmas of the Individual in Public Service (1983).

19 An example is the recent effort by the U.S. Digital Service to develop web interfaces (APIs) for nutrition program eligibility. See Ed Mullen, *Implementing Rules Without a Rules Engine*, 18F (Oct. 9, 2018), https://18f.gsa.gov/2018/10/09/implementing-rules-without-rules-engines/; *Eligibility APIs Initiative: Helping States Turn Federal Eligibility Policy into Action*, Github, https://github.com/18F/eligibility-rules-service (last visited Dec. 13, 2019).

20 These distinctions are admittedly not always clear, particularly when given sparse descriptions.

Endnotes to *Part I. Taking Inventory: A Survey of Federal Agency Use of AI*

1 David Lewis & Jennifer Selin, Sourcebook of United States Executive Agencies (2d ed. 2018).

2 Due to substantive interest, there were four agencies that we included, despite lower employee counts in the ACUS sourcebook: DOJ's Civil Rights Division, the Board of Governors of the Federal Reserve, the Legal Services Corporation, and the Office of the Director of National Intelligence.

3 *See, e.g., Artificial Intelligence: With Great Power Comes Great Responsibility: Hearing Before the H. Comm. on Science, Space, and Technology,* 115th Cong. (2018); *Game Changers: Artificial Intelligence Parts I-III: Hearings Before the Subcomm. on Information Technology of the H. Comm. on Oversight and Government Reform*, 115th Cong. (2018); Jacques Bughin et al., McKinsey Global Institute, Artificial Intelligence: The Next Digital Frontier? (June 2017); Alex Campolo et al., AI Now, AI Now Report (2017); Disruptive Competition Project, Potential Uses of Artificial Intelligence for the Federal Government (2018); McKinsey Global Institute, The Promise and Challenge of the Age of Artificial Intelligence (2018); Aaron Rieke, Miranda Bogen, & David G. Robinson, Upturn & Omidyar Network, Public Scrutiny of Automated Decisions: Early Lessons and Emerging Methods 19 (2018).

4 For each agency, we took the following steps:
1. We conducted a search combining the agency's full name and the terms "artificial intelligence" or "machine learning," with results collected from a minimum of the first three resulting pages. Due to lack of specificity, we did not rely on agency acronyms (*e.g.,* USPTO).
2. We conducted agency site-specific searches for the terms "artificial intelligence" and "machine learning."
3. To identify all agency use cases, we used the boolean "-" operator to eliminate previously identified applications from searches.
4. For each agency, we cross-checked documented use cases on *algorithmtips.org*, an existing online database of noteworthy algorithms used by the federal government.

5 It is possible, for instance, that agencies describe as "machine learning" analytic approaches that are in reality more conventional forms of statistical inference because of the perceived public relations benefit. Where available, we examined technical or other documentation to ascertain whether machine learning was deployed. For our definition of machine learning, see the discussion of the study's "Scope" in the Introduction, *supra*.

6 As an illustration of this challenge, available descriptions and underlying technology can change over time, making quality control efforts over the course of the study period difficult.

7 These policy areas are tied to agencies. "Law enforcement" indicates an agency whose primary mission is in law enforcement, not civil enforcement activities at an agency that may be primarily engaged in, say, energy regulation.

8 We should note that the distinctions between these task categories are not always straightforward. As spelled out in our data collection protocol, in all instances, we attempted to code the task category that "best describes" the use case.

9 This statistic excludes use cases with mixed methods.

Endnotes to *Part II. Case Studies: Regulatory Enforcement at the Securities and Exchange Commission*

1 The Administrative Procedure Act is broadly organized into formal and informal rulemaking and adjudication. 5 U.S.C. §§ 551-559 (2018). In addition, § 553(b)(3)(A) refers to "rules of agency organization" and has been interpreted to encompass internal structuring choices. *Id.* at 70.

2 Margaret Lemos, *Democratic Enforcement? Accountability and Independence for the Litigation State*, 102 Cornell L. Rev. 929, 931 (2017); Roscoe Pound, *Law in Books and Law in Action*, 44 Am. L. Rev. 12 (1910), *reprinted in* American Legal Realism 39, 39-40 (William W. Fiscer III, Morton J. Horwitz & Thomas A. Reed eds., 1993).

3 A. Mitchell Polinsky & Steven Shavell, *The Economic Theory of Public Enforcement of Law*, 38 J. Econ. Lit. 45, 45 (2000).

4 Steven Shavell, *The Fundamental Divergence Between the Private and the Social Motive to Use the Legal System*, 26 J. Legal Stud. 575, 582 (1997).

5 *See generally* Rachel E. Barkow, *Insulating Agencies: Avoiding Capture Through Institutional Design*, 89 Tex. L. Rev. 15, 22-24 (2010); Preventing Regulatory Capture: Special Interest Influence and How to Limit it (Daniel Carpenter & David A. Moss eds., 2014).

6 *See What We Do*, Sec. & Exch. Comm'n, https://www.sec.gov/Article/whatwedo.html (last modified June 10, 2013). SEC commissioners serve staggered five-year terms, and no more than three of them may be from the same political party.

7 *Id.* The agency derives its regulatory authority from the 1933 Securities Act and the 1934 Securities Exchange Act, as supplemented by, among others, the Trust Indenture Act of 1939, the Investment Company Act of 1940, the Investment Advisers Act of 1940, the Sarbanes-Oxley Act of 2002, and the Dodd-Frank Wall Street Reform and Consumer Protection Act of 2010. *Laws that Govern the Securities Industry*, Sec. & Exch. Comm'n, https://www.sec.gov/answers/about-lawsshtml.html (last modified Oct. 1, 2013).

8 *What We Do, supra* note 6. SEC commissioners serve staggered five-year terms, and no more than three of them may be from the same political party.

9 Divisions include: the Division of Enforcement, which assists the Commission by recommending commencement of investigations and enforcement actions; the Division of Corporation Finance, which oversees corporate disclosures, such as earnings reports; the Division of Economic and Risk Analysis, which integrates economic analysis and data analytics into the SEC's work; the Division of Trading and Markets, which oversees key market participants, such as broker-dealers, securities exchanges, and self-regulating organizations (SROs); and the Division of Investment Management, which oversees registered investment advisers and investment companies (*e.g.,* mutual funds). *Id*.

10 An example of a standalone office is the Office of Compliance Inspections and Examinations, which administers nationwide examinations of various market actors and Commission "registrants." *Id*.

11 CIRA was developed within the Office of Risk Assessment at the Division of Economic and Risk Analysis (DERA). *See DERA - Office of Risk Assessment*, Sec. & Exch. Comm'n, https://www.sec.gov/page/dera_ora_page (Nov. 2, 2016). "Corporate issuers" develop and sell securities to finance their operations.

12 *See Fast Answers: Form 10-K*, Sec. & Exch. Comm'n, https://www.sec.gov/fast-answers/answers-form10khtm.html (last modified Nov. 2, 2016); *Fast Answers: Form 10-Q*, Sec. & Exch. Comm'n, https://www.sec.gov/fast-answers/answersform10qhtm.html (last modified Sept. 2, 2011). *See* also 15 U.S.C. §§ 78m, 78o(d) (2018).

13 *See* Scott W. Bauguess, Acting Dir. & Acting Chief Economist, DERA, Address to OpRisk North America: The Role of Big Data, Machine Learning, and AI in Assessing Risks: A Regulatory Perspective (June 21, 2017), https://www.sec.gov/news/speech/bauguess-big-data-ai#_ednref8 [hereinafter OpRisk Keynote]; *see also* Telephone Interview with Scott Bauguess, former Deputy Dir. & Deputy Chief Economist, Sec. & Exch. Comm'n (Feb. 15, 2019) [hereinafter Bauguess Interview I].

14 OpRisk Keynote, *supra* note 13; *see also* Telephone Interview with Scott Bauguess, former Deputy Dir. & Deputy Chief Economist, Sec. & Exch. Comm'n (Nov. 22, 2019) [hereinafter Bauguess Interview II]. For the seminal paper on random forests, see Leo Breiman, *Random Forests*, 45 Machine Learning 5 (2001).

15 Scott W. Bauguess, Deputy Dir. & Deputy Chief Economist, Sec. & Exchange Comm'n, Address to the Midwest Regional Meeting––American Accounting Association: Has Big Data Made Us Lazy? (Oct. 21, 2016), https://www.sec.gov/news/speech/bauguess-american-accounting-association-102116.html; *see also DERA, CIRA, and XBRL at the SEC: Expanding the Availability and Use of XBRL Data*, Fin. Executives Int'l (July 1, 2015), https://daily.financialexecutives.org/dera-cira-and-xbrl-at-the-sec-expanding-the-availability-and-use-of-xbrl-data/ (listing among the CIRA inputs commercial databases as well as XBRL (eXtensible Business Reporting Language) data, which is a computer readable business reporting markup language).

16 *See* Bauguess Interview II, *supra* note 14. Unlike an earlier generation of data-driven tools at the SEC, CIRA is designed to spot fraud before it becomes public. What is now CIRA was originally called the Accounting Quality Model ("AQM") and was used to monitor inappropriate managerial discretion in the usage of accounting accruals. While dubbed by the press as "Robocop," the tool has in fact always depended on human analysis of its outputs. Janet Novack, *How SEC's New RoboCop Profiles Companies for Accounting Fraud*, Forbes (Aug. 9, 2013), https://www.forbes.com/sites/janetnovack/2013/08/09/how-secs-new-robocop-profiles-companies-for-accounting-fraud/#35d3072812d1.

17 ARTEMIS was developed at the Division of Enforcement. ATLAS was developed in the Philadelphia Regional Office by the Office of Compliance Inspections and Examinations in collaboration with the Division of Enforcement.

18 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.

19 Telephone Interview with SEC Staff Members, Complex Fin. Instruments Unit, Enforcement Div., Sec. & Exch. Comm'n (Feb. 15, 2019) [hereinafter SEC Staff Interview I]. On NLP methods, *see* CHRISTOPHER D. MANNING & HINRICH SCHÜTZE, FOUNDATIONS OF STATISTICAL NATURAL LANGUAGE PROCESSING (1999); DANIEL JURAFSKY & JAMES H. MARTIN, SPEECH AND LANGUAGE PROCESSING (2d ed. 2008).

20 *See* Manning & Schütze, *supra* note 19.

21 According to SEC staff, bluesheet data is requested when the total amount of trading exceeds a threshold. The size of individual transactions is not a threshold for the request, though it is a factor in the analysis of the transactions. Although a tender offer is used as a threshold for issuing a bluesheet, the SEC is also interested in the dimensions of longitudinal trading history to identify outliers. Telephone Interview with SEC Staff Members, Complex Fin. Instruments Unit, Enforcement Div., Sec. & Exch. Comm'n (Feb. 15, 2019) [hereinafter SEC Staff Interview I, *supra* note 19.

22 *Id.* The SEC's authority to make bluesheet requests from the broker/dealer community derives from Section 17(a), Rule 17a-25, of the Securities Exchange Act. A sample electronic bluesheet (also referred to as an "EBS") is publicly available through the FINRA website and can be examined to understand the criteria of data requested by the SEC. *See* Fin. Indus. Reg. Auth., 18-04, Regulatory Notice: Electronic Bluesheet Submissions, Attachment A (Jan. 29, 2018), http://www.finra.org/sites/default/files/

notice_doc_file_ref/Regulatory-Notice-18-04.pdf. The trading record, by regulation, includes trading information (the name of the security, whether the transaction was a buy or 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 (2018).

23 Sec. & Exch. Comm'n, Div. of Enforcement, Enforcement Manual § 3.2.2 (2017) [hereinafter Enforcement Manual]. The Securities Exchange Act only authorizes the SEC to request data through bluesheets from the past three years. *Id. See* also Bauguess Interview I, *supra* note 13.

24 Enforcement Manual, *supra* note 23.

25 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. Press Release, Fin. Indus. Reg. Auth., FINRA Fines Deutsche Bank Securities Inc. $6 Million for Submitting Inaccurate and Late Blue Sheet Data (June 29, 2016) (on file with the author). 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. Press Release, Sec. & Exch. Comm'n, Citigroup Provided Incomplete Blue Sheet Data for 15 Years (July 12, 2016) (on file with the author). Note, however, that these enforcement actions are likely a patchwork approach to validating bluesheet data. Because inaccurate bluesheets are often not detected and charged for years after their submission, the resulting delays in the validation process may create the risk of the SEC using inaccurate training data for its AI/ML models.

26 For instance, in a takeover situation, the SEC tends to organize data according to buy volume. For increased trading before a critical announcement, such as an FDA drug approval, the SEC sorts the data by sell volume to identify potential insider trading. A retail investor with a diversified set of index funds suddenly leveraging in a biotechnology company before an FDA approval would be flagged as an outlier.

27 SEC Staff Interview I, *supra* note 19.

28 "Ground truth" is a term used in statistical analysis, including machine learning, to refer to provable information derived from direct observation rather than inference.

29 *Id*.

30 The features may include how often the trader normally trades the stock in question, how often she trades other stocks, how many shares were traded in comparison to other trades, and the time between the announcement and trade. *Id*.

31 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.

32 The tool was developed within the Division of Investment Management.

33 Fast Answers: Form ADV, Sec. & Exchange Comm'n (Mar. 11, 2011), https://www.sec.gov/fast-answers/answersformadvhtm.html.

34 *Id*.

35 OpRisk Keynote, *supra* note 13.

36 Telephone Interview with SEC Staff, Office of Research & Data Servs., Div. of Econ. & Risk Analysis and Office of Analytics & Research, Sec. & Exch. Comm'n (Feb. 20, 2019) [hereinafter SEC Staff Interview II]. During this pre-processing step, a pair of algorithms processes the data by converting the PDF forms into a block of text and then splits the block into sections that answer specific questions in the form. *Id*.

37 *See* David M. Blei, Andrew Y. Ng & Michael I. Jordan, *Latent Dirichlet Allocation*, J. Machine Learning Res. 993 (2003).

38 A standard bag of words model works by constructing a frequency matrix summing the number of times each word is repeated in a document. As an example, the document "John bought stocks. Mary bought stocks" would be converted to "BoW = {"John":1,"bought":2,"stocks":2,"Mary":1}." LDA is described as an iterative generative statistical model because it is run for multiple epochs to find the top K topics that best represented the collection of documents. LDA outputs the probability that a word occurs given a specific topic. A topic is defined as a probability distribution over a set of words. LDA then scores each of the documents according to how closely they fit the top K topics. These topics are clusters and then used as inputs to a supervised learning algorithm. There is no public information about the specifics of the system's implementation, but clustering on these topics could possibly be done by comparing their distance from each other using a cosine similarity metric or the Kullback–Leibler divergence metric. *See* Manning & Schütze, *supra* note 19, at 513, 540-41.

39 SEC Staff Interview II, *supra* note 36.

40 *Id. See also* Ceshine Lee, *Feature Importance Measures for Tree Models — Part I*, The Artificial Imposter (Oct. 28, 2017), https://medium.com/the-artificial-impostor/feature-importance-measures-for-tree-models-part-i-47f187c1a2c3. The limitations of these and other efforts to render machine learning models more explainable is treated in Part III's discussion of transparency and accountability challenges.

41 Interview with SEC Staff, Div. of Enforcement, Sec. & Exch. Comm'n (Feb. 20, 2019) [hereinafter SEC Staff Interview III].

42 Danielle Ensign et al., *Runaway Feedback Loops in Predictive Policing*, Cornell U. (Dec. 22, 2017), https://arxiv.org/abs/1706.09847. For a more general survey of potential problems with predictive policing, *see* Andrew Guthrie Ferguson, *Policing Predictive Policing*, 94 Wash U.L. Rev. 1109 (2017).

43 Deven R. Desai & Joshua A. Kroll, *Trust But Verify: A Guide to Algorithms and the Law*, 31 Harv. J.L. & Tech. 21 (2017) (noting that systems require "ongoing monitoring and evaluation to ensure the model remains accurate given that the real world changes").

44 Erik Hemberg et al., *Tax Non-Compliance Detection Using Co-Evolution of Tax Evasion Risk and Audit Likelihood*, ICAIL '15 (2015).

45 As an example, the FPS tool in use at CMS appears to train models using data from claims that were *successfully* prosecuted for fraud, either through administrative action or by referral to law enforcement. Telephone Interview with Raymond Wedgeworth, Dir., Data Analytics & Sys. Grp., Ctrs. for Medicare & Medicaid Servs. (Mar. 1, 2019). Fraudulent claims that were not successfully prosecuted may not be included in training data. As an initial matter, this once again raises the possibility of encoded, human-level arbitrariness or bias. Unless explicitly accounted for, predictive models may be biased towards previously prosecuted groups or areas, running into similar issues as predictive policing. Further problems can result from ad hoc procedures for incorporating newly discovered forms of wrongdoing. At CMS, when investigators unearth a new form of fraud outside the FPS, they report it, and the technical team managing the system engages with investigators to determine whether it is worthwhile to add the new activity to the system. *Id.* If not done carefully, these updating efforts can aggravate overfitting by producing a narrow set of training data covering only a subset of claims successfully identified as fraudulent. As we note below, enforcement agencies can mitigate these problems by engaging in more systematic training of their systems through randomized claim sampling that includes both cases identified as problematic and cases that are not.

46 *SEC Approves Plan to Create Consolidated Audit Trail*, Sec. & Exch. Comm'n (Nov. 15, 2016), https://www.sec.gov/news/pressrelease/2016-240.html.

47 *Rule 613 (Consolidated Audit Trail)*, Sec. & Exch. Comm'n, https://www.sec.gov/divisions/marketreg/rule613-info.htm (last modified Apr. 19, 2019). *See also*, *Perspectives: Consolidated Audit Trail: The Wait Is Over,* Deloitte, https://www2.deloitte.com/us/en/pages/financial-services/articles/sec-rule-613-consolidated-audit-trail-national-market-system-nms-plan-banking-securities.html (last visited Nov. 3, 2019).

48 SEC Staff Interview III, *supra* note 41.

49 See Bauguess Interview II, *supra* note 14.

50 *Id*.

51 *See* David Freeman Engstrom & Daniel E. Ho, *Algorithmic Accountability in the Administrative State,* 37 Yale J. on Reg. (forthcoming 2020).

52 Most NLP tools depend on "word embeddings"—that is, a process by which words are mapped to vectors of real numbers that permit computation of their similarity to other words.

53 Tim Loughran & Bill McDonald, *When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks*, 66 J. Fin. 35, 36 (2011).

54 Oliver E. Williamson, *Public & Private Bureaucracies: A Transaction Cost Perspective,* 15 J.L. Econ. & Org. 306, 319 (1999).

55 *See* David Freeman Engstrom & Daniel E. Ho, *Artificially Intelligent Government: A Review and Agenda*, in Big Data Law (Roland Vogl ed., forthcoming 2020); *see also* Coglianese & Lehr, *supra* note 16, at 24; Robert L. Glicksman, David L. Markell & Claire Monteleoni, *Technological Innovation, Data Analytics, and Environmental Enforcement*, 44 Ecology L.Q. 41, 47 (2017).

56 Ensuring a human remains "in the loop" does not guarantee the application of human judgment, given the well-documented issue of "automation bias"—*i.e.,* the human tendency to increasingly and unreasonably defer to automated outputs over time. R. Parasuraman & D.H. Manzey, *Complacency and Bias in Human Use of Automation: An Attentional Integration*, 52 Hum. Factors 381, 391 (2010); Linda J. Skitka, Kathleen L. Mosier, & Mark Burdick, *Does Automation Bias Decision-Making?*, 51 Int'l. J. Hum.-Computer Stud. 991 (1991).

57 *See* Cassandra Handan-Nader, Daniel E. Ho & Larry Y. Liu, *Deep Learning with Satellite Imagery to Enhance Environmental Enforcement*, in Data-Driven Insights and Decisions: A Sustainability Perspective (Prasanna Balaprakash & Jennifer B. Dunn eds., 2020); Daniel E. Ho & Cassandra Handan-Nader, *Deep Learning to Map Concentrated Animal Feeding Operations,* 2 Nature Sustainability 298 (2019).

58 Interview with Jeff Butler, Director of Research Databases, Internal Revenue Serv. (Feb. 11, 2019) (on file with author) [hereinafter Butler Interview]. Unsupervised models are designed to find latent patterns in unlabeled data (*i.e.,* data for which there is "no associated response yi" for observations of xi). Gareth James et al., An Introduction to Statistical Learning with Applications in R 26 (2013). Unsupervised learning techniques have become more popular in recent years with the rapid developments in neural networks—a set of algorithms that are well-suited to "discovering latent structures within unlabeled, unstructured data." *A Beginner's Guide to Neural Networks and Deep Learning*, Skymind, https://skymind.ai/wiki/neural-network#define (last visited Apr. 8, 2019); Geoffrey Hinton et al., *The "Wake-Sleep" Algorithm for Unsupervised Neural Networks*, 268 Science 1158, 1158-61 (1995). Based loosely on the biological structure of neurons firing in the brain, neural networks consist of "layers" of "nodes." *Id*.

59 Butler Interview, *supra* note 58. Butler in particular focused on Generative Adversarial Networks (GANs). Ian J. Goodfellow et al., *Generative Adversarial Networks*, Cornell U. (June 10, 2014), https://arxiv.org/abs/1406.2661. GANs consist of a generator neural network and a discriminator neural network that are pitted against one another in order to anonymize data. Alec Radford et al., *Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks*, Cornell U. (Jan. 7, 2016), https://arxiv.org/abs/1511.06434; Tim Salimans et al., *Improved Techniques for Training GANs*, Cornell U. (June 10, 2016), https://arxiv.org/abs/1606.03498. The generator is tasked with creating data instances and the discriminator tries to determine whether the new data instance is generated data or real data from the training dataset. Using GANs, an agency may be able to generate and release synthetic data.

60 Gov't Accountability Office, GAO-18-544, Tax Fraud and Noncompliance 5 (2018). The RRP system ingests a wide range of data, from taxpayer and employer filings to federal and state prison records, and it uses a mix of supervised and unsupervised machine learning models to generate fraud risk scores for all national individual tax returns claiming a refund. U.S. Dep't of the Treasury, Annual Privacy, Data Mining, and Section 803 Report 24 (2018).

61 *See Heckler v. Chaney*, 470 U.S. 821 (1985) (holding that agency decisions not to enforce are not subject to review); Fed. Trade Comm'n v. Standard Oil Co., 449 U.S. 232, 242 (1980) (holding that an agency's decision to proceed with an enforcement action is not immediately challengeable).

62 *See, e.g.,* Anthony D. Joseph et al., Adversarial Machine Learning (2019); Jane R. Bambauer & Tal Zarsky, *The Algorithm Game*, 94 Notre Dame L. Rev. 1, 10 (2018); Joshua A. Kroll et al., *Accountable Algorithms,* 165 U. Pa. L. Rev. 699 (2017); Daniel Lowd & Christopher Meek, *Adversarial Learning*, Proc. Eleventh ACM SIGKDD Int'l Conf. Knowledge Discovery Data Mining 641, 641 (2005).

63 Sean Farhang, The Litigation State: Public Regulation and Private Lawsuits in the United States (2010). Indeed, one explanation for that shift, achieved via legislatively created private rights of action and whistleblower schemes, is a legislative desire to surface privately held information about misconduct while alleviating strain on public budgets. 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).

64 On the risks of hollowing out, *see* Peter H. Schuck, Why Government Fails So Often: And How It Can Do Better (2014); Paul Verkuil, Valuing Bureaucracy: The Case for Professional Government (2017).

65 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).

66 Bambauer & Zarsky, *supra* note 62, at 11.

67 *See* David Freeman Engstrom & Daniel E. Ho, *Enforcement by Algorithm* (working paper 2020) (on file with author).

Endnotes to *Part II. Case Studies: Law Enforcement at Customs and Border Protection*

1 *See* NYU Policing Project, Privacy Audit & Assessment of Shotspotter, Inc.'s Gunshot Detection Technology, https://www.policingproject.org/shotspotter (last visited Nov. 3, 2019).

2 Axon AI & Policing Tech. Ethics Board, Automated License Plate Readers 6 (Oct. 2019), https://www.policingproject.org/axon.

3 *See* Ali Winston & Ingrid Burrington, *A Pioneer in Predictive Policing Is Starting a Troubling New Project*, The Verge (Apr. 26, 2018), https://www.theverge.com/2018/4/26/17285058/predictive-policing-predpol-pentagon-ai-racial-bias.

4 *See* Matt Stroud, *Chicago's Predictive Policing Tool Just Failed a Major Test*, The Verge (Aug. 19, 2016), https://www.theverge.com/2016/8/19/12552384/chicago-heat-list-tool-failed-rand-test.

5 Ali Winston, *Palantir Has Secretly Been Using New Orleans to Test its Predictive Policing Technology*, The Verge (Feb. 27, 2018), https://www.theverge.com/2018/2/27/17054740/palantir-predictive-policing-tool-new-orleans-nopd; *see also* Andrew G. Ferguson, *Policing Predictive Policing*, 94 Wash. U.L. Rev. 1109, 1142 (2017).

6 *See* John Eligon & Timothy Williams, *Police Program Aims to Pinpoint Those Most Likely to Commit Crimes*, N.Y. Times (Sept. 25, 2015), https://www.nytimes.com/2015/09/25/us/police-program-aims-to-pinpoint-those-most-likely-to-commit-crimes.html.

7 *The Debate Over Facial Recognition Technology's Role in Law Enforcement*, NPR (July 10, 2019, 5:48 PM ET), https://www.npr.org/2019/07/10/740480966/the-debate-over-facial-recognition-technologys-role-in-law-enforcement ("LA County has been using facial recognition technology since 2009.").

8 *See* Clare Garvie, Alvaro Bedoya, & Jonathan Frankle, *The Perpetual Line-Up: Unregulated Police Face Recognition in America*, Geo. L. Ctr. on Privacy & Tech. (Oct. 18, 2016), https://www.perpetuallineup.org.

9 *See, e.g.,* Nick Wingfield, *Amazon Pushes Facial Recognition to Police. Critics See Surveillance Risk*, N.Y. Times (May 22, 2018), https://www.nytimes.com/2018/05/22/technology/amazon-facial-recognition.html; Ian Wren & Scott Simon, *Body Camera Maker Weighs Adding Facial Recognition Technology*, NPR (May 12, 2018), https://www.npr.org/2018/05/12/610632088/what-artificial-intelligence-can-do-for-local-cops.

10 *See* Levi Sumagaysay, *Berkeley Bans Facial Recognition*, Mercury News (Oct. 16, 2019, 4:23 PM), https://www.mercurynews.com/2019/10/16/berkeley-bans-facial-recognition/.

11 Chris Mills Rodrigo, *Booker Introduces Bill Banning Facial Recognition Tech in Public Housing*, The Hill (Nov. 1, 2019), https://thehill.com/policy/technology/468582-booker-introduces-bill-banning-facial-recognition-tech-in-public-housing.

12 John Buntin, *Social Media Transforms the Way Chicago Fights Gang Violence*, Governing (Oct. 2013), https://www.governing.com/topics/public-justice-safety/gov-social-media-transforms-chicago-policing.html. Risk prediction at the policing stage has been the subject of much criticism. *See, e.g.,* John Eligon & Timothy Williams, *supra* note 6.

13 *See State v. Loomis*, 881 N.W.2d 749 (Wis. 2016).

14 *See id.* (raising a due process claim against a risk prediction sentencing tool).

15 U.S. Dep't of Homeland Sec., Biometric Pathway: Transforming Air Travel, Version 3.0 1 (Dec. 1, 2016), https://epic.org/foia/dhs/cbp/biometric-entry-exit/Biometric-Pathway.pdf.

16 *See Boots on the Ground or Eyes in the Sky: How Best to Utilize the National Guard to Achieve Operational Control: Hearing Before the H. Comm. on Homeland Security, Subcomm. On Border and Maritime Security*, 112th Cong. (Apr. 17, 2012) (statement of Ronald Vitiello, U.S. Customs and Border Protection Office of Border Patrol Deputy Chief), https://www.dhs.gov/news/2012/04/17/written-testimony-us-customs-and-border-protection-house-homeland-security.

17 *About CBP: History*, Customs & Border Protection, https://www.cbp.gov/about/history (last updated July 30, 2019).

18 *Id.*

19 *See About CPB: Leadership/Organization*, Customs & Border Protection, https://www.cbp.gov/about/leadership-organization/executive-assistant-commissioners-offices (last updated Mar. 22, 2017).

20 *Id.*

21 *Id.*

22 *Summary of Laws Enforced by CBP*, Customs & Border Protection, https://www.cbp.gov/trade/rulings/summary-laws-enforced/us-code (last updated Mar. 8, 2014).

23 *See Mission Statement,* Customs & Border Protection, https://www.cbp.gov/about (last updated Nov. 21, 2016) ("On a typical day, Customs and Border Protection welcomes nearly one million visitors, screens more than 67,000 cargo containers, arrests more than 1,100 individuals, and seizes nearly 6 tons of illicit drugs.").

24 *See, e.g.,* Marcy Mason, *Stopping Smugglers: How CBP's Aircraft Search Team Uncovers Internal Conspiracies with the Airlines*, CBP: Frontline, https://www.cbp.gov/frontline/stopping-smugglers-how-cbps-aircraft-search-team-uncovers-internal-conspiracies-airlines (last visited Nov. 3, 2019).

25 *About CBP: History*, *supra* note 17. In 2017, the agency completed 635 unmanned drone missions along America's borders. Matt Novak, *U.S. Border Patrol Flew More Drone Missions Last Year than Ever Before*, Gizmodo (Sept. 26, 2018), https://gizmodo.com/u-s-border-patrol-flew-more-drone-missions-last-year-t-1829323612.

26 *See* H.R. 1625, 115th Cong. § 230 (2018), https://www.congress.gov/bill/115th-congress/house-bill/1625/text.

27 Exec. Order No. 13,780, 82 Fed. Reg. 13209 (Mar. 6, 2017), https://www.govinfo.gov/content/pkg/FR-2017-03-09/pdf/2017-04837.pdf; *see also* U.S. Dep't of Homeland Sec. & U.S. Dep't of Justice, Exec. Order 13,780: Protecting the Nation from Foreign Terrorist Entry into the United States Initial Section 11 Report (Jan. 2018), https://www.dhs.gov/sites/default/files/publications/Executive%20Order%2013780%20Section%2011%20Report%20-%20Final.pdf.

28 *See CBP to Implement a Facial Comparison Technical Demonstration at Anzalduas International Bridge for Vehicle Travelers*, Customs & Border Protection (Aug. 29, 2018), https://www.cbp.gov/newsroom/local-media-release/cbp-implement-facial-comparison-technical-demonstration-anzalduas.

29 Dep't of Homeland Sec. Science and Tech. Directorate, Risk Prediction Program (Nov. 20, 2011), https://www.dhs.gov/sites/default/files/publications/Risk%20Prediction-508_0.pdf.

30 Marcy Mason, *Biometric Breakthrough: How CBP Is Meeting its Mandate and Keeping America Safe*, CBP: Frontline, *https://www.cbp.gov/frontline/cbp-biometric-testing* (last visited Nov. 3, 2019).

31 *Id.*

32 *Id.*

33 *Id.*

34 For a historical summary of facial recognition methods, *see* Daniel Sáez Trigueros, Li Meng & Margaret Hartnett, *Face Recognition: From Traditional to Deep Learning Methods*, Cornell U. (Oct. 31, 2018), https://arxiv.org/abs/1811.00116. For an overview of modern deep learning approaches, *see* Mei Wang & Weihong Deng, *Deep Face Recognition: A Survey*, Cornell U. (Feb. 12, 2019), https://arxiv.org/abs/1804.06655.

35 *What are Biometrics?*, Kaspersky Labs, https://usa.kaspersky.com/resource-center/definitions/biometrics (last visited Apr. 7, 2019).

36 *See* Davey Alba, *The U.S. Government Will Be Scanning Your Face at 20 Top Airports, Documents Show*, Buzzfeed News (Mar. 11, 2019), https://www.buzzfeednews.com/article/daveyalba/these-documents-reveal-the-governments-detailed-plan-for ("I think it's important to note what the use of facial recognition [in airports] means for American citizens," Jeramie Scott, director of EPIC's Domestic Surveillance Project, told BuzzFeed News in an interview. "It means the government, without consulting the public, a requirement by Congress, or consent from any individual, is using facial recognition to create a digital ID of millions of Americans.").

37 Off. of Inspector Gen., U.S. Dep't of Homeland Sec., Pub. No. OIG-18-80, Progress Made, but CBP Faces Challenges Implementing a Biometric Capability to Track Passenger Departure Nationwide 3 (Sept. 21, 2018), https://www.oig.dhs.gov/sites/default/files/assets/2018-09/OIG-18-80-Sep18.pdf.

38 U.S. Customs & Border Protection, Southern Border Pedestrian Field Test: Summary Report 8 (2016), https://epic.org/foia/dhs/cbp/biometric-entry-exit/Southern-Border-Pedestrian-Field-Test-Report.pdf [hereinafter Test Summary Report].

39 Biometric Pathway, *supra* note 15, at 1 ("The use of face as the primary modality, with the large gallery of available biometrics, removes the need to segment travelers and provides a previously unavailable method to facilitate travel for everyone, not just the smaller population of in-scope travelers for whom fingerprints are available. Additionally, the use of facial recognition does not require the collection of new information; Customs and Border Protection will leverage information travelers have already provided to the U.S. government.").

40 *Id.* at 1.

41 Mason, *supra* note 30.

42 *EPIC v. CBP (Biometric Entry/Exit Program),* Elec. Privacy Info. Ctr., https://epic.org/foia/dhs/cbp/biometric-entry-exit/ (last visited Apr. 7, 2019) ("In 2017, Customs and Border Protection launched the Traveler Verification Service (TVS). Under this program, a passenger's flight check-in prompts the TVS to compile a 'gallery' of pre-existing photographs of the passenger. These photographs may include photographs captured by the Department of State from U.S. passports and U.S. visas, as well as photographs from previous encounters with Customs and Border Protection or the Department of Homeland Security. Before a passenger boards an aircraft, a camera takes a 'live' photograph of the passenger, which the TVS compares to the passenger's gallery to verify the passenger's identity.").

43 In addition to Unisys, other recent vendors include Government Acquisitions, Inc. (for Facial Recognition matching algorithms), FS Partners LLP (for Facial Recognition Cameras), and GOVPLACE (for Facial Recognition cameras and software). *See* USA Spending, https://www.usaspending.gov (Awards No. 70B04C18F00000039, HSBP1017J00203, and HSBP1017J00203, respectively).

44 NEC's Video Face Recognition Technology Ranks First in NIST Training, NEC (Mar. 16, 2017), https://www.nec.com/en/press/201703/global_20170316_01.html.

45 Stephen Mayhew, Unisys *Integrates NEC's Facial Recognition Software in CBP Project at JFK Airport*, Biometric Update (May 8, 2016), https://www.biometricupdate.com/201605/unisys-integrates-necs-facial-recognition-software-in-cbp-project-at-jfk-airport.

46 Memorandum of Understanding, Customs & Border Protection (2017) https://epic.org/foia/dhs/cbp/biometric-entry-exit/MOU-Biometric-Pilot-Project.pdf [hereinafter Memorandum of Understanding].

47 Dep't of Homeland Sec., Biometric Entry-Exit 4-5 (2017), https://acconline.org//documents/3A1_Biometrics_Hardin.pdf.

48 *Id.* at 5.

49 *See CBP Advances Biometric Exit Mission as Orlando International Airport Becomes First US Airport to Commit to Facial Recognition Technology*, Customs & Border Protection (June 21, 2018), https://www.cbp.gov/newsroom/national-media-release/cbp-advances-biometric-exit-mission-orlando-international-airport.

50 As of March 2019, these airports included Atlanta, Chicago, Seattle, San Francisco, Las Vegas, Los Angeles, Washington (Dulles and Reagan), Boston, Fort Lauderdale, Houston Hobby, Dallas/Fort Worth, JFK, Miami, San Jose, Orlando, and Detroit. CBP plans to expand facial recognition to additional airports. Alba, *supra* note 36.

51 Mason, *supra* note 30 ("Since June 2016 . . . passengers like the young Mexican woman have been found daily. 'She was typical of the people who have entered without inspection,' said Frazier. 'Most days we find a minimum of two or three undocumented people, but sometimes we find as many as eight to 10 boarding a flight.'").

52 *See EPIC v. CBP*, *supra* note 42 ("Without any formal procedure in place, Customs and Border Protection has frequently changed the FAQs provided on the agency's website with regard to opt-out procedures.").

53 *See* Dep't. of Homeland Sec., *supra* note 37, at 8 ("Customs and Border Protection allowed U.S. citizens to decline participation in the pilot. In such cases, Customs and Border Protection officers would permit the travelers to bypass the camera and would instead check the individuals' passports to verify U.S. citizenship. When a U.S. citizen opted to participate in the pilot but did not successfully match with a gallery photo, the Customs and Border Protection officer would examine the individual's passport but did not collect fingerprints.").

54 Biometric Pathway, *supra* note 15, at 3.

55 *Id.* at 4.

56 Dep't. of Homeland Sec., Report 2018-01 of the DHS Data Privacy and Integrity Advisory Committee (DPIAC): Privacy Recommendations in Connection with the Use of Facial Recognition Technology (Dec. 2018), https://www.dhs.gov/sites/default/files/publications/Report%202018-01-Draft%20Report%20on%20Privacy%20Recommendations%20in%20Connection%20with%20the%20Use%20of%20Facial%20Recognition%20Technology.pdf.

57 *EPIC v. CBP*, *supra* note 42.

58 Biometric Pathway, *supra* note 15.

59 *See* Test to Collect Facial Images from Occupants in Moving Vehicles at the Anzalduas Port of Entry (Anzalduas Biometric Test), 83 Fed. Reg. 56862 (Nov. 14, 2018), https://www.federalregister.gov/documents/2018/11/14/2018-24850/test-to-collect-facial-images-from-occupants-in-moving-vehicles-at-the-anzalduas-port-of-entry.

60 *CBP to Implement a Facial Comparison Technical Demonstration at Anzalduas International Bridge for Vehicle Travelers*, Customs & Border Protection (Aug. 19, 2018), https://www.cbp.gov/newsroom/local-media-release/cbp-implement-facial-comparison-technical-demonstration-anzalduas.

61 Automated Targeting System of Records, 72 Fed. Reg. 43650 (Sep. 5, 2007), https://www.federalregister.gov/documents/2007/08/06/E7-15197/privacy-act-of-1974-us-customs-and-border-protection-automated-targeting-system-system-of-records.

62 *DHS Exempts Dossiers Used for "Targeting" From the Privacy Act*, Papers Please!: The Identity Project, https://papersplease.org/wp/2010/02/08/dhs-exempts-dossiers-used-for-targeting-from-the-privacy-act/ (last visited Apr. 7, 2019).

63 Risk Prediction Program, *supra* note 29.

64 Risk Prediction Program Fact Sheet, DHS Science & Tech. Directorate (2014), https://www.dhs.gov/publication/risk-prediction-program.

65 Risk Prediction Program, *supra* note 29.

66 *About Us: Christopher M. Boner*, METRON: SCI. SOLUTIONS, https://web.archive.org/web/20160319061341/http://www.metsci.com:80/About-Us/Management/Christopher-M-Boner (last visited Apr. 7, 2019).

67 *U.S. Customs and Border Protection Awards Contract to* UNISYS *to Help Agency Assess Potential Threats from Travelers and Cargo Crossing into U.S.*, UNISYS (May 9, 2018), https://www.unisys.com/offerings/security-solutions/news%20release/us-customs-border-protection-awards-unisys-contract-to-assess-threats.

68 *Linesight: Advanced Targeting Analytics Solution for Border Security*, UNISYS, https://www.unisys.com/offerings/industry-solutions/public-sector-industry-solutions/justice-law-enforcement-and-border-security-solutions/linesight (last visited Apr 7, 2019).

69 Shana Dines, *Interim Report on the Automated Targeting System: Documents Released through EFF's FOIA Efforts*, EFF (July 31, 2009), https://www.eff.org/wp/interim-report-automated-targeting-system-documents-released-through-effs-foia-efforts; *see also* U.S. DEP'T OF HOMELAND SEC., PRIVACY OFF., 2017 Data Mining Report to Congress (Oct. 2018), https://www.dhs.gov/sites/default/files/publications/2017-dataminingreport_0.pdf; U.S. DEP'T OF HOMELAND SEC., PRIVACY IMPACT ASSESSMENT UPDATE FOR THE AUTOMATED TARGETING SYSTEM (Jan. 13, 2017), https://www.dhs.gov/sites/default/files/publications/privacy-pia-cbp006-ats-december2018.pdf, [hereinafter Privacy Impact Assessment Update Automated Targeting].

70 *See About Us, supra* note 66.

71 2017 DATA MINING REPORT TO CONGRESS, *supra* note 69.

72 *Id*.

73 *Metron, Inc. Gov't Contract: HSHQDC12C00040*, USA SPENDING, https://www.usaspending.gov/#/award/24238797 (last visited Apr. 7, 2019).

74 *DHS Awards Virginia Company $200K to Begin Automated Machine Learning Prototype Test,* U.S. DEP'T OF HOMELAND SEC. (Aug. 20, 2018), https://www.dhs.gov/science-and-technology/news/2018/08/20/news-release-dhs-awards-va-company-200k-begin-automated.

75 *See* Privacy Act of 1974, 5 U.S.C. § 552a (2018); U.S. CUSTOMS & BORDER PROTECTION, Automated Targeting System, System of Records, 72 Fed. Reg. 43650, 43650 (Aug. 6, 2007) ("In the case of cargo and conveyances, this screening results in a risk assessment score. In the case of travelers, however, it does not result in a risk assessment score.").

76 PRIVACY IMPACT ASSESSMENT UPDATE AUTOMATED TARGETING (Jan. 13, 2017), *supra* note 69.

77 Dines, *supra* note 69 ("While the DHS resolutely denies using numerical 'scores' to assign risk assessments to passengers, there is evidence in certain documents released pursuant to our FOIA request that suggest otherwise. [A] letter from the Executive Director of National Targeting and Security to the Directors of Field Operations and Preclearance Operations (agency unknown) regarding the 'Requirements for Access to the Automated Targeting System—Passenger' . . [states] 'Authorized Customs and Border Protection employees can access risk-scored passenger information . . . .' In a Customs and Border Protection training presentation, titled 'Targeting in the Passenger Environment' . . . a redacted slide reveals the heading 'ATS Passenger Examples (cont.): Passenger Arriving Flights - Risk Scored.'").

78 *U.S. Customs and Border Protection Awards Contract to Unisys*, *supra* note 67.

79 *See* PATRICK J. GROTHER ET AL., NAT'L INST. OF STANDARDS & TECH., REPORT ON THE EVALUATION OF 2D STILL-IMAGE FACE RECOGNITION ALGORITHMS, NIST INTERAGENCY REPORT 7709 2 (Aug. 24, 2011), http://ws680.nist.gov/publication/get_pdf.cfm?pub_id=905968. *But see NIST Evaluation Shows Advance in Face Recognition Software's Capabilities*, NAT'L INST. OF STANDARDS & TECH. (Nov. 30, 2018), https://www.nist.gov/news-events/news/2018/11/nist-evaluation-shows-advance-face-recognition-softwares-capabilities (explaining recent reductions in facial recognition error rates).

80 *Compare* Aaron Holmes, *These Clothes Use Outlandish Designs to Trick Facial Recognition Software into Thinking You're Not a Human*, BUS. INSIDER (Oct 12, 2019, 7:59 AM), https://www.businessinsider.com/clothes-accessories-that-outsmart-facial-recognition-tech-2019-10 (listing various masks, clothing, and accessories that can evade facial recognition) *with* Russell Brandom, *Your Phone's Biggest Vulnerability Is Your Fingerprint*, THE VERGE (May 2, 2016, 8:00 AM), https://www.theverge.com/2016/5/2/11540962/iphone-samsung-fingerprint-duplicate-hack-security (listing various ways to spoof a fingerprint reader, including silicon fingerprint molds).

81 *See* BIOMETRIC PATHWAY, *supra* note 15, at 2.

82 Test SUMMARY REPORT, *supra* note 38, at 7.

83 *See, e.g.,* Anirban Chakraborty et al., *Adversarial Attacks and Defences: A Survey*, CORNELL U. (Sept. 28, 2018), https://arxiv.org/abs/1810.00069; Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy, *Explaining and Harnessing Adversarial Examples*, ICLR '15 (Dec. 2014).

84 Alexey Kurakin, Ian Goodfellow & Samy Bengio, *Adversarial Examples in the Physical World,* CORNELL U. (July 8, 2016), https://arxiv.org/abs/1607.02533.

85 Evan Ackerman, *Three Small Stickers in Intersection Can Cause Tesla Autopilot to Swerve into Wrong Lane*, IEEE SPECTRUM (Apr. 1, 2019), https://spectrum.ieee.org/cars-that-think/transportation/self-driving/three-small-stickers-on-road-can-steer-tesla-autopilot-into-oncoming-lane.

86 Nicholas Carlini & David Wagner, *Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods*, CORNELL U. (Nov. 1, 2017), https://arxiv.org/abs/1705.07263.

87 *See, e.g.,* Aditi Raghunathan et al., *Certified Defenses Against Adversarial Examples,* CORNELL U. (Jan. 29, 2018), https://arxiv.org/abs/1801.09344.

88 The VA can, for example, integrate facial recognition with its new VA.gov benefits portal. Adam Mazmanian, *VA.gov Relaunches as Front Door to Benefits, Services*, FCW (Nov. 12, 2018), https://fcw.com/articles/2018/10/02/redesigned-va-site-sammie.aspx ("Mobile users won't be able to use their fingerprint or facial recognition logins that are built into devices, but that is in the offing.").

89 CTRS. FOR MEDICARE & MEDICAID SERVS., REPORT TO CONGRESS: FRAUD PREVENTION SYSTEM SECOND IMPLEMENTATION YEAR ii (2014), https://www.cms.gov/About-CMS/Components/CPI/Widgets/Fraud_Prevention_System_2ndYear.pdf.

90 U.S. DEP'T. OF HOMELAND SEC., PRIVACY IMPACT ASSESSMENT FOR THE TRAVELER VERIFICATION SYSTEM: DHS/CBP/PIA-056 (Nov. 14, 2018), https://www.dhs.gov/sites/default/files/publications/PIA%20for%20Traveler%20Verification%20Service.pdf [hereinafter Privacy Impact Traveler].

91 *See* Vishra Patel, *Airport Passenger Processing Technology: A Biometric Airport Journey*, EMBRY-RIDDLE AERONAUTICAL U. SCHOLARLY COMMONS (Apr. 2018), https://commons.erau.edu/cgi/viewcontent.cgi?article=1384&context=edt.

92 *See* Global Info. Services, Privacy Impact Assessment: Consular Consolidated Database (2018), https://www.state.gov/documents/organization/242316.pdf ("The CCD stores information about U.S. citizens and legal permanent residents (hereafter 'U.S. persons'), as well as foreign nationals (hereafter 'non-U.S. persons') such as nonimmigrant and immigrant visa applicants. The PII in CCD includes, but is not limited to: Names, Home/business addresses, Birthdates, Biometric data (fingerprints and facial images), Arrests and convictions, Social media indicators.").

93 *See* Mike Levine & Justin Fishel, *Security Gaps Found in Massive Visa Database*, ABC News (Mar. 31, 2016), https://abcnews.go.com/US/exclusive-security-gaps-found-massive-visa-database/story?id=38041051 ("State Department documents describe CCD as an 'unclassified but sensitive system.' Connected to other federal agencies like the FBI, Department of Homeland Security and Defense Department, the database contains more than 290 million passport-related records, 184 million visa records and 25 million records on U.S. citizens overseas."); see also Alba, *supra* note 36 ("Customs and Border Protection took images from the State Department that were submitted to obtain a passport and decided to use them to track travelers in and out of the country.").

94 Privacy Impact Traveler, *supra* note 90, at 16.

95 For the ADIS program alone, CBP shares data with the following non-DHS entities: the U.S. Department of State Bureau of Consular Affairs, the Federal Bureau of Investigation, the Social Security Administration, and the U.S. Intelligence Community. U.S. Dep't of Homeland Sec. & U.S. Customs & Border Protection, Pub. No. DHS/CBP/PIA-024(b), Privacy Impact Assessment for the Arrival and Departure Information System (ADIS) 5, 30-32 (2017), https://www.dhs.gov/sites/default/files/publications/privacy-pia-cbp024b-adis-april2017.pdf. This system does not itself run algorithms to discover a predictive pattern or an anomaly. *See Arrival & Departure Information System,* U.S. Dep't. of Homeland Sec. (Feb. 21, 2019), https://www.dhs.gov/publication/arrival-and-departure-information-system.

96 Memorandum of Understanding, *supra* note 46, at 6.

97 Alba, *supra* note 36.

98 See, *e.g.,* Document Production in Response to the Electronic Privacy Information Center's 2017 FOIA Request, https://epic.org/foia/dhs/cbp/afi/14-04-08-CBP-FOIA-20150205-Production-p1.pdf (last visited Dec. 13, 2019).

99 *See, e.g., Security Policies*, Dep't of Homeland Sec., https://www.dhs.gov/publication/security-training-contract-policy (last visited Apr. 7, 2019); Dep't of Homeland Sec., DHS Sensitive Systems Policy Directive 4300A 2 (2017), https://www.dhs.gov/sites/default/files/publications/Sensitive%20Systems%20Policy%20Directive%204300A.pdf.

100 Jason Kelley, *Skip the Surveillance by Opting out of Face Recognition at Airports,* EFF (Apr. 24, 2019), https://www.eff.org/deeplinks/2019/04/skip-surveillance-opting-out-face-recognition-airports ("These questions [about whether travelers can opt out] should be simple to answer, but we haven't gotten simple answers.").

101 *See, e.g.,* Jacob Snow, *Amazon's Face Recognition Falsely Matched 28 Members of Congress with Mugshots*, Am. Civil Liberties Union (July 26, 2018), https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28; Joy Boulamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 Proc. Machine Learning Res. 77 (2018).

102 Inioluwa Deborah Raji & Joy Buolamwini, *Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products,* Conf. on Artificial Intelligence, Ethics & Soc'y (2019), https://dam-prod.media.mit.edu/x/2019/01/24/AIES-19_paper_223.pdf.

103 Amazon's facial recognition service, for example, "made no errors in recognizing the gender of lighter-skinned men [but] it misclassified women as men 19 percent of the time . . . and mistook darker-skinned women for men 31 percent of the time." Natasha Singer, *Amazon Is Pushing Facial Technology that a Study Says Could Be Biased*, N.Y. Times (Jan. 24, 2019), https://www.nytimes.com/2019/01/24/technology/amazon-facial-technology-study.html.

104 *See* Jon Fingas, *Chinese Facial Recognition System Confuses Bus Ad with a Jaywalker*, Engadget (Nov. 22, 2018), https://www.engadget.com/2018/11/22/chinese-facial-recognition-confuses-bus-ad-with-jaywalker/.

105 Alexandra Chouldechova & Aaron Roth, *The Frontiers of Fairness in Machine Learning,* Cornell U. (Oct. 20, 2018), https://arxiv.org/abs/1810.08810.

106 Raji & Boulamwini, *supra* note 102, at 4.

107 *See* Dines, *supra* note 69 ("The data reviewed under the ATS system includes seven large government databases, plus the Passenger Name Record data from the airlines (which includes data like whether you've ordered a Muslim or Hindu or Jewish special meal."); *see also* Jennifer Lynch, *HART: Homeland Security's Massive New Database Will Include Face Recognition, DNA, and Peoples' "Non-Obvious Relationships,"* Elec. Frontier Found. (June 7, 2018), https://www.eff.org/deeplinks/2018/06/hart-homeland-securitys-massive-new-database-will-include-face-recognition-dna-and.

108 If they use a suspect classification, risk prediction tools may implicate the Equal Protection Clause of the Fourteenth Amendment. The risk prediction algorithm held to be constitutional by the Wisconsin Supreme Court in *Loomis* considered gender, but the plaintiff did not challenge this use of gender under the Equal Protection Clause. *See State v. Loomis*, 881 N.W.2d 749, 766 (Wis. 2016). Scholars who have analyzed the case remain divided on whether such a challenge would have been successful. *Compare* John Lightbourne, *Damned Lies and Criminal Sentencing Using Evidence-Based Tools*, 15 Duke L. & Tech. Rev. 327, 337 (2017) (arguing that the Equal Protection Clause should provide a remedy) *with* Leah Wisser, *Pandora's Algorithmic Black Box: The Challenges of Using Algorithmic Risk Assessments in Sentencing*, 56 Am. Crim. L. Rev. 1811 (2019) ("The Equal Protection Clause of the Fourteenth Amendment does not adequately address the problematic nature of algorithmic risk assessments.").

109 *See U.S. v. Flores-Montano*, 541 U.S. 149, 152-53 (2004) (noting that the government interest in controlling entry is so strong at the border that searches are "reasonable simply by virtue of the fact that they occur at the border" (quoting *U.S. v. Ramsey,* 431 U.S. 606, 616 (1977)).

110 Davey Alba, *These Senators Want Homeland Security to "Pause" its Airport Facial Recognition Program*, *Buzzfeed* (Mar. 12, 2019), https://www.buzzfeednews.com/article/daveyalba/these-senators-want-homeland-security-to-pause-its-facial.

111 *Id*.

112 Harrison Rudolph et al., *Not Ready for Take off: Face Scans at Airport Departure Gates*, Geo. L. Ctr. on Privacy & Tech. (Dec. 21, 2017), https://www.airportfacescans.com/.

113 5 U.S.C. § 553 (2018).

114 *See* Elec. Privacy Info. Ctr. v. Dep't of Homeland Sec., 653 F.3d 1, 8 (D.C. Cir. 2011).

115 5 U.S.C. § 553(b)(3)(B).

116 *Id.* § 553(a)(1).

117 Exec. Order No. 13,780, 82 Fed. Reg. 13209 (Mar. 6, 2017) ("The Secretary of Homeland Security shall expedite the completion and implementation of biometric entry exit tracking system for in-scope travelers to the United States, as recommended by the National Commission on Terrorist Attacks Upon the United States.").

118 5 U.S.C. § 553(b)(3)(B).

Endnotes to *Part II. Case Studies: Formal Adjudication at the Social Security Administration*

1 Jerry L. Mashaw, Bureaucratic Justice: Managing Social Security Disability Claims (1985); David Ames, Cassandra Handan-Nader, Daniel E. Ho & David Marcus, *Due Process and Mass Adjudication: Crisis and Reform*, 72 STAN. L. REV. (forthcoming 2020); 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).

2 Ames et al., *supra* note 1.

3 5 U.S.C. § 554 (2018) (applying formal adjudicatory provisions for adjudications required to be "on the record after opportunity for an agency hearing").

4 Michael Asimow, Admin. Conference of the U.S., Evidentiary Hearings Outside the Administrative Procedure Act (2016).

5 5 U.S.C. §§ 554, 556, 557.

6 *Id.*

7 Historically, ALJs were appointed by the agency through a competitive selection process run by the Office of Personnel Management (OPM). In Lucia v. Securities and Exchange Commission, 138 S. Ct. 2044 (2018), the Supreme Court found that ALJs in the SEC were inferior officers under the appointments clause of the Constitution, hence requiring appointment by the President or the SEC. Executive Order 13,843, in turn, provided that ALJs be placed under the Excepted Service, eliminating the competitive selection process. Exec. Order. No. 13,843: Excepting Administrative Law Judges From the Competitive Service, 83 Fed. Reg. 32,755 (July 13, 2018). ALJs are subject to removal only for "for good cause" through a formal hearing in front of the Merit Systems Protection Board (MSPB). 5 U.S.C. § 7521 (2018).

8 Asimow, *supra* note 4; Paul R. Verkuil, *Reflections Upon the Federal Administrative Judiciary*, 39 UCLA L. Rev. 1341 (1991).

9 *Schemes*, Adjudication Res., ACUS & Stan. L.S., http://acus.law.stanford.edu/schemes (last visited Nov. 9, 2019).

10 Roughly 92 Veterans Law Judges sit in the Board of Veterans Appeals, part of the Department of Veterans Affairs (VA), and hear appeals of benefits determinations by the VA. VLJs are appointed by the President, paid according the ALJ salary scale and removable under the same MSPB mechanism as for ALJs for cause. 31 U.S.C. §7101A.

11 Some 390 Immigration Judges work for the Executive Office of Immigration Review in the Department of Justice and decide immigration cases. They are appointed by the Attorney General and can be removed for poor performance reviews. *Immigration Judge*, Dep't of Justice (June 9, 2017), https://www.justice.gov/legal-careers/job/immigration-judge; 4 Executive Off. for Immigration Review & Nat'l Ass'n of Immigration Judges, Labor Agreement Between the National Association of Immigration Judges and U.S. Dep't of Justice, Executive Off. for Immigration Review Art. 22.6.1 (2016); *id*. at Ar. 22.2.

12 Asimow, *supra* note 4, at 18.

13 The SSDI program is established in Title II of the Act. 42 U.S.C. § 401ff (2018).

14 The SSI program is established in Title XVI of the Act. 42 U.S.C. § 1381ff (2018).

15 42 U.S.C. §§ 423(d)(1)(A), 1382c(a)(3)(A) (2018).

16 *See* 20 C.F.R. §§ 404.1520, 416.920 (2019). *See also Bowen v. Yuckert*, 482 U.S. 137 (1987) (approving the sequential process even though it might preclude consideration of the severity of an impairment in light of age and work history).

17 20 C.F.R. §§ 404.1520(a)(4)(i), 416.920(a)(4)(i).

18 20 C.F.R. §§ 404.1520(a)(4)(ii), 416.920(a)(4)(ii).

19 20 C.F.R. §§ 404.1520(a)(4)(iii), 416.920(a)(4)(iii). The conditions are listed in 20 C.F.R. § 404(P) app. 1. The Listings describe for each major body system impairments considered severe enough to satisfy requirements under the Social Security Act; they are updated from time to time to reflect progress in medical knowledge.

20 20 C.F.R. §§ 404.1520(a)(4)(iv), 416.920(a)(4)(iv). For discussion of the residual functional capacity determination, made using the grid regulations, and the issues associated with it, see, for example Heckler v. Campbell, 461 U.S. 458 (1983); *Sykes v. Apfel*, 228 F.3d 259 (3d Cir. 2000). The SSA's Occupational Information System reflects current occupations and their requirements.

21 20 C.F.R. §§ 404.1520(a)(4)(v), 416.920(a)(4)(v). If the agency finds that the claimant can perform other work, it must provide evidence that such other work as it asserts the claimant can perform exists in the national economy. 20 C.F.R. § 404.1560(c)(2). The agency has determined that such a showing is not required where the claimant is shown to be able to perform her past relevant work, a determination upheld by the Court. See Barnhart v. Thomas, 540 U.S. 20 (2003). The regulations also provide for several situations in which older workers are categorically unable to adjust to alternative work. *See, e.g.,* 20 C.F.R. § 404.1562.

22 *Barnhart*, 540 U.S. at 28-29.

23 Soc. Sec. Admin., FY 2018 Congressional Justification (2017).

24 *Hearing on Examining Changes to Social Security's Disability Appeals Process Before the Subcomm. On Soc. Sec'y of the H. Comm. On Ways & Means*, 115th Cong. (2018) (statement of Patricia Jonas, Deputy Comm'r, Office of Analytics, Review, and Oversight, Social Security Administration).

25 Between 2007 and 2015, concerted efforts to address the problem reduced the average processing time from 512 to 450 days but did not reduce the number of pending cases, which increased from 743,800 to 1 million. *See* Off. of the Inspector Gen., Soc. Sec. Admin., Audit Report A-12-15-15005, The Social Security Administration's Efforts to Eliminate the Hearings Backlog (Sept. 2015).

26 *See, e.g.,* U.S. Gov't Accountability Office, GAO-18-37, Social Security Disability: Additional Measures and Evaluation Needed to Enhance Accuracy and Consistency of Hearings Decisions (2017); *see also* Harold J. Krent & Scott Morris, *Inconsistency and Angst in District Court Resolution of Social Security Disability Appeals,* 67 Hastings L.J. 367 (2016).

27 *Compare* Ass'n of Admin. Law Judges, Inc. v. Heckler, 594 F. Supp. 1132 (D.D.C. 1984) (finding Bellmon Review program invalid but declining to provide injunctive relief because the SSA had rolled back the program), with *Ass'n of Admin. Law Judges, Inc. v. Colvin*, 777 F.3d 402 (7th Cir. 2015) (holding that requirement to hear a particular number of cases without regard to outcome did not violate the APA).

28 Jonah B. Gelbach & David Marcus, Admin. Conference of the U.S., A Study of Social Security Disability Litigation in the Federal Courts 47-48 (2016); Ames et al., *supra* note 1; Jonah B. Gelbach & David Marcus, *Rethinking Judicial Review of High Volume Agency Adjudication,* 96 Tex. L. Rev. 1097 (2018); Paul Verkuil, *Meeting the Mashaw Test for Consistency in Administrative Adjudication*, in Administrative Law from the Inside Out: Essays on Themes in the Work of Jerry L. Mashaw 239 (Nicholas R. Parrillo ed., 2017).

29 Felix F. Bajandas & Gerald K. Ray, Admin. Conference of the U.S., Implementation and Use of Electronic Case Management Systems in Federal Agency Adjudication (2018).

30 After early experimentation, SSA established the Analytics Center of Excellence ("ACE") in 2015 to provide an institutional knowledge base for the agency in developing technical solutions to its core challenges. ACE seeks to "nurture and promote a culture of evidence-based policies and decision-making across the agency" by talent management, training, and collaboration with business owners and outside technology experts. ACE has authority to develop a phased hiring approach, hiring both external candidates skilled in data science and internal candidates with sufficient institutional knowledge to help implement SSA's technical initiatives. ACE developed a four-week "comprehensive training package" for all ACE analysts including training for institutional knowledge, analytical

techniques, and specific programs. SSA also developed a formal analytics training program, the Gerald Ray Academy ("GRA"). The GRA trains staff on techniques to conduct data analysis and provides practical experience with applied analytical techniques. See Soc.Sec. Admin., Open Government Plan 4.0 (2016).

31 Bajandas & Ray, *supra* note 29, at 47. Bajandas describe a probit analysis to "identify cases with similar characteristics without first reviewing the records." *Id.* Our interviews suggest that the clustering tool in operation is more likely to be a form of unsupervised learning.

32 *Id.* at 47-48.

33 *See* Administrative Review Process for Adjudicating Initial Disability Claims, 71 Fed. Reg. 16,424 (Mar. 31, 2006) (to be codified at 20 C.F.R. §§ 404, 405, 416, 422).

34 *Id.* at 16,429.

35 *Id*.

36 *Id.* at 16,430. QDD is not available for applications submitted entirely on paper. *Program Operations Manual System (POMS):DI 23022.030,* Soc. Sec. Admin. (Apr. 6, 2018), https://secure.ssa.gov/apps10/poms.nsf/lnx/0423022030. Because nearly all, if not all, initial applications are submitted electronically, either by the claimant or by the office when a claimant applies in person or by phone, few to no applications are submitted entirely on paper at this point.

37 *Id.* ("The predictive model will not necessarily identify specific conditions. Instead, as described above, it will consider a variety of factors, including medical history, treatment protocols, and medical signs and findings.").

38 *Id.* Compare the Compassionate Allowances ("CAL") program, which identifies claimants with any of 225 conditions, including certain cancers, and allows for a disability determination within days. The CAL selection software "identifies cases for CAL processing based solely on the claimant's alleged medical condition(s) listed on the SSA-3368 (Disability Report––Adult) or SSA-3820 (Disability Report––Child). If the claimant alleges a medical condition (by name, synonym, or abbreviation) that is on the CAL list, the selection software identifies the case for CAL processing." *Program Operations Manual System (POMS):DI 23022.010,* Soc. Sec. Admin. (Apr. 6, 2018), https://secure.ssa.gov/apps10/poms.nsf/lnx/0423022010.

39 *See* 20 CFR §§ 404.1619, 416.1019 (2018).

40 *Id.* As late as 2012, an OIG report recommended that SSA "implement a program to automate the initial disability claim decision that would only require human review for denied claims," suggesting that even at that date QDD had not been implemented. *See* Off. of the Inspector Gen., Soc. Sec. Admin, Evaluation Report A-14-12-11222, The Social Security Administration's Implementation of the Future Systems Technology Advisory Panel's Recommendations (2012).

41 Frank S. Bloch et al., *The Social Security Administration's New Disability Adjudication Rules: A Significant and Promising Reform*, 92 Cornell L. Rev. 235, 238 (2007).

42 Bajandas & Ray, *supra* note 29, at 48.

43 *Id*.

44 Interview with Gerald K. Ray, former Admin. Appeals Judge & Deputy Exec. Dir., Office of Appellate Operations, Soc. Sec. Admin. (Oct. 23, 2018) (on file with authors).

45 Gerald K. Ray, Presentation at "A Roundtable Discussion on the Use of Artificial Intelligence in the Federal Administrative Process," NYU School of Law (Feb. 25, 2019)

46 Interview with Kurt Glaze, Program Analyst, Analytics Ctr. Of Excellence, Soc. Sec. Admin. (Oct. 24, 2018) (on file with authors).

47 *Id*.

48 Ames et al., *supra* note 1.

49 Soc. Sec. Admin., Updated Compassionate and Responsive Service (CARES) and Anomaly Plan (2017).

50 Office of the Inspector Gen., Soc. Sec. Admin., Audit Report A-12-18-50353, The Social Security Administration's Use of Insight Software to Identify Potential Anomalies in Hearing Decisions (2019). [hereinafter "Insight Audit Report"]

51 Interview with Kurt Glaze, *supra* note 46.

52 *Id*.

53 Jonas, *supra* note 24.

54 Gerald 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, 1593 (2015).

55 *Id*.

56 Interview with Kurt Glaze, *supra* note 46.

57 Insight Audit Report *supra* note 50.

58 We set aside here some statutory constraints that may apply to specific agencies. The BVA, for instance, is required to proceed in docket order, limiting the applicability of certain case triage systems.

59 Interview with Jae Song, Economist, Division of Econ. Research, Soc. Sec. Admin. (Nov. 16, 2018) (on file with authors).

60 Robin E. Kobayashi, SSA's Proposal to Replace the Outdated Dictionary of Occupational Titles (2009).

61 Beaulieu-Joneset al., *Privacy-Preserving Generative Deep Neural Networks Support Clinical Data Sharing*, Cardiovascular Quality & Outcomes 12.7 (2019).

62 *BERT Explained: A List of Frequently Asked Questions,* Let the Machines Learn (June 12, 2019), https://yashuseth.blog/2019/06/12/bert-explained-faqs-understand-bert-working/.

63 Ho et al., *supra* note 1.

64 Insight Audit Report, *supra* note 50, at 11.

65 James Ridgeway, Presentation at "A Roundtable Discussion on the Use of Artificial Intelligence in the Federal Administrative Process," NYU School of Law (Feb. 25, 2019)

66 *See* Citron, *supra* note 5, at 1261.

67 We acknowledge that SSA's decision tree aimed primarily to capture existing policy. But to the extent that such rules-based decision trees are important for deploying AI-based systems in other forms of adjudication, the system may shift to being more rules-based.

68 *See Heckler v. Campbell*, 461 U.S. 458 (1983).

69 *See, e.g., K.W. v. Armstrong*, 789 F.3d 962, 967-68, 971-74 (9th Cir. 2015); Ark. Dep't of Human Servs. v. Ledgerwood, 530 S.W.3d 336 (Ark. 2017).

70 *See* 5 U.S.C. § 553(a)(2) (2018).

71 *See Heckler v. Campbell*, 461 U.S. 458 (1983).

72 See Smolen v. Chater, 80 F.3d 1273, 1288 (9th Cir. 1996).

73 *See id.; Thompson v. Schweiker*, 665 F.2d 936, 941 (9th Cir. 1982) (quoting Gold v. Sec'y of Health Educ. & Welfare, 463 F.2d 38, 43 (2d Cir. 1972) (referring to a duty "to scrupulously and conscientiously probe into, inquire of, and explore for all the relevant facts").

74 *See, e.g., Tonapetyan v. Halter*, 242 F.3d 1144 (9th Cir. 2001); *Pratts v. Chater*, 94 F.3d 34, 37 (2d Cir. 1996).

75 Ames et al., *supra* note 1.

76 *See* David Freeman Engstrom & Daniel E. Ho, *Artificially Intelligent Government: A Review and Agenda*, in Big Data Law (Roland Vogl ed., forthcoming 2020)

Endnotes to *Part II. Case Studies: Informal Adjudication at the U.S. Patent and Trademark Office*

1 *See* 5 U.S.C. §§ 551(6), (7) (2018) (defining an adjudication as the process for formulating an order, which is a final disposition "in a matter other than rule making but including licensing").

2 Due to the variety and lack of standard procedures, it is challenging to report "case volume" for such decisions across the administrative state. Verkuil provided a rough estimate that 90% of federal agency adjudication is informal. Paul R. Verkuil, *A Study of Informal Adjudication Procedures*, 43 U. Chi. L. Rev. 739, 741 (1976). To be sure, Verkuil distinguished between formal adjudications under the APA and included Type B adjudications in informal adjudication. *Id.* This might mean that the 90% figure is an upper bound, but there is little rigorous data to support the fraction of all instances of adjudication that are informal. Id.

3 Michael Asimow, Admin. Conference of the U.S., Evidentiary Hearings Outside the Administrative Procedure Act 4 (2016). . We hence also implicitly adopt Asimow's typology: "The term 'informal adjudication' should be reserved for Type C adjudication in which decisions are not required to be based on evidentiary hearings." *Id.* at 3.

4 35 U.S.C. § 1 (2018) (establishing the USPTO). The trademark and patent laws governing the USPTO are codified in Chapter 22 of Title 15 of the United States Code (trademark) and Title 35 of the United States Code (patent). Trademark and patent regulations promulgated by the USPTO are codified in Title 37 of the Code of Federal Regulations, with Parts 1, 3, 4, 5, 11, 41, 42, and 90 specifically pertaining to patents, and Parts 2, 3, 6, 8, 10, and 11 specifically pertaining to trademarks.

5 35 U.S.C. § 131 authorizes the director of the USPTO to cause examination of patent applications, and 35 U.S.C. § 132 authorizes an examiner to reject a patent application if the examiner identifies certain deficiencies that would bar the application from being granted.

6 37 C.F.R. § 2.61(a) (2018) (requiring that the USPTO notify and advise applicants of the reasons for refusals of trademark applications); *Possible Grounds for Refusal of a Mark,* U.S. Patent & Trademark Off. (July 11, 2016), https://www.uspto.gov/trademark/additional-guidance-and-resources/possible-grounds-refusal-mark (describing example grounds for refusal of a mark, such as likelihood of confusion with respect to existing marks). 15 U.S.C. § 1062(a) authorizes the director of the USPTO to cause examination of trademark applications, and 15 U.S.C. § 1062(b) authorizes an examining attorney at the USPTO to refuse a trademark application if the examining attorney identifies certain deficiencies in the application that would bar the applicant from registering her trademark.

7 U.S. Patent & Trademark Off., FY 2018 Performance and Accountability Report 12, 32 (2018) [hereinafter FY 2018 Performance].

8 *See Alphabetical Index to Code*, U.S. Patent & Trademark Off. (Oct. 15, 2018), http://tess2.uspto.gov/tmdb/dscm/dsc_ai.htm.

9 *See CPC Scheme and Definitions*, Coop. Patent Classification, https://www.cooperativepatentclassification.org/cpcSchemeAndDefinitions.html (last visited Mar. 7, 2019).

10 *See* U.S. Patent & Trademark Off., Trademark Manual of Examining Procedure § 104 (2018) [hereinafter TMEP]; *Design Search Code Manual*, U.S. Patent & Trademark Off., http://tess2.uspto.gov/tmdb/dscm/index.htm (last visited Nov. 11, 2019).

11 *Design Search Code Manual*, *supra* note 10.

12 *See PUMA––Trademark Details*, Justia, https://trademarks.justia.com/854/79/puma-85479965.html.

13 Coop. Patent Classification, www.cooperativepatentclassification.org (last visited Feb. 23, 2019).

14 *Id*.

15 Coop. Patent Classification, Guide to the CPC (Cooperative Patent Classification) 4.0 § 3.1 (2017).

16 Jessica Manno, *A Day in the Life of a Patent Examiner: Searching,* U.S. Patent & Trademark Off. (May 3, 2018), https://www.uspto.gov/sites/default/files/documents/20180503_PPAC_Day_in_the_Life.pdf.

17 U.S. Patent & Trademark Off., Manual of Patent Examining Procedure § 902.03(e) (2018). The USPTO also provides other search tools as well, including the Patent Linguistic Utility Service (a word frequency-based search system). U.S. Patent & Trademark Off., PTOP-008-00, Privacy Impact Assessment for the Patent Search System––Primary Search & Retrieval (PSS-PS) System 2 (2018) [hereinafter Privacy Impact Assessment for the Patent Search System].

18 Privacy Impact Assessment, *supra* note 17, at 2; *Public Search Facility,* U.S. Patent & Trademark Off. (Dec. 7, 2018), https://www.uspto.gov/learning-and-resources/support-centers/public-search-facility/public-search-facility. Boolean Retrieval allows a user to "pose any query which is in the form of a Boolean expression of terms, that is, in which terms are combined with the operators AND, OR, and NOT." Christopher D. Manning et al., Introduction to Information Retrieval 4 (2008). This involves recording term-document matrices, often through an inverted index for efficiency, over the entire corpus. When a new query is issued, each term in the Boolean expression is looked up in an index, and the relevant intersection or union of documents constitute the retrieved set. *Id*.

19 TMEP, *supra* note 10, § 704.01.

20 *Id.* § 104.

21 *Id*.

22 *See generally Trademark Electronic Search System (TESS)*, U.S. Patent & Trademark Off. (Nov. 11, 2019), http://tess2.uspto.gov/.

23 U.S. Patent & Trademark Off., Manual of Patent Examining Procedure (MPEP) § 706.02(j) (9th ed. 2018) [hereinafter MPEP].

24 TMEP, *supra* note 10, § 704.01; *Introduction*, USPTO Design Search Code Manual (Oct. 15, 2018), http://tess2.uspto.gov/tmdb/dscm/index.htm#intro.

25 TMEP, *supra* note 10, § 705.

26 FY 2018 Performance, *supra* note 7, at 19.

27 Arti Rai, *Machine Learning at the Patent Office: Lessons for Patents and Administrative Law*, 104 Iowa L. Rev. 2617, 2619 (2019).

28 FY 2018 Performance, *supra* note 7, at 19.

29 See the initiatives described by the Office of Patent Quality Assurance. *About the Office of Patent Quality Assurance*, USPTO, https://www.uspto.gov/patent/office-patent-quality-assurance-0#step2 (last visited Dec. 13, 2019).

30 Michael D. Frakes & Melissa F. Wasserman, *Reconsidering Patent Examiner's Time Allocations,* Law360 (Oct. 13, 2016), https://www.law360.com/articles/850828/reconsidering-patent-examiner-s-time-allocations; *see also* 35 U.S.C. § 8 (2018) (authorizing the Director of the USPTO to "revise and maintain the classification by subject matter of United States letters patent, and such other patents and printed publications as may be necessary or practicable, for the purpose of determining with readiness and accuracy the novelty of inventions for which applications for patent are filed").

31 Andrew Chin, *Search for Tomorrow: Some Side Effects of Patent Office Automation*, 87 N.C. L. Rev. 1617, 1620 (2009).

32 FY 2018 Performance, *supra* note 7, at 56-67.

33 *Id.* at 60-61.

34 *Id.* at 58-59.

35 *Id.* at 95.

36 U.S. Patent & Trademark Off.ce, PTOC-016-00, Privacy Impact Assessment: USPTO Serco Patent Processing System (PPS) 1 (2018); *Serco Processes 4 Millionth Patent Application for U.S. Patent and Trademark Office,* Serco (Nov. 15, 2018), https://www.prnewswire.com/news-releases/serco-processes-4-millionth-patent-application-for-us-patent-and-trademark-office-300751330.html ("Since 2006, Serco has performed classification and other analysis services through awarded contracts including Pre-Grant

Publication (PGPubs) Classification Services, Initial Classification and Reclassification (ICR) Services, and Full Classification Services (FCS) contracts.") [hereinafter Serco Processes 4 Millionth Patent Application].

37 *Serco Processes 4 Millionth Patent Application*, *supra* note 36; *see* Cathy Weiss, *Artificial Intelligence: Challenges Presented by Patents*, SERCO (Dec. 26, 2018), https://sercopatentsearch.com/post?name=artificial-intelligence-challenges-presented-by-patents.

38 Modifications are not uncommon. In 2018 alone, the CPC issued 129 Notices of Change that added codes, removed codes, or revised classification rules. Weiss, *supra* note 37.

39 *Id*.

40 MPEP, *supra* note 23, §§ 719.05, 904 (describing that an examiner must make available to the applicant notes indicating the nature of her search).

41 Indeed, as inventors continue to innovate, non-standard terms may continue to creep into the lexicon of patents and patent applications, thus magnifying the drawbacks of Boolean search systems.

42 Arthi Krishna et al., *Examiner Assisted Automated Patents Search*, AAAI FALL SYMP. SERIES: COGNITIVE ASSISTANCE IN GOV'T & PUB. SECTOR APPLICATIONS 153, 153-54 (2016).

43 Rai, *supra* note 27, at 2626-37. Thomas A. Beach, *Japan Patent Information Organization Presentation: USPTO Bulk Data,* U.S. PATENT & TRADEMARK OFF. 22-25, http://www.japio.or.jp/english/fair/files/2016/2016e09uspto.pdf (last visited Feb. 23, 2019).

44 U.S. PATENT & TRADEMARK OFF., PATENT PUBLIC ADVISORY COMMITTEE QUARTERLY MEETING: IT UPDATE (2018).

45 *See Emerging Technologies in USPTO Business Solutions*, U.S. PATENT & TRADEMARK OFF. 18, https://www.wipo.int/edocs/mdocs/globalinfra/en/wipo_ip_itai_ge_18/wipo_ip_itai_ge_18_p5.pdf (last visited Nov. 11, 2019); Andrei Iancu, *Remarks by Director Iancu at 2018 National Lawyers Convention* U.S. PATENT & TRADEMARK OFF. (Nov. 15, 2018), https://www.uspto.gov/about-us/news-updates/remarks-director-iancu-2018-national-lawyers-convention.

46 Rai, *supra* note 27, at 2622-23.

47 Coding of Design Marks in Registrations, 75 Fed. Reg. 81,587-81,588 (Dec. 28, 2010), https://www.regulations.gov/document?D=PTO-T-2010-0090-0001.

48 *Id.*

49 *Id*.

50 *Id.*

51 *General Guidelines for Coding Design Marks*, U.S. PATENT & TRADEMARK OFF. (Oct. 11, 2013), http://tess2.uspto.gov/tmdb/dscm/dsc_gl.htm#generalglines.

52 *See Emerging Technologies in USPTO Business Solutions*, *supra* note 45, at 14.

53 *See id*. at 14.

54 *See id.* at 20.

55 *See id*. at 19-20.

56 The training objective in this case can be an autoencoding objective, for instance.

57 *See USPTO's Challenge to Improve Patent Search With Artificial Intelligence*, GOVTRIBE (Sept. 13, 2018, 4:29 PM), https://govtribe.com/opportunity/federal-contract-opportunity/uspto-s-challenge-to-improve-patent-search-with-artificial-intelligence-rfiptoaipatentseach.

58 *See* U.S. GOV'T ACCOUNTABILITY OFF., GAO-16-479, INTELLECTUAL PROPERTY: PATENT OFFICE SHOULD STRENGTHEN SEARCH CAPABILITIES AND BETTER MONITOR EXAMINERS' WORK 16 (2016).

59 *See id*. at 52.

60 So long as these terms have occurred in these contexts in the training data. Neural word embeddings are trained using a large amount of text that can originate from patent and nonpatent literature.

61 *See* U.S. GOV'T ACCOUNTABILITY OFF., *supra* note 58, at 25.

62 Attention maps can be used to determine such alignments.

63 *See* U.S. GOV'T ACCOUNTABILITY OFF., *supra* note 58, at 15.

64 Currently, the USPTO uses the Scientific and Technical Information Center (STIC) to collect human translations.

65 Rai, *supra* note 27, at 2638 ("To the extent that the AI-assisted search used by the Patent Office does not account for potentially rapid change in the average skill of practitioners itself spurred by AI, it will fall short.").

66 WIPO reported a 20% accuracy on their trademark classification test set. Christophe Mazenc, *Machine Learning Applied to Trademark Classification and Search*, WORLD INTELL. PROP. ORG. 10, 20, https://www.wipo.int/edocs/mdocs/globalinfra/en/wipo_ip_itai_ge_18/wipo_ip_itai_ge_18_p17.pdf.

67 *See generally* TMEP, *supra* note 10, § 1400.

68 There are many deep learning models that carry out this type of optical character recognition (OCR) in the wild.

69 The term "explainability" in the legal context and the term "explainability" in the computer science context may be co-extensive in purpose but otherwise are distinctive terms with different meanings. For example, an "explainable" algorithm in the computer science context need not necessarily be "explainable" in the context of administrative law, and vice versa.

70 5 U.S.C. § 555(e) (2017).

71 Arti Rai, *Machine Learning at the Patent Office: Lessons for Patents and Administrative Law* 1-2 (Dec. 29, 2018) (unpublished manuscript) (on file with authors).

72 MPEP, *supra* note 23, § 719.05.

73 *Id.* In fact, issued patents are legally presumed to be valid. 35 U.S.C. § 282(a) (2017).

74 *Id.* These search notes often include the CPC classifications searched as well as specific search terms and strategy. For example, an examiner could note the specific individuals whom she has spoken to, *e.g.,* other examiners. The examiner could also record the specific queries entered into the search system.

75 TMEP, *supra* note 10, § 710.02. These search terms could include characters, words, and design codes, and the search history could include several entries as well as a duration of time that the examining attorney spent searching.

76 *About Us,* PATENT OFF. PROF'L ASS'N, http://www.popa.org/forms/about-us/ (last visited Mar. 29, 2019).

77 THE NATIONAL TREASURY EMPLOYEES UNION: CHAPTER 245, http://www.nteu245.org/ (last visited Mar. 29, 2019).

78 Rai, *supra* note 71, at 16.

79 *Id*.

80 *See id*.

81 *See id*. (describing that the USPTO's previous Sigma tool appeared to be more effective for those with computer science backgrounds).

82 Megan McLoughlin, *A Better Way to File Patent Application*, IPWATCHDOG (Apr. 14, 2016), http://www.ipwatchdog.com/2016/04/14/better-way-file-patent-applications/id=68302/.

83 37 C.F.R. § 1.56 (2018).

84 *See Therasense, Inc. v. Becton, Dickinson & Co*., 649 F.3d 1276, 1285 (Fed. Cir. 2011).

85 15 U.S.C. § 1051(b)(3)(D) (2017); 37 C.F.R. § 2.33(b)(2); TMEP, *supra* note 10, § 804.02. Notably, trademark applicants only have an obligation not to defraud the USPTO, not an affirmative duty to disclose material information to the USPTO. *See generally* Susan M. Richey, *The Second Kind of Sin: Making the Case for a Duty to Disclose Facts Related to Genericism and Functionality in the Trademark Office*, 67 WASH. & LEE L. REV. 137, 140 n.7 (2010).

86 35 U.S.C. § 112 (2018). Section 112 requires "full, clear, concise, and exact terms" for a written description of the invention in an application as well as claims that "distinctly claim[] the subject matter which the inventor or a joint inventor regards as the invention." *Id*.

87 *Vendor Information*, U.S. PATENT & TRADEMARK OFF. (Feb. 29, 2019), https://www.uspto.gov/about-us/vendor-information.

88 48 C.F.R. § 9.504(a)(1)-(2) (2018).

89 *Id.* § 9.504.

Endnotes to *Part II. Case Studies: Regulatory Analysis at the Food and Drug Administration*

1 Maeve P. Carey, *Counting Regulations: An Overview of Rulemaking, Types of Federal Regulations, and Pages in the Federal Register*, Cong. Res. Serv. 2 (2019), https://fas.org/sgp/crs/misc/R43056.pdf.

2 *See Postmarketing Surveillance Programs*, U.S. Food & Drug Admin. (Nov. 17, 2016), https://www.fda.gov/drugs/surveillance/postmarketing-surveillance-programs.

3 A concrete example of the FDA's use of FAERS analysis to refine its regulation of infusion pumps. After receiving thousands of adverse event reports, the FDA launched an Infusion Pump Improvement Initiative, eventually imposing new requirements on infusion pump manufacturers and promulgating a new guidance document. *See White Paper: Infusion Pump Improvement Initiative*, U.S. Food & Drug Admin. (April 2010), https://www.fda.gov/medical-devices/infusion-pumps/white-paper-infusion-pump-improvement-initiative#causes; *Infusion Pumps Total Product Life Cycle: Guidance for Industry and FDA Staff*, U.S. Food & Drug Admin. (Dec. 2, 2014), https://www.fda.gov/media/78369/download.

4 Robert Ball, *Why Is FDA Interested in Natural Language Processing (NLP) of Clinical Texts?: Applications to Pharmacovigilance and Pharmacoepidemiology,* U.S. Food & Drug Admin. 4 (June 15, 2017), https://pharm.ucsf.edu/sites/pharm.ucsf.edu/files/cersi/media-browser/Ball.pdf.

5 *See, e.g.,* Leihong Wu et. al., *A Deep Learning Model To Recognize Food Contaminating Beetle Species Based on Elytra Fragments*, 166 Computers & Elec. in Agric. 105002 (2019); Randy L. Self, Michael G. McLendon, Christopher M. Lock, *Determination of Decomposition in Salmon Products by Mass Spectrometry with Sensory-driven Multivariate Analysis*, 39 J. of Food Safety 5 (2019); Leihong Wu, Xiangwen Liu, Joshua Xu, HetEnc: *A Deep Learning Predictive Model for Multi-type Biological Dataset*, 20 BMC Genomics 638 (2019); Meng Hu et. al., *Predictive Analysis of First Abbreviated New Drug Application Submission for New Chemical Entities Based on Machine Learning Methodology*, 106 Clinical Pharmacology 1 (2019); Xiajing Gong, Meng Hu, Liang Zhao, *Big Data Toolsets to Pharmacometrics: Application of Machine Learning for Time-to-Event Analysis,* 11 Clinical Translational Sci. 3 (2018).

6 *See generally* David Lazer et al., *The Parable of Google Flu: Traps in Big Data Analysis*, 343 Sci. 1203 (2014); Kristen M. Altenburger & Daniel E. Ho, *When Algorithms Import Private Bias into Public Enforcement: The Promise and Limitations of Statistical De-biasing Solutions,* 175 J. Institutional & Theoretical Econ. 98 (2018).

7 *Fact Sheet: FDA at a Glance*, U.S. Food & Drug Admin., https://www.fda.gov/AboutFDA/Transparency/Basics/ucm553038.htm (last visited Apr. 4, 2019).

8 21 U.S.C. § 301 (2018).

9 21 U.S.C. §§ 355(k)(3)(b), (k)(3)(b)(ii)-(iii) ("The Secretary shall, not later than 2 years after the date of the enactment of the Food and Drug Administration Amendments Act of 2007 [enacted Sept. 27, 2007], in collaboration with public, academic, and private entities . . . develop validated methods for the establishment of a postmarket risk identification and analysis system to link and analyze safety data from multiple sources" and "convene a committee of experts, including individuals who are recognized in the field of protecting data privacy and security, to make recommendations to the Secretary on the development of tools and methods for the ethical and scientific uses for, and communication of, postmarketing data . . . including recommendations on the development of effective research methods for the study of drug safety questions.").

10 *See* Lars Noah, *Governance by the Backdoor: Administrative Law(lessness?) at the FDA*, 93 Neb. L. Rev. 89, 90 (2014) (noting the FDA's "shift from the promulgation of binding rules to the issuance of nonbinding guidance documents"); *Search for FDA Guidance Documents*, U.S. Food & Drug Admin. (Nov. 8, 2019), https://www.fda.gov/regulatory-information/search-fda-guidance-documents; *see also* Michael S. Greve & Ashley C. Parrish, *Administrative Law Without Congress*, 22 Geo. Mason L. Rev. 501, 532 (2015); K.M. Lewis, *Informal Guidance and the FDA*, 66 Food & Drug L.J. 507, 509 (2011). In the first decades after the passage of the FDCA, there was some uncertainty over the level of formality required by the rulemaking procedures provided by the FDCA, in particular in applying the "formal rulemaking" provisions of the Administrative Procedure Act (APA). *See* Noah, *supra*, at 94-95. By the mid-1990s, the agency had largely turned away from rulemaking in favor of issuing non-binding guidance documents, at least in part to avoid more arduous rulemaking procedures. *See id.* at 90-92.

11 *See About Warning and Close-Out Letters*, U.S. Food & Drug Admin. (Apr. 29, 2019), https://www.fda.gov/ICECI/EnforcementActions/WarningLetters/ucm278624.htm (last visited Nov. 21, 2018).

12 *See generally Regulatory Procedures Manual*, U.S. Food & Drug Admin. (Dec. 12, 2017), https://www.fda.gov/iceci/compliancemanuals/regulatoryproceduresmanual/default.htm.

13 *See e.g.,* Lars Noah, *The Little Agency that Could (Act with Indifference to Constitutional and Statutory Strictures)*, 93 Cornell L. Rev. 901, 924 (2008) (noting that FDA has "struggl[ed] to protect the public health with its limited statutory powers and often inadequate resources"); Lewis, *supra* note 10, at 538 ("[The] FDA operates under severe resource constraints."); Nicholas R. Parrillo, Admin. Conf. of the U.S., Federal Agency Guidance: An Institutional Perspective 53 (2017) (noting an interviewee who described FDA as "resource-constrained").

14 *See, e.g., The FDA's Drug Review Process: Ensuring Drugs Are Safe and Effective,* U.S. Food & Drug Admin. (Nov. 24, 2017), https://www.fda.gov/Drugs/ResourcesForYou/Consumers/ucm143534.htm.

15 *See, e.g., The FDA's Drug Review Process: Continued*, U.S. Food & Drug Admin. (Aug. 24, 2015), https://www.fda.gov/Drugs/ResourcesForYou/Consumers/ucm289601.htm.

16 *See, e.g., Generic Drugs: Questions & Answers*, U.S. Food & Drug Admin. (June 1, 2018), https://www.fda.gov/Drugs/ResourcesForYou/Consumers/QuestionsAnswers/ucm100100.htm; *Abbreviated New Drug Application (ANDA)*, U.S. Food & Drug Admin., https://www.fda.gov/drugs/developmentapprovalprocess/howdrugsaredevelopedandapproved/approvalapplications/abbreviatednewdrugapplicationandagenerics/default.htm (last visited Feb. 17, 2019).

17 *See, e.g., A History of Medical Device Regulation & Oversight in the United States*, U.S. Food & Drug Admin. (June 24, 2019), https://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/Overview/ucm618375.htm.

18 *Id.*

19 *See e.g.,* 21 U.S.C. § 355(k)(3)(B)(ii) (2018) ("The secretary shall . . . develop validated methods for the establishment of a postmarket risk identification and analysis system to link and analyze safety data from multiple sources." Similarly, subsections 355(k)(3)(C)(i)(II), (IV) and (VI) state, "The Secretary shall . . . establish and maintain procedures . . . to identify certain trends and patterns with respect to data accessed by the system . . . [and] to enable the program to export data in a form appropriate for further aggregation, statistical analysis, and reporting." Subsection 355(k)(4)(A) states, "The Secretary shall . . . establish collaborations with public, academic, and private entities . . . to provide for advanced analysis of drug safety data described in paragraph (3)(C).") (emphasis added).

20 *See generally* U.S. Gov't Accountability Off., Report to Requesters: Drug Safety, Improvements Needed in FDA's Postmarket Decision-Making and Oversight Process (2006); *see also* Rebecca S. Eisenberg & W. Nicholson Price II, *Promoting Healthcare Innovation on the Demand Side*, 4 J.L. & Biosciences 3, 41-44 (2017) (providing a brief overview of the development of FDA's postmarket surveillance authority).

21 Eisenberg & Price, *supra* note 20 at 42 n.257.

22 *But see, e.g.,* Alison M. Pease et al., *Postapproval Studies of Drugs Initially Approved by FDA on the Basis of Limited Evidence: Systematic Review*, 357 British Med. J. 1680 (2017).

23 Megan Molteni, *Medicine Is Going Digital. FDA Is Racing to Catch Up*, WIRED (May 22, 2017), https://www.wired.com/2017/05/medicine-going-digital-fda-racing-catch/; *see also* Evan Sweeney, *FDA Begins Filling Positions for its New Digital Health Unit,* FierceHealthcare (Sept. 11, 2017, 11:31 AM), https://www.fiercehealthcare.com/regulatory/fda-begins-filling-positions-for-its-new-digital-health-unit.

24 U.S. Food & Drug Admin., Digital Health Action Plan 7 (2017), https://www.fda.gov/downloads/MedicalDevices/DigitalHealth/UCM568735.pdf.

25 U.S. Dep't of Health & Human Servs., U.S. Food & Drug Admin., Justification of Estimates for Appropriations Committees: Fiscal Year 2019 8-9 (2018).

26 *Id.; see also* Digital Health Action Plan, *supra* note 24.

27 Scott Gottlieb, *FDA's Comprehensive Effort to Advance New Innovations: Initiatives to Modernize for Innovation*, U.S. Food & Drug Admin. (Aug. 29, 2018), https://www.fda.gov/NewsEvents/Newsroom/FDAVoices/ucm619119.htm.

28 *Id.*

29 *Id.*

30 Other databases at the FDA could be used to exploit AI/ML tools. *See e.g., FDA's Sentinel Initiative,* U.S. Food & Drug Admin., https://www.fda.gov/Safety/FDAsSentinelInitiative/default.htm (last visited Nov. 24, 2018) (Sentinel is "the FDA's national electronic system which has transformed the way researchers monitor the safety of FDA-regulated medical products, including drugs, vaccines, biologics, and medical devices."); *FDA Officials Discuss Sentinel Challenges*, FDANews (Feb. 14, 2018), https://www.fdanews.com/articles/185596-fda-officials-discuss-sentinels-challenges ("The biggest challenge for the FDA's 10-year-old Sentinel . . . is dealing with a wide range of partners and data sources, according to CDER Deputy Director Robert Ball. Increased use of natural language processing and machine learning will help meet this challenge, Ball said. . . ."); *Validation of Anaphylaxis Using Machine Learning*, Sentinel Initiative (Oct. 11, 2018), https://www.sentinelinitiative.org/sentinel/methods/validation-anaphylaxis-using-machine-learning (discussing an ongoing "pilot project" which will use "natural language processing" and "machine learning" to "improve health outcome of interest (HOI) [in this case, anaphylaxis] detection algorithms that may later be used in the larger Sentinel Distributed Database").

31 *Postmarket Surveillance Programs*, U.S. Food & Drug Admin. (Nov. 17, 2016), https://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/ucm090385.htm.

32 *See* 21 C.F.R. § 314.80 (2018) (describing postmarket adverse event reporting requirements); *see also* Anne Tobenkin, U.S. Food & Drug Admin., An Introduction to Drug Safety Surveillance and FDA Adverse Event Reporting System 28 (Apr. 10, 2018), https://www.fda.gov/media/112445/download.

33 Tobenkin, *supra* note 32, at 25.

34 *Id.*

35 *Id.* at 23.

36 *Questions and Answers on FDA's Adverse Event Reporting System (FAERS)*, U.S. Food & Drug Admin. (June 4, 2018), https://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm.

37 Qais Hatim et al., U.S. Food & Drug Admin., Modeling Text Analysis to Empower FAERS Adverse Event Assessment 5-6 (2018); *see also* Qais Hatim et al., *Modeling and Text Analysis to Empower FAERS Adverse Event Assessment*, https://phusewiki.org/docs/2018_US%20Connect18/AB%20STREAM/AB09%20FINAL.pdf (last visited Dec. 6, 2019).

38 Hatim, *supra* note 37, at 8.

39 *Id.* at 9. After applying some custom filtering, the final number of labeled records came to 304,000.

40 *See id.* for the full details of the approach.

41 Text mining is an unsupervised method to obtain the frequencies of important terms across documents within a corpus, yielding a term-by-document matrix. The Project applied numerical analysis techniques to express the large matrix more compactly. *Id.* at 8. The terms with the highest frequencies were deemed not to add any value to the analysis (examples of such terms are "patient" and "liver"). *Id.* at 9.

42 Topic modeling looks at "collections of terms that describe and characterize a main theme or idea." *Id.* at 9.

43 *Id.* at 11-13. The FDA project incorporated expert knowledge by adding in custom topics. *Id.* at 9.

44 *Id.* at 13.

45 *Id.* at 18-20.

46 *See* Lichy Han et al., *Development of an Automated Assessment Tool for MedWatch Reports in the FDA Adverse Event Reporting System*, 24 J. Am. Med. Informatics Assoc. 913 (2017).

47 The area under the receiver operating characteristic curve was 0.66.

48 Hatim, *supra* note 37, at 27.

49 *Id.*

50 These terms were derived from the L1 regularized logistic regression approach.

51 As one FDA official noted, the FDA needs "regulatory grade data. It must be pristine; [but] NLP doesn't come anywhere close to that." Additionally, it, perhaps, bears repeating that classical rule-based techniques for NLP may be more suitable for generating causal hypotheses, but they do not have high performance on prediction tasks. On the other hand, machine learning techniques that benefit from the large FDA datasets, such as neural network methods, may successfully uncover patterns and statistical relationships in the data. But these are typically less interpretable, and causal inference is even more difficult than with more structured rule-based systems. Together, this can create something of a regulatory catch-22 where performance and understanding are inversely proportional.

52 Hatim, *supra* note 37, at 1, 22.

53 *Use of Natural Language Processing to Extract Information from Clinical Text*, U.S. Food & Drug Admin. (June 14, 2017), https://www.fda.gov/ScienceResearch/SpecialTopics/RegulatoryScience/ucm548651.htm.

54 *See* Han et al., *supra* note 46.

55 None of the interviewed officials at the FDA suggested that human capital was the barrier to successful NLP deployment at FDA. According to one FDA official, "We have the best AI researchers but . . . we're stuck [due to various practical barriers]." Another official noted that the FDA has very good fellows and that, although pay may be a barrier for some, there are individuals working at FDA, including career officials and those with statistical backgrounds, who have taken to AI/ML.

56 Hatim, *supra* note 37, at 22.

57 The critical importance of standardized or otherwise fit-for-purpose data is also something we heard from officials across a number of federal agencies.

58 *See, e.g.,* Scott Gottlieb & Jeffrey E. Shuren, *Statement from FDA Commissioner Scott Gottlieb, M.D. and Jeff Shuren, M.D., Director of the Center for Devices and Radiological Health, on FDA's Updates to Medical Device Safety Action Plan to Enhance Post-market Safety*, U.S. Food & Drug Admin. (Nov. 20, 2018), https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm626286.htm. Particularly in the device space, FDA is exploring ways to lower the premarket barrier while ratcheting up postmarket scrutiny. *See generally*, U.S. Food & Drug Admin., Developing a Software Precertification Program: A Working Model (2019), https://www.fda.gov/downloads/MedicalDevices/DigitalHealth/DigitalHealthPreCertProgram/UCM629276.pdf.

59 *Postmarket Surveillance Programs*, *supra* note 31; Tobenkin, *supra* note 32, at 14.

60 *See, e.g., The Public's Stake in Adverse Event Reporting*, U.S. Food & Drug Admin., https://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/ucm179586.htm (last visited Feb. 17, 2019) (describing postmarket safety monitoring as a "critical part of FDA's responsibilities").

61 In 2018, the FDA announced that "81 percent of open PMRs [post-marketing requirements] (1,056/1,298) and 76 percent of open PMCs [post-marketing commitments] (248/326) [were] progressing on schedule" and that the "FDA Center for Drug Evaluation and Research (CDER) completed review for 1,468 of the 1,553 PMRs and PMCs in the *backlog* and the FDA Center for Biologics Evaluation and Research (CBER) completed review of 71 of 83 PMRs and PMCs in the *backlog*." Scott Gottlieb, *Statement by FDA Commissioner Scott Gottlieb, M.D., on the FDA's Efforts to Hold Industry Accountable for Fulfilling Critical Post-marketing Studies of the Benefits, Safety of New Drugs*, U.S. Food & Drug Admin. (Nov. 16, 2018), https://www.fda.gov/news-events/press-announcements/statement-fda-commissioner-scott-gottlieb-md-fdas-efforts-hold-industry-accountable-fulfilling (emphasis added).

62 For example, devices that incorporate a learning component might need to be re-authorized by the FDA under the current regulatory regime each time they change. For that (and other) reasons, the FDA is working on designing an approval process that considers a company's track record and focuses heavily on postmarket surveillance, so products get to market faster. *See, e.g.,* U.S. Food & Drug Admin., Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)––Discussion Paper and Request for Feedback 3 (2019), https://www.fda.gov/downloads/MedicalDevices/DigitalHealth/SoftwareasaMedicalDevice/UCM635052.pdf?utm_campaign=2019-04-02%20Discussion%20Paper%20on%20Regulating%20Artificial%20Intelligence&utm_medium=email&utm_source=Eloqua ("The traditional paradigm of medical device regulation was not designed for adaptive AI/ML technologies, which have the potential to adapt and optimize device performance in real-time to continuously improve healthcare for patients.") [hereinafter Proposed Regulatory Framework for Modifications]. *See also* Developing a Software Precertification Program, *supra* note 58.

63 *See, e.g.,* Scott Gottlieb, *Remarks Before the Bipartisan Policy Center,* U.S. Food & Drug Admin. (Jan. 28, 2019), https://www.fda.gov/NewsEvents/Speeches/ucm629942.htm.

64 Viz.AI is "a computer-aided triage software that uses an artificial intelligence algorithm to analyze images for indicators associated with a stroke." *FDA Permits Marketing of Clinical Decision Support Software for Alerting Providers of a Potential Stroke in Patients,* U.S. Food & Drug Admin. (Feb. 13, 2018), https://www.fda.gov/newsevents/newsroom/pressannouncements/ucm596575.htm (Once the software analyzes CT images of the brain, the application then "send[s] a text notification to a neurovascular specialist if a suspected large vessel blockage has been identified.") [hereinafter Viz.AI].

65 *FDA Permits Marketing of Artificial Intelligence Algorithm for Aiding Providers in Detecting Wrist Fractures*, U.S. Food & Drug Admin. (May 24, 2018), https://www.fda.gov/newsevents/newsroom/pressannouncements/ucm608833.htm [hereinafter OsteoDetect].

66 *FDA Permits Marketing of Artificial Intelligence-based Device to Detect Certain Diabetes-related Eye Problems*, U.S. Food & Drug Admin. (Apr. 11, 2018), https://www.fda.gov/newsevents/newsroom/pressannouncements/ucm604357.htm (IDx-DR is "a software program that uses an artificial intelligence algorithm to analyze images of the eye taken with a retinal camera called the Topcon NW400.") [hereinafter IDx-DR].

67 E*valuation of Automatic Class III Designation (De Novo) Summaries*, U.S. Food & Drug Admin., http://www.fda.gov/about-fda/cdrh-transparency/evaluation-automatic-class-iii-designation-de-novo-summaries (last visited Nov. 26, 2019) ("The Food and Drug Administration Modernization Act of 1997 (FDAMA) added the De Novo classification option as an alternate pathway to classify novel medical devices that had automatically been placed in Class III after receiving a "not substantially equivalent" (NSE) determination in response to a premarket notification [510(k)] submission.").

68 *See* Viz.AI, *supra* note 64.

69 *Id.* ("The company submitted a retrospective study of 300 CT images that assessed the independent performance of the image analysis algorithm and notification functionality of the Viz.AI Contact application against the performance of two trained neuro-radiologists for the detection of large vessel blockages in the brain. Real-world evidence was used with a clinical study to demonstrate that the application could notify a neurovascular specialist sooner in cases where a blockage was suspected."); IDx-DR, *supra* note 66 (The FDA "evaluated data from a clinical study of retinal images obtained from 900 patients with diabetes at 10 primary care sites[, and concluded that] IDx-DR was able to correctly identify the presence of more than mild diabetic retinopathy 87.4 percent of the time and was able to correctly identify those patients who did not have more than mild diabetic retinopathy 89.5 percent of the time."); OsteoDetect, *supra* note 65 (OsteoDetect was approved based on "a retrospective study of 1,000 radiograph images that assessed the independent performance of the image analysis algorithm for detecting wrist fractures and the accuracy of the fracture localization of OsteoDetect against the performance of three board certified orthopedic hand surgeons. Imagen also submitted a retrospective study of 24 providers who reviewed 200 patient cases. Both studies demonstrated that the readers' performance in detecting wrist fractures was improved using the software, including increased sensitivity, specificity, positive and negative predictive values, when aided by OsteoDetect, as compared with their unaided performance according to standard clinical practice.").

70 *See, e.g.,* Proposed Regulatory Framework for Modifications, *supra* note 62.

71 Amy Abernathy, *Statement on New Steps to Advance Digital Health Policies That Encourage Innovation and Enable Efficient and Modern Regulatory Oversight*, U.S. Food & Drug Admin. (Sept. 26, 2019), https://www.fda.gov/news-events/press-announcements/statement-new-steps-advance-digital-health-policies-encourage-innovation-and-enable-efficient-and ("We believe that an appropriate regulatory framework that takes into account the realities of how technology advances plays a crucial role in the efficient development of digital health technologies . . . . As part of this plan, we've accomplished several key initiatives, including launching and testing the digital health software precertification pilot program ('Pre-Cert') and taking steps to modernize our policies.").

72 Conor Hale, *FDA Lays Out Plans for a New Review Framework for AI and Machine Learning-based Devices*, FierceBiotech (Apr. 3, 2019), https://www.fiercebiotech.com/medtech/fda-lays-out-plans-for-a-new-review-framework-for-ai-and-machine-learning-based-devices ("Those changes would require manual validation and verification of the updates . . . . The agency's future approach may require scrutinizing manufacturers' prespecified plans for modifications, including through algorithm retraining and updates, as well as their ability to manage and control the resulting risks.").

73 *Id.*

74 *See Information Exchange and Data Transformation (INFORMED)*, U.S. Food & Drug Admin. (July 12, 2018), https://www.fda.gov/aboutfda/centersoffices/officeofmedicalproductsandtobacco/oce/ucm543768.htm (The FDA's website contains an outline of INFORMED's "research portfolio.").

75 Sean Khozin et al., *INFORMED: An Incubator at the US FDA for Driving Innovation in Data Science and Agile Technology*, 17 Nature Rev. Drug Discovery 529, 529 (2018) [hereinafter Khozin, *Incubator*]; *see also Dr. Sean Khozin on FDA Initiative to Analyze Data from Real-World Pipelines*, Am. Journal Managed Care (Jan. 23, 2017), https://www.ajmc.com/interviews/dr-sean-khozin-on-fda-initiative-to-analyze-data-from-real-world-pipelines (INFORMED is "an oncology data science initiative, and it has [two] primary components . . . . [T]he first component of our program is aggregating and standardizing the clinical trial data into a common standard, and doing meta-analyses and predictive analytics . . . . The second component of the program is to start leveraging new pipelines of data outside of clinical trials, and there are several new pipelines of data that we're looking at. One of the main ones is data from electronic health records, real-world evidence, if you will . . . We're also looking at other pipelines of data, we are looking at the utility of biosensors and mobile sensor technologies to be able to better capture the patient's experience . . . . And other new pipelines of data include raw genomic data, and that's where you really start to go into big data analytics and data science.") [*Khozin on FDA,* Am. Journal Managed Care].

76 Am. *Khozin on FDA*, *supra* note 75; *see also* Khozin, *Incubator*, *supra* note 75, at 530 ("Our current objectives are twofold: first, to continue to expand and maintain organizational and technical infrastructure for data science and big data analytics; and second, to support systems thinking in oncology regulatory science research, with a focus on the development and utilization of novel solutions for improving efficiency, reliability and productivity in related workflows.").

77 *INFORMED: An Interview with Sean Khozin, MD, MPH*, Lung Cancer News (Dec. 19, 2017), http://www.lungcancernews.org/2017/12/19/informed-an-interview-with-sean-khozin-md-mph/ [hereinafter Lung Cancer News].

78 Khozin, *Incubator*, *supra* note 75, at 530.

79 Sean Khozin, *What is Regulatory Science?*, SK, https://www.seankhozin.com (last visited Dec. 6, 2019).

80 Khozin, *Incubator*, *supra* note 75, at 530 ("The success of INFORMED has been largely the result of an entrepreneurial and collaborative model and the active building of an interactive community of government, academic, non-profit and industry partners."); *see also* Lung Cancer News, *supra* note 77. ("As an incubator, INFORMED conducts collaborative research with innovators in professional organizations, academia, nonprofits, and industry. For example, in the domain of real-world evidence generation, we have research collaborations with the American Society of Clinical Oncology's CancerLinQ and a start-up called Flatiron Health. In the area of data sharing, we're collaborating with a nonprofit called Project Data Sphere on open access data. We are also developing a framework for decentralized sharing of data at scale with IBM Watson Health based on blockchain, which allows users to access and add to a secure, shared ledger or spreadsheet of data. We're also working with data science experts at MIT and Stanford on innovations based on artificial intelligence and algorithmic analytics that can help the drug development and the life sciences communities.").

81 *See National Evaluation System For Health Technology (NEST)*, U.S. Food & Drug Admin. (Oct. 29, 2019), https://www.fda.gov/about-fda/cdrh-reports/national-evaluation-system-health-technology-nest ("NEST" is designed to "generate evidence across the total product lifecycle of medical devices by strategically and systematically leveraging real-world evidence and applying advanced analytics to data tailored to the unique data needs and innovation cycles of medical devices."); *Medical Device Safety Action Plan: Protecting Patients, Promoting Public Health*, U.S. Food & Drug Admin. 6 (2018), https://www.fda.gov/media/112497/download (It is managed "by the non-profit Medical Device Innovation Consortium (MDIC) through the NEST Coordinating Center (NESTcc).").

82 *FDA Launches New Digital Tool to Help Capture Real World Data from Patients to Help Inform Regulatory Decision-Making*, U.S. Food & Drug Admin. (Nov. 6, 2018), https://www.fda.gov/NewsEvents/Newsroom/FDAInBrief/ucm625228.htm [hereinafter FDA Launches New Digital Tool]; *see also FDA's MyStudies Application (App)*, U.S. Food & Drug Admin. (Jan. 28, 2019), https://www.fda.gov/Drugs/ScienceResearch/ucm624785.htm; Zachary Wyner et al., *FDA MyStudies App: A Patient Centered Outcomes Research Trust Fund Enabler for Distributed Clinical Trials and Real World Evidence Studies*, U.S. Food & Drug Admin. (2018), https://www.fda.gov/media/119835/download.

83 *FDA Launches New Digital Tool, supra* note 82.

Endnotes to *Part II. Case Studies: Public Engagement at the Federal Communications Commission and Consumer Financial Protection Bureau*

1 Hila Mehr, *Artificial Intelligence for Citizen Services and Government*, Harv. Kennedy Sch. Ash Ctr. for Democratic Governance & Innovation (Aug. 2017), https://ash.harvard.edu/files/ash/files/artificial_intelligence_for_citizen_services.pdf.

2 *See* Katy Harris, *The Emergence of Civic Tech*, Knight Found. (Dec. 4, 2013), https://knightfoundation.org/articles/emergence-civic-tech/.

3 For other examples of NLP applications for citizen engagement, *see* Karel Verhaeghe, *Natural Language Processing at CitizenLab: How Machine Learning Can Transform Citizen Engagement Projects*, citizenlab (Apr. 29, 2019), https://www.citizenlab.co/blog/product-update/natural-language-processing-at-citizenlab-how-machine-learning-can-transform-citizen-engagement-projects/.

4 Maeve P. Carey, *Counting Regulations: An Overview of Rulemaking, Types of Federal Regulations, and Pages in the Federal Register*, Cong. Res. Serv. (Oct. 4, 2016), https://fas.org/sgp/crs/misc/R43056.pdf.

5 FDMS.gov, https://www.fdms.gov/fdms/public/aboutus (last visited Dec. 8, 2019).

6 Jacob Kastrenakes, *FCC Received a Total of 3.7 Million Comments on Net Neutrality*, Verge (Sept. 16, 2014, 6:06 PM EDT), https://www.theverge.com/2014/9/16/6257887/fcc-net-neutrality-3-7-million-comments-made. An agency official with the Department of Labor, for instance, confirmed the general trend of increasing numbers of comments.

7 5 U.S.C. § 553 (2018).

8 Carey, *supra* note 4. *But see Portland Cement Ass'n v. Ruckelshaus* 486 F.2d 375, 392 (D.C. Cir. 1973) (holding that an EPA standard was inadequate because interested persons had not received the methodology in time to comment on it).

9 Executive Order 12866: Regulatory Planning and Review, 58 Fed. Reg. 51,735 (Sept. 30, 1993).

10 Am. Radio Relay League, Inc. v. FCC, 524 F.3d 227, 237 (D.C. Cir. 2008).

11 *U.S. v. Nova Scotia Food Prod. Corp.*, 568 F.2d 240, 253 (2d Cir. 1977).

12 For a recent example of such a challenge, see *East Bay Sanctuary Covenant v. Trump*, 354 F. Supp. 3d 1094 (N.D. Cal. 2018), appeal docketed, No. 18-17436 (9th Cir. filed Dec. 26, 2018) (arguing that the Trump Administration violated notice-and-comment procedures in changing port of entry rules).

13 Michael A. Livermore et al., *Computationally Assisted Regulatory Participation*, 93 Notre Dame L. Rev. 977, 983-84 (2018).

14 *See, e.g.,* Cary Coglianese & David Lehr, *Regulating by Robot: Administrative Decision Making in the Machine-Learning Era*, 105 Geo. L.J. 1147, 1198 (2017); Livermore, *supra* note 13 (2018); Melissa Mortazavi, *Rulemaking Ex Machina*, 117 Colum. L. Rev. 202 (2017).

15 Public-private research firm MITRE, for example, is developing a comment clustering tool for the Department of Health and Human Services which it is using as a prototype and has not yet been fully transferred to the agency. *See* Telephone Interview with L. Karl Branting and Chris Giannella, MITRE Corporation (Feb. 19, 2019).

16 *See* Livermore, *supra* note 13, at 979.

17 Elise Hu, *John Oliver Helps Rally 45,000 Net Neutrality Comments to FCC*, NPR (June 3, 2014, 11:56 AM ET), https://www.npr.org/sections/alltechconsidered/2014/06/03/318458496/john-oliver-helps-rally-45-000-net-neutrality-comments-to-fcc; *The FCC (@FCC)*, Twitter (June 2, 2014, 1:44 PM), https://twitter.com/FCC/status/473565753463959552.

18 Hu, *supra* note 17.

19 Michelle Castillo, *John Oliver's Plea for Net Neutrality May Have Provoked Hackers to Knock Out FCC Website*, CNBC (May 9, 2017, 11:18 AM EDT), https://www.cnbc.com/2017/05/09/fcc-john-oliver-net-neutrality-plea-may-have-brought-down-fcc-site.html.

20 Devin Coldewey, *Net Neutrality Activists, Not Hackers, Crashed the FCC's Comment System*, TechCrunch (Aug. 7, 2018), https://techcrunch.com/2018/08/07/net-neutrality-activists-not-hackers-crashed-the-fccs-comment-system/.

21 Paul Hitlin & Skye Toor, *Public Comments to the Federal Communications Commission About Net Neutrality Contain Many Inaccuracies and Duplicates*, Pew Res. Ctr. (Nov. 29, 2017), http://www.pewinternet.org/2017/11/29/public-comments-to-the-federal-communications-commission-about-net-neutrality-contain-many-inaccuracies-and-duplicates/.

22 *Id.*

23 Brian Naylor, *As FCC Prepares Net-Neutrality Vote, Study Finds Millions of Fake Comments*, NPR (Dec. 14, 2017, 5:00 AM ET), https://www.npr.org/2017/12/14/570262688/as-fcc-prepares-net-neutrality-vote-study-finds-millions-of-fake-comments.

24 Emprata, FCC Restoring Internet Freedom Docket 17-108 Comments Analysis (Aug. 30, 2017).

25 Bo Pang et al., *Thumbs Up?: Sentiment Classification Using Machine Learning Techniques*, ACL Anthology 1 (July 2002), https://www.aclweb.org/anthology/W02-1011. The vast majority of academic works on sentiment analysis use datasets of movie reviews or tweets due to the availability of large amounts of (often highly opinionated) data. It is possible that some of the more ambiguous and difficult language constructs that thwart simple NLP tools would be more prevalent in these datasets than in AI/ML notice-and-comment comments.

26 *See* Trevor Hastie, Robert Tibshirani & Jerome Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction 672 (2d ed. 2016).

27 *See* Pang et al., *supra* note 25.

28 *See* Yoon Kim, *Convolutional Neural Networks for Sentence Classification*, Proc. 2014 Conf. Empirical Methods Nat. Language Processing (EMNLP) 1746 (2014), https://www.aclweb.org/anthology/D14-1181.pdf.

29 Emprata, *supra* note 24, at 6.

30 *Id.* at 23.

31 *Id.*

32 *Id.*

33 *Id.* at 23-24.

34 *Id.*

35 The sample was characterized by class imbalance, containing fewer comments supporting repeal. The comments in the training data that supported repeal consisted mostly of form letters, as commenters who supported repeal tended to use form letters more often. *Id.* at 25.

36 Jeff Kao, *More Than a Million Pro-Repeal Net Neutrality Comments Were Likely Faked*, Hackernoon (Nov. 23, 2017), https://hackernoon.com/more-than-a-million-pro-repeal-net-neutrality-comments-were-likely-faked-e9f0e3ed36a6.

37 *Id.*

38 *Id.*

39 *See* Alec Radford et al., *Better Language Models and Their Implications*, OpenAI (Feb. 14, 2019), https://blog.openai.com/better-language-models/.

40 *See* Issie Lapowsky, *How Bots Broke the FCC's Public Comment System*, WIRED (Nov. 28, 2017, 12:19 PM), https://www.wired.com/story/bots-broke-fcc-public-comment-system/.

41 The foundation's DearFCC tool used custom, automatically generated text to allow human users to "craft a unique comment" on the FCC's net neutrality proposal with "just two clicks." Rainey Reitman, Electronic

Frontier Foundation, *Launching DearFCC: The Best Way to Submit Comments to the FCC about Net Neutrality* (May 8, 2017), https://www.eff.org/deeplinks/2017/05/launching-dearfcc-best-way-submit-comments-fcc-about-net-neutrality.

42 Emprata, *supra* note 24, at 6.

43 *Id.* at 5.

44 *Id.* at 14.

45 *Id.* at 10.

46 *Id.* at 14.

47 *Id.* at 9.

48 *Id.* at 7 (emphasis added).

49 *Id.* at 20.

50 *Id.*

51 *Id.*

52 *See* Consumer Financial Protection Bureau, https://www.consumerfinance.gov/ (last visited Dec. 13, 2019).

53 *Strategic Plan, Budget, and Performance Plan and Report*, Consumer Fin. Protection Bureau 47 (Mar. 2014), https://files.consumerfinance.gov/f/strategic-plan-budget-and-performance-plan-and-report-FY2013-15.pdf ("CONSUMER RESPONSE SYSTEM—NATURAL LANGUAGE PROCESSING: Gain greater efficiency and effectiveness in complaint handling to respond to the anticipated increase of interactions with consumers as the Bureau adds to the number of available services and these services become better known to the public.").

54 *Id.* at 18.

55 *See How We Use Complaint Data*, Consumer Fin. Protection Bureau, https://www.consumerfinance.gov/complaint/data-use/ (last visited Dec. 13, 2019).

56 *See Consumer Complaint Database*, Consumer Fin. Protection Bureau, https://www.consumerfinance.gov/data-research/consumer-complaints/search/?from=0&has_narrative=true&searchField=all&searchText=&size=25&sort=created_date_desc (last visited Dec. 13, 2019).

57 *See* Off. of Consumer Response, *Narrative Scrubbing Standard,* Consumer Fin. Protection Bureau (Mar. 2015), https://files.consumerfinance.gov/a/assets/201503_cfpb_Narrative-Scrubbing-Standard.pdf.

58 Interview with Lewis Kirvan, Consumer Financial Protection Bureau (Feb. 25, 2019) (on file with authors).

59 *Id.*

60 LDA allows people to find k number of topic clusters and cluster the documents into these topics. STM's differentiating feature is that it "permits users to incorporate arbitrary metadata, defined as information about each document, into the topic model." Bettina Grün & Kurt Hornik, *Topicmodels: An R Package for Fitting Topic Models*, 40 J. of Stat. Software 1 (2011). This enables the CFPB to define which features they value and want to categorize topics on.

61 *See* Telephone Interview with Robert Waterman, Compliance Specialist, Department of Labor, Wage and Hour Division (Apr. 4, 2019).

62 *Id.*

63 *See* Branting & Giannella Interview, *supra* note 15; Telephone Interview with Vlad Eidelman, Vice Pres. of Res., FiscalNote (Feb. 5, 2019).

64 *See* Livermore, *supra* note 13, at 988 ("The State Department's Keystone XL oil pipeline decision received more than 2.5 million comments; the Federal Communications Commission received over 1.25 million comments on its net neutrality rules; and the EPA received over 4 million comments on its proposed Clean Power Plan.").

65 *About Us*, Regulations.gov, https://test.regulations.gov/aboutProgram.

66 *See* Eidelman Interview, *supra* note 63.

67 *See Searchable Electronic Rulemaking System*, FEC.Gov, sers.fec.gov/fosers/ (last visited Apr. 7, 2019).

68 *See* Branting & Giannella Interview, *supra* note 15.

69 *See* Edward Walker, Grassroots for Hire: Public Affairs Consultants in American Democracy (2014); Philip N. Howard, New Media Campaigns and the Managed Citizen (2005).

70 As Livermore has noted, use of automation to filter out bot-generated comments raises unresolved questions about the core purpose of notice-and-comment rulemaking. If the purpose of rulemaking is to sharpen an agency's analysis—that is, to improve the agency's analytic rationality—then even purely bot-generated comments, separate from any type of political elite or grassroots mobilization effort, should be welcomed. If, however, the aim of rulemaking is to provide a measure of the weight of public support—that is, to serve a plebiscitory function—then bot-generated comments presumptively lack any value. *See* Livermore, *supra* note 13, at 990-92

71 5 U.S.C. § 553(c) (2018).

72 *Bolling v. Sharpe*, 347 U.S. 497, 500 (1954) (holding that the Fifth Amendment requires that the federal government is also bound by the equal protection clause).

73 *See* Julia Angwin et al., *Machine Bias*, Pro Publica (May 23, 2016) https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing; *see also* Tom Simonite, *How Coders Are Fighting Bias in Facial Recognition Software*, WIRED (Mar. 28, 2018, 7:00 AM), https://www.wired.com/story/how-coders-are-fighting-bias-in-facial-recognition-software/.

74 *See* Dell Cameron & Jason Prechtel, *How an Investigation of Fake FCC Comments Snared a Prominent D.C. Media Firm*, Gizmodo (Feb. 21, 2019, 12:15 PM), https://gizmodo.com/how-an-investigation-of-fake-fcc-comments-snared-a-prom-1832788658.

75 FiscalNote's private clients include Amazon, Johnson & Johnson, and Intel, among others. *Global Issues Management*, FiscalNote, https://www.fiscalnote.com/solutions/global-issues-management.

Endnotes to *Part II. Case Studies: Autonomous Vehicles for Mail Delivery at the United States Postal Service*

1 *The United States Postal Service Delivers the Facts*, U.S. Postal Serv. (Apr. 2019), https://about.usps.com/news/delivers-facts/usps-delivers-the-facts.pdf.

2 Lydia Saad, *Postal Service Still Americans' Favorite Federal Agency*, Gallup News (May 13, 2019), https://news.gallup.com/poll/257510/postal-service-americans-favorite-federal-agency.aspx.

3 39 U.S.C. § 101(a) (2018).

4 *Id.*

5 *See Universal Service and the Postal Monopoly: A Brief History*, U.S. Postal Service (Oct. 2008), https://about.usps.com/universal-postal-service/universal-service-and-postal-monopoly-history.pdf.

6 *FY2018 Annual Report to Congress*, U.S. Postal Service 25 (2018), https://about.usps.com/who-we-are/financials/annual-reports/fy2018.pdf.

7 Task Force on the United States Postal System, *United States Postal Service: A Sustainable Path Forward* 2 (2018), https://home.treasury.gov/system/files/136/USPS_A_Sustainable_Path_Forward_report_12-04-2018.pdf.

8 *Id.* at 2-4.

9 *Id.* at 24.

10 Lori Rectanus, *U.S. Postal Service: Key Considerations for Restoring Fiscal Sustainability*, U.S. Gov't Accountability Off. 1 (2017), https://www.gao.gov/assets/690/682534.pdf.

11 *FY2018 Annual Report to Congress, supra* note 6, at 27.

12 *U.S. Postal Service: Projected Capital Spending and Processes for Addressing Uncertainties and Risks*, U.S. Gov't Accountability Off. 11 (2018), https://www.gao.gov/assets/700/692859.pdf.

13 *FY2018 Annual Report to Congress, supra* note 6, at 3.

14 U.S. Postal Service, Off. of the Inspector Gen., Fuel Consumption and Cost Risk Mitigation 1 (2017), https://www.uspsoig.gov/sites/default/files/document-library-files/2017/NL-AR-17-004.pdf. The Postal Service spent $502 million on delivery vehicle fuel in FY2018, up from $409 million in FY2017 and $359 million in FY2016. In FY2018, delivery vehicle fuel accounted for 3.4% of operating expenses. FY2018 Annual Report to Congress, *supra* note 6, at 40.

15 *FY2018 Annual Report to Congress, supra* note 6, at 22.

16 Bureau of Lab. Stat., *Injuries, Illnesses, and Fatalities, U.S. Dep't of Lab.* (2017), https://www.bls.gov/iif/oshwc/cfoi/cftb0314.htm.

17 U.S. Postal Service, Off. of the Inspector Gen., Autonomous Vehicles for the Postal Service 7 (2017), https://www.uspsoig.gov/sites/default/files/document-library-files/2017/RARC-WP-18-001.pdf [hereinafter Autonomous Vehicles for the Postal Service].

18 *High-Risk Series: Substantial Efforts Needed to Achieve Greater Progress on High-Risk Areas*, U.S. Gov't Accountability Off. 99-102 (2019), https://www.gao.gov/assets/700/697245.pdf.

19 *See generally* Task Force on the United States Postal System, *supra* note 7.

20 *See generally* James M. Anderson et al., *Autonomous Vehicle Technology: A Guide for Policymakers*, RAND Corp. (2016), https://www.rand.org/pubs/research_reports/RR443-2.html.

21 *See generally Automated Vehicles 3.0: Preparing for the Future of Transportation*, U.S. Dep't of Transp. (2018), https://www.transportation.gov/sites/dot.gov/files/docs/policy-initiatives/automated-vehicles/320711/preparing-future-transportation-automated-vehicle-30.pdf.

22 *See generally* Autonomous Vehicles for the Postal Service, *supra* note 17.

23 Sam.gov, https://www.fbo.gov/index.php?s=opportunity&mode=form&tab=core&id=c5e74cf8db0539949cec2dc4b506558f&_cview=0 (last visited Dec. 13, 2019).

24 The agency's UAV RFI also indicated that one of the reasons the Postal Service is interested in UAVs is "to collect geodetic/spatial data" to incorporate into autonomous ground vehicles.

25 *See generally* Autonomous Vehicles for the Postal Service, *supra* note 17.

26 E-mail from Kimberly Frum, USPS spokesperson (Nov. 6, 2019) (on file with author); e-mail from Han Dinh, USPS Program Director for Vehicle Engineering (Oct. 26, 2019) (on file with author).

27 *United States Postal Service Autonomous Vehicles Capability*, U.S. Postal Serv. (Feb. 22, 2019), https://beta.sam.gov/opp/02bc426f89d161981688de9ba68008a2/view#general.

28 E-mail from Kimberly Frum, *supra* note 26.

29 *Driving Test: Autonomous Truck Pilot Begins*, LINK (May 21, 2019), https://link.usps.com/2019/05/21/driving-test/.

30 *TuSimple Self-Driving Truck Service to the United States Postal Service*, tu simple (May 21, 2019), https://www.tusimple.com/wp-content/uploads/2019/05/USPS-TuSimple-Press-Release-FINAL.pdf.

31 Murray Slovick, *TuSimple Completes Self-Driving Truck Test for the USPS*, ElectronicDesign (June 24, 2019), https://www.electronicdesign.com/automotive/tusimple-completes-self-driving-truck-test-usps.

32 *Id.*

33 E-mail from Kimberly Frum, *supra* note 26.

34 *See generally* U.S. Postal Service, Off. of the Inspector Gen., Summary Report: Public Perception of Self-Driving Technology for Long-Haul Trucking and Last-Mile Delivery (2017), https://www.uspsoig.gov/sites/default/files/document-library-files/2017/RARC-WP-17-011.pdf.

35 86% of survey respondents were aware of self-driving cars, but just 40% had heard of delivery applications. *Id.* at 5.

36 *Id.* at 12.

37 *Id.* at 13.

38 *Id.* at 1.

39 *Executive Counsel Installed*, Nat'l Ass'n of Letter Carriers (Dec. 24, 2018), https://www.nalc.org/news/bulletin/pdf2018/Bulletin-18-16-red.pdf. Similarly, following Amazon's release of an autonomous delivery robot in January 2019, the NALC president said, "As this technology matures, the Postal Service will likely consider adopting it. So, NALC must develop the expertise to understand and respond to how our work is organized by changing technology." *Amazon Rolls out New Delivery Fleet*, Postal Record (Mar. 2019), https://www.nalc.org/news/the-postal-record/2019/march-2019/document/Amazon.pdf.

40 Autonomous Vehicles for the Postal Service, *supra* note 17, at 28.

41 E-mail from Kimberly Frum, USPS spokesperson (Nov. 22, 2019) (on file with author).

42 E-mail from Jordan Schultz, Communications and PAC Manager of NRLCA (Nov. 22, 2019) (on file with author).

43 Christopher Jackson, *City Delivery Updates*, Postal Record (June 2019), https://www.nalc.org/news/the-postal-record/2019/june-2019/document/Director-of-City-Delivery.pdf.

44 *USPS Is Testing Self-Driving Trucks*, APWU (June 24, 2019), https://apwu.org/news/usps-testing-self-driving-trucks.

45 The OIG public opinion report, for instance, characterized the business case as "undeniable."

46 Patrick Olsen, *Cadillac Tops Tesla in Consumer Reports' First Ranking of Automated Driving Systems*, Consumer Reports (Oct. 24, 2018), https://www.consumerreports.org/autonomous-driving/cadillac-tops-tesla-in-automated-systems-ranking/.

47 David Z. Morris, *Tesla Could Deliver 'Full Self-Driving' Within Weeks*, Fortune (Nov. 20, 2019), https://fortune.com/2019/11/20/tesla-full-self-driving-car-tsla-stock/.

48 Alex Davies, *Amazon Dives into Self-Driving Cars with a Bet on Aurora*, Wired (Feb. 7, 2019), https://www.wired.com/story/amazon-aurora-self-driving-investment-funding-series-b/; Elizabeth Woyke, *FedEx Bets on Automation as it Prepares to Fend Off Uber and Amazon*, MIT Tech. Rev. (Feb. 3, 2017), https://www.technologyreview.com/s/602896/fedex-bets-on-automation-as-it-prepares-to-fend-off-uber-and-amazon/.

49 *UPS Invests in Autonomous Trucking Company, Tests Self-Driving Tractor Trailers*, UPS (Aug. 15, 2019), https://pressroom.ups.com/pressroom/ContentDetailsViewer.page?ConceptType=PressReleases&id=1565871221437-794; *TuSimple Announces Oversubscribed Round of Series D Investment at $215 Million*, tusimple (Sept. 17, 2019), https://www.tusimple.com/wp-content/uploads/2019/09/TuSimple-Series-D-Extension-Press-Release.pdf.

50 Morgan Forde, *TuSimple Plans Fully Driverless Deliveries in 2021*, supplychaindive (Oct. 14, 2019), https://www.supplychaindive.com/news/tusimple-driverless-deliveries-2021/564947/.

51 NHTSA, *Automated Driving Systems*, U.S. Dep't of Transp., https://www.nhtsa.gov/vehicle-manufacturers/automated-driving-systems (last visited Dec. 10, 2019) ("Advanced vehicle technologies hold the promise not only to change the way we drive but to save lives. The continuing evolution of automotive technology aims to deliver even greater safety benefits and Automated Driving Systems (ADS) that––one day––could potentially handle the whole task of driving when we don't want to or can't do it ourselves. Fully automated cars and trucks that drive us, instead of us driving them, are a vision that seems on the verge of becoming a reality.").

52 *Autonomous Vehicles: Self-Driving Vehicles Enacted Legislation,* Nat'l Conf. of State Legis. (Oct. 9, 2019), http://www.ncsl.org/research/transportation/autonomous-vehicles-self-driving-vehicles-enacted-legislation.aspx.

53 *See* Andrew J. Hawkins, *Congress takes another stab at passing self-driving car legislation*, The Verge, July 28, 2019, https://www.theverge.com/2019/7/28/8931726/congress-self-driving-car-bill-redo-2019. The House of Representatives passed H.R.3388, the Safely Ensuring Lives Future Deployment and Research In Vehicle Evolution Act ("SELF DRIVE Act"), in September 2017. A complementary bill, S.1885, American Vision for Safer Transportation through Advancement of Revolutionary Technologies Act ("AV START Act") was introduced in November 2017.

54 David Shepardson, *U.S. Congress seeks to jump start stalled self-driving car bill*, Reuters, July 30, 2019, https://www.reuters.com/article/us-autos-selfdriving-congress/u-s-congress-seeks-to-jump-start-stalled-self-driving-car-bill-idUSKCN1UP2HV. For analysis of the discussion draft, *see* Susan H. Lent et al, *New Autonomous Vehicle Legislation Proposed—What You Should Know*, Akin Gump Strauss Hauer & Feld, Nov. 22, 2019, https://www.akingump.com/en/news-insights/new-autonomous-vehicle-legislation-proposed-what-you-should-know.html.

55 *Highly Automated Vehicles: Federal Perspectives on the Deployment of Safety Technology: Hearing before the S. Comm. on Commerce, Sci., and Transp.,* 116th Cong. (2019).

56 *See* National Conference of State Legislatures, *Autonomous Vehicles State Bill Tracking Database*, www.ncsl.org/research/transportation/autonomous-vehicles-legislative-database.aspx.

57 *See, e.g.,* Bryant Walker Smith, *Automated Driving and Product Liability*, 2017 Mich. St. L. Rev. 1 (2017); Kenneth S. Abraham & Robert L. Rabin, *Automated Vehicles and Manufacturer Responsibility for Accidents: A New Legal Regime for a New Era*, 105 Va. L. Rev. 127 (2019).

58 Lent et al, *supra* note 54.

59 *See* S.1885, *supra* note 53 ("Compliance with a motor vehicle safety standard prescribed under this chapter does not exempt a person from liability at common law."); H.R.3388, *supra* note 53 ("Compliance with a motor vehicle safety standard prescribed under this chapter does not exempt a person from liability at common law . . . . Nothing in this section shall be construed to preempt common law claims.)

60 Mark Geistfeld, *A Roadmap for Autonomous Vehicles: State Tort Liability, Automobile Insurance, and Federal Safety Regulation*, 105 Cal. L. Rev. 1611 (2017).

61 For a general discussion of the privacy implications of autonomous vehicles, see Dorothy J. Glancy, *Privacy in Autonomous Vehicles*, 52 Santa Clara L. Rev. 1171 (2012).

62 Lent et al, *supra* note 54.

63 *See, e.g.,* S.1885, *supra* note 53 (ordering the Secretary of Transportation to establish a "Data Access Advisory Committee" composed of various stakeholders "to discuss and make policy recommendations to Congress with respect to the ownership of, control of, or access to, information or data that vehicles collect, generate, record, or store in an electronic form that is retrieved from a highly automated vehicle or automated driving system.").

64 *See, e.g.,* H.R.3388, *supra* note 53 (requiring any vehicle manufacturer to publish a written "privacy plan" that discloses the manufacturer's practices "with respect to the way that information about vehicle owners or occupants is collected, used, shared, or stored")

65 Norton Rose Fulbright White, *The Privacy Implications of Autonomous Vehicles*, July 17, 2017, https://www.dataprotectionreport.com/2017/07/the-privacy-implications-of-autonomous-vehicles/.

66 *See* H.B. 1197, 66th Legis. Ass. Sess. (ND 2019) (failed to pass), https://www.legis.nd.gov/assembly/66-2019/bill-actions/ba1197.html ("The owner of an autonomous vehicle owns any data or information stored by the autonomous vehicle or gathered by the use of the autonomous vehicle.")

67 5 U.S.C. § 552a (2012).

68 Glancy, *supra* note 61, at 1202 ("Moreover, to the extent that federal agencies collect or receive information about identifiable users of autonomous vehicles, the Privacy Act of 1974 would apply.").

69 5 U.S.C. § 552a (2012) (emphasis added).

70 Autonomous Vehicles For The Postal Service, *supra* note 17, at 9 (noting that lidar technology does not capture "faces, license plates, or other privacy red flags).

71 There is an ongoing debate about whether data collected by law enforcement license plate readers qualifies as "individual" information under the Privacy Act. *See, e.g.,* Johanna Zmud et al, *License Plate Reader Technology: Transportation Uses and Privacy Risks* (2016), https://scholarship.law.tamu.edu/cgi/viewcontent.cgi?article=1920&context=facscholar.

72 To alleviate privacy concerns about "vehicle-to-vehicle" communications, the National Highway Traffic Safety Administration recently proposed excluding "reasonably linkable" data — including license plate numbers and driver/owner names — from such communications. *See* Federal Motor Vehicle Safety Standards; V2V Communications, 82 Fed. Reg. 3854 (proposed Jan. 12, 2017) (to be codified at 49 C.F.R. 571).

73 Autonomous Vehicles for the Postal Service, *supra* note 17.

74 *NHTSA, Automated Vehicles for Safety,* U.S. Dep't of Transp., https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety.

75 U.S. Energy Info. Admin., *Autonomous Vehicles: Uncertainties and Energy Implications*, U.S. Dep't of Energy (2018), https://www.eia.gov/outlooks/aeo/pdf/AV.pdf.

76 Autonomous Vehicles for the Postal Service, *supra* note 17, at 8-9.

77 *Id.*

78 *Id.* at 9.

Endnotes to *Part III. Implications and Recommendations*

Building Internal Capacity

1 Some of the sections below rely on David Freeman Engstrom & Daniel E. Ho, *Artificially Intelligent Government: A Review and Agenda*, in Big Data Law (Roland Vogl ed., forthcoming 2020).

2 *See* Oliver E. Williamson, *Public & Private Bureaucracies: A Transaction Cost Perspective*, 15 J.L. Econ. & Org. 306, 319 (1999).

3 For general literature on contracting out, including historical perspective, *see* Paul R. Verkuil, Outsourcing Sovereignty: Why Privatization of Government Functions Threatens Democracy and What We Can Do About It (2007); John D. Donohue, The Privatization Decision: Public Ends, Private Means (1991); Jon D. Michaels, Constitutional Coup: Privatization's Threat to the American Republic (2017).

4 *See* Verkuil, *supra* note 3. For a recent review of the vast literature on the costs and benefits, and the political and other determinants of contracting out, see Jonathan Levin & Steven Tadelis, *Contracting for Government Services: Theory and Evidence from U.S. Cities,* 58 J. Indus. Econ. 507 (2010).

5 *See* John D. Donahue, *Transformation of Government Work: Causes, Consequences, and Distortions*, in Government by Contract: Outsourcing and American Democracy 41, 49 (Jody Freeman & Martha Minow eds., 2009) (distinguishing between commodity tasks that are "well defined, relatively easy to evaluate, and available from competitive private suppliers" and custom tasks, which lack these features).

6 *See* Joan Puigcerver, *Are Multidimensional Recurrent Layers Really Necessary for Handwritten Text Recognition?* (2017), http://www.jpuigcerver.net/pubs/jpuigcerver_icdar2017.pdf. There has been some work on extracting figures and charts from PDF documents, as well as *classifying* those figures (for example, to distinguish bar-charts from pie-charts). However, current AI techniques are not equipped to understand what a figure is meant to *represent*. *See* Yan Liu et al., *Review of Chart Recognition in Document Images*, 8654 Visualization & Data Analysis 1, 1 (2013), https://www.spiedigitallibrary.org/conference-proceedings-of-spie/8654/865410/Review-of-chart-recognition-in-document-images/10.1117/12.2008467.full?SSO=1.

7 Additionally, even when building a particular tool may not require an immediate legacy system update, expanding those tools or integrating with other systems can require infrastructure upgrades. For example, the IRS is considering integrating its Return Revenue Program with an older linear scoring algorithm, the Discriminant Inventory Function System. Interview with Jeff Butler, Director of Research Databases, Internal Revenue Serv. (Feb. 11, 2019) (on file with author) [hereinafter Butler Interview]. Unsupervised models are designed to find latent patterns in unlabeled data (*i.e.,* data for which there is "no associated response yi" for observations of xi). The IRS conducted its first audits using the Discriminant Inventory Function (DIF) System in 1969, which significantly increased the agency's audit efficiency. William J. Hunter & Michael A. Nelson, *An IRS Production Function*, 49 Nat'l Tax J. 105, 108-09 (1996). The DIF was developed with the data provided from the Taxpayer Compliance Measurement Program that began operating in 1964, reviewed delinquent returns and accounts, and conducted detailed field audits. Combining programs at widely differing levels of technological maturity can require significant resources, and agencies should identify and upgrade legacy systems early on in order to avoid future setbacks. As an alternative, agencies may be able to build on top of existing technology by working closely with developers to design prototypes with existing systems in mind. Agencies such as the SEC and SSA have in some cases sought to integrate their tools with existing agency technology.

8 Presentation at "A Roundtable Discussion on the Use of Artificial Intelligence in the Federal Administrative Process," NYU School of Law (Feb. 25, 2019). Other agencies face the same challenge in upgrading legacy systems. Many applications used by the SSA, for example, utilize a programming language (COBOL) which was standardized 30 years ago and is now unfamiliar to most developers. As it transitions to modern languages like Java, the SSA has struggled to find personnel familiar with both languages. *See* Amelia Brust, *SSA on Track to Modernize IT Systems Over Next Five Years*, Fed. News Network (Apr. 17, 2018), https://federalnewsnetwork.com/it-modernization-month-2018/2018/04/social-security-on-schedule-to-modernize-it-systems-by/. The IRS is home to the federal government's two oldest databases and has determined that 52% of its hardware is "aged," or operating past useful life. Taxpayer Advoc. Serv., 1 Fiscal Year 2019 Objectives Report to Congress 49 (2019), https://taxpayeradvocate.irs.gov/reports/fy-2019-objectives-report-to-congress/full-report; Internal Revenue Serv., Strategic Plan FY2018-2022 5 (2018), https://www.irs.gov/about-irs/irs-strategic-plan. The PTO faces similar challenges. *See* U.S. Patent & Trademark Off., Patent Public Advisory Committee, 2018 Annual Report 8 (2018), https://www.uspto.gov/sites/default/files/documents/PPAC_2018_Annual_Report_2.pdf (recognizing "recent patent system outages, slow access times on PAIR data, and erroneous messages that are given to public users").[herinafter Presentation]

9 Amelia Brust, *SSA on Track to Modernize IT Systems over Next Five Years*, Fed. News Network (Apr. 17, 2018), https://federalnewsnetwork.com/it-modernization-month-2018/2018/04/social-security-on-schedule-to-modernize-it-systems-by/.

10 The federal regime also includes area-specific data constraints. For instance, the Health Insurance Portability and Accountability Act (HIPAA), of 1996, Pub. L. No. 104-191, 110 Stat. 1936 (codified as amended in scattered sections of 18, 26, 29 and 42 U.S.C.), sets forth privacy and security standards for protecting personal health information. Other sectoral laws include the Gramm-Leach-Bliley Act (GLBA), the Family Educational Rights and Privacy Act (FERPA), and the Fair Credit Reporting Act (FCRA).

11 Computer Matching and Privacy Protection Act, 5 U.S.C. § 552a(b), (e)(3) (2018).

12 *See id*. at §§ 552a(a)(8), 552a(o)-(r). In particular, 5 U.S.C. § 552a(p) requires independent verification before "adverse action" can be taken or, for information regarding the identification and amount of benefits granted, that "there is a high degree of confidence" in the information's accuracy, while 5 U.S.C. § 552a(p)(3)(A) requires "notice from such agency containing a statement of its findings and informing the individual of the opportunity to contest such findings." That said, the Privacy Act might be thought weak because its exemption for "routine uses" creates a loophole, permitting an array of data-sharing subject only to the requirement that an agency publish an entry in the Federal Register describing the use and, more generally, a "system of records" notice. 5 U.S.C. § 552a(e)(4)(D).

13 *See* Paperwork Reduction Act of 1980, 44 U.S.C. § 101.

14 44 U.S.C. § 3516 note (2000) (requiring agency action to ensure the "quality, objectivity, utility, and integrity of information").

15 *See, e.g.,* Transportation Recall Enhancement, Accountability and Documentation (TREAD) Act, Pub. L. No. 106-414, 114 Stat. 1800 (2000) (amending various provisions of 49 U.S.C. §§ 30101-30170 to require manufacturer disclosure of certain defect information in response to Congressional concern over high-profile tire defects).

16 *See, e.g.,* Transportation Recall Enhancement, Accountability and Documentation (TREAD) Act, Pub. L. No. 106-414, 114 Stat. 1800 (2000). For a recent effort to establish a voluntary system, see 83 Fed. Reg. 50872 (Oct. 10, 2018); see also Pilot Program for Collaborative Research on Motor Vehicles with High or Full Driving Automation, 83 Fed. Reg. 59353 (proposed Nov. 23, 2018) (to be codified at 49 C.F.R. pts. 555, 571, 591).

17 *See* NHTSA, *Review of the National Automotive Sampling System: Report to Congress,* U.S. Dep't of Transp. 30-32 (2015), https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812128. NHTSA has proposed but not finalized a Federal Motor Vehicle Safety Standards revision to mandate event data recorder ("EDR") installation for all light vehicles. *See* 77 Fed. Reg. 74,144 (Dec. 13, 2012) (to be codified at 40 C.F.R. pt. 571).

18 One academic study conducted for the EPA, for example, relied on self-reported discharge monitoring reports whose accuracy could not be independently validated. *See* Miyuki Hino et al., *Machine Learning for Environmental Monitoring*, 2018 Nature Sustainability 583, 583 (2018). The EPA has begun to explore tools that can identify fraudulent discharge monitoring reports.

19 Presentation, *supra* note 8.

20 Although the IRS Return Revenue Program processes both electronic and paper filings, the missing data in paper filings can limit the tool's effectiveness. Gov't Accountability Office, GAO-18-544, Tax Fraud and Noncompliance 8, 22-24 (2018).

21 *DERA, CIRA, and XBRL at the SEC: Expanding the Availability and Use of XBRL Data*, Fin. Executives Int'l (July 1, 2015), https://daily.financialexecutives.org/dera-cira-and-xbrl-at-the-sec-expanding-the-availability-and-use-of-xbrl-data/.

22 The NHTSA, for example, collects consumer complaints using a largely unstructured form. The agency could fully structure the input form, requiring complainants to select from a series of dropdown menus. However, this could extend or complicate the submission process or sacrifice information that cannot fit into a structured format. Furthermore, not all data should be standardized. The development team behind HHS Accelerate, a central procurement system, pointed out its decision to retain the divergent labeling schemes. Telephone Interview with HHS Accelerate Development Team (Mar. 22, 2019). Although this made it difficult to restructure and consolidate the purchasing data, the team recognized that purchasing departments had adopted labeling systems for specific organizational and legal reasons.

23 *See* Federal Information Security Management Act (FISMA), Public Law 107-347, 116 Stat. 2899; Memorandum from Clay Johnson III, Deputy Dir. For Mgmt., Off. of Mgmt. & Budget, on Safeguarding Against and Responding to the Breach of Personally Identifiable Information (May 22, 2007), https://epic.org/apa/ferpa/OMB-Directive.pdf (requiring all agencies to "to develop and implement a breach notification policy within 120 days"). For a state overview of data security and data disposal laws, *see Data Security Laws––State Government*, Nat'l Conf. of State Legis. (Feb. 22, 2019), http://www.ncsl.org/research/telecommunications-and-information-technology/data-security-laws-state-government.aspx; *Data Disposal Laws*, Nat'l Conf. of State Legis. (Jan. 4, 2019), http://www.ncsl.org/research/telecommunications-and-information-technology/data-disposal-laws.aspx.

24 The conventional view is that FISMA creates liability only for the intentional agency disclosures of data, but some courts have found that even negligent failures to prevent hacks are actionable. *See* AFGE v. Hawley, 543 F. Supp. 2d 44 (D.D.C. 2008). Security problems are real, and the U.S. federal government in particular has suffered high-profile data breaches. See, *e.g.,* Zolan Kanno-Youngs & David E. Sanger, *Border Agency's Images of Travelers Stolen in Hack*, N.Y. Times (June 10, 2019), https://www.nytimes.com/2019/06/10/us/politics/customs-data-breach.html. *See* generally Michael Froomkin, *Government Data Breaches*, 24 Berkeley Tech. L.J. 1019 (2009).

25 For instance, the Department of Veterans Affairs developed a strategy in its health care partnership with Alphabet's DeepMind that uses cryptographic hashes to obscure veterans' sensitive personal information and thus permit data-sharing. *See* Tom Simonite, *The VA Wants to Use DeepMind's AI to Prevent Kidney Disease*, Wired (Jan. 21, 2019, 7:00 AM), https://www.wired.com/story/va-wants-deepminds-ai-prevent-kidney-disease/.

26 Butler Interview, *supra* note 7. GANs pit a generator neural network against a discriminator neural network. The generator creates data instances and the discriminator determines whether the new data instance represents real data from the training dataset. *See* Ian J. Goodfellow et al., *Generative Adversarial Networks*, Cornell U. (June 10, 2014), https://arxiv.org/abs/1406.2661; Alec Radford et al., *Unsupervised Representation Learning With Deep Convolutional Generative Adversarial Networks*, Cornell U. (Nov. 19, 2015), https://arxiv.org/abs/1511.06434; Tim Salimans et al., *Improved Techniques for Training GANs*, Cornell U. (June 10, 2016), https://arxiv.org/abs/1606.03498.

27 J. Christopher Giancarlo, Keynote Address at FinTech Week, Georgetown Univ. Law Sch. (Nov. 7, 2018) (transcript available at https://www.cftc.gov/PressRoom/SpeechesTestimony/opagiancarlo59).

28 *See* Cary Coglianese & David Lehr, *Transparency and Algorithmic Governance*, 71 Admin. L. Rev. 4 (2019) ("[D]evelopment of machine-learning algorithms, especially for the kinds of specialized applications to which they would be applied by government officials, is a challenging endeavor. . . . It also requires knowledge of how policy choices can be embedded in mathematical choices made while designing the algorithm.").

29 As detailed in Part II's SSA case study, the Insight tool flags over 30 error types, from citation to a non-existing legal provision to potential inconsistencies in reasoning (*e.g.,* finding a functional impairment that would prevent a disability applicant from engaging in a posited form of employment).

30 As detailed in Part II's SEC case study, if historical enforcement patterns are used as training data, the system may unnecessarily confine enforcement actions to a subset of violations (*e.g.,* by triggering "runaway feedback loops") or fight the last war at the expense of spotting new evasions by sophisticated actors.

31 *Hiring Authorities*, U.S. Off. of Personnel Mgmt., https://www.opm.gov/policy-data-oversight/hiring-information/hiring-authorities/ (last visited Dec. 10, 2019).

32 Mark D. Reinhold, Assoc. Dir. Of Emp. Servs., Off. of Personnel Mgmt., Data Scientist Title Guidance (June 27, 2019), https://www.chcoc.gov/content/data-scientist-titling-guidance.

33 Margaret M. Weichert, Acting Dir., Off. of Personnel Mgmt., Delegation of Direct-Hire Appointing Authority for IT Positions (Apr. 5, 2019), https://chcoc.gov/content/delegation-direct-hire-appointing-authority-it-positions. These direct hire positions are still subject to public notice requirements and prioritized hiring requirements for displaced agency employees. Agencies may hire employees under the direct hiring authority for a four-year period, renewable for up to eight years.

34 In determining the right metrics, agencies should avoid focusing disproportionately on "quantifiable" projects. For example, CMS would benefit greatly from tools that identify types of fraud that humans have failed to consistently identify or do not have the capacity to monitor—yet the impact of such a tool would be difficult to quantify as compared to a tool that simply makes existing human processes more efficient. CMS has acknowledged that a Fraud Prevention System assessed based on return on investment could lead to biased enforcement: the program may more easily detect money from local providers who do not possess sophisticated offshore networks for hiding their ill-gotten gains. Centers for Medicare & Medicaid Servs., *Report to Congress: Fraud Prevention System Second Implementation Year*, Dep't of Health & Hum. Servs. 16 (2014), https://www.cms.gov/About-CMS/Components/CPI/Widgets/Fraud_Prevention_System_2ndYear.pdf. A return-on-investment system is likely to prioritize such easily recoverable fraud, which the CMS notes is, "an undesirable result, given both types of fraud have no place." *Id*. at iii.

35 Meredith Somers, *FPS 2.0 More User-Friendly, But Not for Medicare Frauds*, Fed. News Network (Oct. 17, 2016), https://federalnewsnetwork.com/technology-main/2016/10/FPS-2-0-user-friendly-not-medicare-frauds/ ("[Agencies] need to know down to the level of a particular [predictive] model, what the return on investment is so [the agency] can make refinements, retire models that aren't as worthwhile as others and continue to develop and prioritize models based on historical trends.").

36 Telephone Interview with Raymond Wedgeworth, Dir., Data Analytics & Sys. Grp., Ctrs. for Medicare & Medicaid Servs. (Mar. 1, 2019). Yet this process remains imperfect: although the CMS Fraud Prevention System has demonstrated some success, the agency struggles to systematically track the program's success and account for variable performance. James Swann, *Are Medicare Anti-Fraud Efforts Flawed?*, Bloomberg BNA (Oct. 11, 2017), https://www.bna.com/medicare-antifraud-efforts-n73014470774/. Establishing the right metrics with which to measure successful tools can enable agencies to retire tools that fail to meet agency standards and deploy new ones.

37 U.S. Customs & Border Protection, Southern Border Pedestrian Field Test: Summary Report 8 (2016). Similarly, CMS has spent nearly $200 million in contracts to develop its Fraud Prevention System, which is already on its second iteration. U.S. Gov't Accountability Off., GAO-17-710, Medicare: CMS Fraud Prevention System Uses Claims Analysis to Address Fraud 4 (2017).

38 The Return Revenue Program, a tax fraud detection system, experienced a 12% increase in its error rate increase from 2016 to 2017. The system is also highly variable: adding two additional filters to the system led to a 495% increase in flagged tax returns in 2018. Taxpayer Advoc. Serv., *supra* note 8, at 29.

39 Presentation, *supra* note 8.

40 This may include creating provisions for partial deployment, setting expectations, and establishing multiple layers of testing.

41 Proposals range from graph mining to chatbots. Butler Interview, *supra* note 7

42 Sean Khozin et al., *INFORMED: An Incubator at the US FDA for Driving Innovation in Data Science and Agile Technology*, 17 Nature Rev. Drug Discovery 529, 529 (2018).

43 *Content of Premarket Submissions for Management of Cybersecurity in Medical Devices,* U.S. Food & Drug Admin. 11 (Oct. 18, 2018), https://www.fda.gov/media/119933/download. The FDA held a public workshop at the end of January 2019 focusing on the cybersecurity guidance.

44 *See, e.g.,* U.S. Food & Drug Admin., Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)––Discussion Paper and Request for Feedback 3 (2019).

45 *Medical Device Cybersecurity: Regional Incident Preparedness and Response Playbook*, MITRE Corp. (2018), https://www.mitre.org/sites/default/files/publications/pr-18-1550-Medical-Device-Cybersecurity-Playbook.pdf.

46 Suzanne B. Schwartz, *The Medical Device Ecosystem and Cybersecurity––Building Capabilities and Advancing Contributions*, U.S. Food & Drug Admin. (Nov. 1, 2019), https://www.fda.gov/NewsEvents/Newsroom/FDAVoices/ucm624749.htm. The FDA has executed two Memorandum of Understandings "of new medical device vulnerability information sharing analysis organizations." *See Memorandum Of Understanding Between the National Health Information Sharing & Analysis Center, Inc. (NH-ISAC), Medisao and the U.S. Food and Drug Administration Center for Devices and Radiological Health,* U.S. Food & Drug Admin. (Oct. 1, 2018), https://www.fda.gov/AboutFDA/PartnershipsCollaborations/MemorandaofUnderstandingMOUs/OtherMOUs/ucm622056.htm; *Memorandum of Understanding Between the Health Information Sharing & Analysis Center, Inc. (H-ISAC), Sensato Critical Infrastructure ISAO (Sensato-ISAO) and the U.S. Food and Drug Administration Center For Devices and Radiological Health,* U.S. Food & Drug Admin. (Oct. 1, 2018), https://www.fda.gov/AboutFDA/PartnershipsCollaborations/MemorandaofUnderstandingMOUs/OtherMOUs/ucm622055.htm.

47 *See* Michael Veale et al., *Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making,* CHI Conf. Hum. Factors Computing Syss. Procs. (2018) (interviewing 27 public servants and contractors who emphasized the importance of augmenting models with "in-house" knowledge and described that organizational pressures lead to the production of more transparent models).

48 For the "accountability by design" framing, see Margot E. Kaminski, *Binary Governance: Lesson's from GDPR's Approach to Algorithmic Accountability*, 92 S. Cal. L. Rev. 24, n.125 (forthcoming 2020). For those who advocate a move away from individual-level conceptions of transparency or remedial approaches, *see* Joshua A. Kroll et al., *Accountable Algorithms,* 165 U. Pa. L. Rev. 633, 661 (2017); Deven R. Desai & Joshua A. Kroll, *Trust But Verify: A Guide to Algorithms and the Law*, 31 Harv. J.L. & Tech. 1, 5 (2017); Lillian 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 (2017); Tal Z. Zarsky, *Transparent Predictions*, 2013 U. Ill. L. Rev. 1503 (2013); Mike Ananny & Kate Crawford, *Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability*, 20 New Media & Soc'y 973, 980 (2018). "Accountability by design" is a riff on "privacy by design," an influential movement in privacy law circles to stimulate a "philosophy and approach of embedding privacy in the design specifications of various technologies." *See* Ann Cavoukian, *Privacy by Design* 1 (2009), https://www.ipc.on.ca/wp-content/uploads/Resources/7foundationalprinciples.pdf. Part of the explanation for this trend in thinking is the impossibility of full transparency over a specific decision's provenance. Part of it arises out of a standard set of observations about the limits of private, litigation-centered enforcement by rights-bearing individuals. *See, e.g.,* David Freeman Engstrom, *Agencies as Litigation Gatekeepers,* 123 Yale L.J. 616 (2013) (reviewing literature on pros and cons of private enforcement); Stephen B. Burbank & Sean Farhang, Rights and Retrenchment: The Counterrevolution Against Federal Litigation (2017) (reviewing constriction of private enforcement as a regulatory mode through legislation, procedural rulemaking, and judicial decisions). But the trend also grows out of a more general recognition that one-off, ex post challenges to decisions, even if numerous and leveled at regular intervals, may not reach systemic sources of error and so may not be as effective as internally driven, critically reflective system design at the outset, before a model is running, and continued systemic monitoring, testing, and experimentation thereafter. *See* Kaminski, *supra*, at 1558; Kroll et al., *supra*, at 640. For a more general argument for why individual, rights-based enforcement may be insufficient to correct systemic error within mass adjudicatory systems, *see* 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). Deirdre K. Mulligan & Kenneth A. Bamberger, *Saving Governance-by-Design*, 106 Cal. L. Rev. 697, 759 (2018) ("Existing governance institutions often lack these tools, and substantive regulatory capacity—breadth of authority, competence, and vision on the one hand, and expertise on the other—must be built to support the rational use of technology to govern.").

49 Among the tools are: organizing code into testable modules; writing and running test cases; and incorporating code that crashes a system when it encounters an error rather than continuing in an errant state and automatically generates audit logs. *See* Kroll et al., *supra* note 48, at 644-56; *see also* Danielle Keats Citron, *Technological Due Process*, 85 Wash. L. Rev. 1249, 1277, 1305, 1310 (2008) (offering a similar set of prescriptions, including coding of "audit trails" and testing prior to implementation). Testing can take static (observational) and dynamic (testing with natural and synthetic inputs) forms. One of the more interesting arguments for "internal" constraints on algorithmic decision-making is that, while marquee uses of algorithmic decisions systems—*e.g.,* the criminal justice context—will draw certain judicial scrutiny. Indeed, they already have. But algorithmic decision-making in less hot-button areas may evade review, creating a slow burn of biased or arbitrary decisions that, in aggregate, exact a significant toll but are not salient enough to attract judicial or other challenges. *Cf.* Citron, *supra*, at 1256.

50 The classic statement is Jerry L. Mashaw, Bureaucratic Justice: Managing Social Security (1985). More recent statements include Gillian Metzger & Kevin M. Stack, *Internal Administrative Law*, 115 Mich. L. Rev. 1239 (2017); Christopher J. Walker, Administrative *Law Without Courts,* 65 UCLA L. Rev. 1620 (2018); Adrian Vermeule, Law's Abnegation: From Law's Empire to the Administrative State (2016); Elizabeth Magill, *Foreword: Agency Self-Regulation*, 77 Geo. Wash. L. Rev. 859 (2009).

Transparency and Accountability

51 This norm pervades American administrative law, both in the Administrative Procedure Act, see 5 U.S.C. § 557(c)(3)(A) (2018) ("All [agency] decisions [with respect to procedures requiring a hearing] . . . shall include a statement of . . . findings and conclusions, and the reasons or basis therefor . . . ."), and in judicial decisions, see Judulang v. Holder, 565 U.S. 42, 45 (2011) ("When an administrative agency sets policy, it must provide a reasoned explanation for its action."); FCC v. Fox Television Stations, Inc., 556 U.S. 502, 515 (2009) (noting "the requirement that an agency provide reasoned explanation for its action"). Similar versions can be found in many Western legal systems. *See* Henrik Palmer Olsen et al., *What's in the Box? The Legal Requirement of Explainability in Computationally Aided Decision-Making in Public Administration*, 162 iCourts Working Paper Series 14-22 (2019).

52 *See, e.g.,* Jenna Burrell, *How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms*, 3 Big Data & Soc'y 1 (2016).

53 A more robust accounting of a decision's provenance would also convey the minimum change necessary to yield a different outcome and provide an explanation for similar cases with different outcomes and different cases with similar outcomes. *See* Finale Doshi-Velez et al., *Accountability of AI Under the Law: The Role of Explanation*, Cornell U. (Nov. 3, 2017), https://arxiv.org/abs/1711.01134. For a real-world example, the Fair Credit Reporting Act requires credit reporting firms to disclose to consumers the four factors driving their credit score ranked in order of significance. 15 U.S.C. § 1681g(f)(1) (2018).

54 In addition, and contrary to common perception, algorithmic tools are human-machine "assemblages," not self-executing creations. Ananny & Crawford, *supra* note 48, at 983. Analysts must make myriad decisions, including how to partition data, what model types to specify, what datasets, target variables, 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). It can thus be difficult to pinpoint whether arbitrary or biased outputs result from tainted code and data, or from numerous other human-made design choices. Andrew Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 Fordham L. Rev. 678 (2018); Kroll et al., *supra* note 48, at 679-82..

55 Selbst & Barocas, *supra* note 54, at 43, 64. For similar efforts to categorize specific and systemic modes of explanation, see Edwards & Veale, *supra* note 48, at 55-59; 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).

56 Aaron Rieke, Miranda Bogen, & David G. Robinson, Upturn & Omidyar Network, Public Scrutiny of Automated Decisions: Early Lessons and Emerging Methods 18 (2018).

57 Selbst & Barocas, *supra* note 54, at 64.

58 *See* Coglianese & Lehr, *Transparency*, *supra* note 16, at 4.

59 Ananny & Crawford, *supra* note 48, at 983; Citron, *supra* note 49, at 1249-54; John Danaher, *The Threat of Algocracy: Reality, Resistance and Accommodation*, 29 Phil. & Tech. 245, 257 (2016); Virginia Eubanks, Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor (2018); Cathy O'Neil, Weapons of Math Destruction (2017); Tal Z. Zarsky, *Automated Predictions: Perception, Law, and Policy*, 15 Comms. ACM 35 (2012).

60 An example of the latter is a ceiling on the number of "terminal leaves" in a random-forest machine learning model. Selbst & Barocas, *supra* note 54, at 33.

61 Dillon Reisman et al., AI Now, Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability (2018).

62 For a similar "hard" versus "soft" formulation, see Ananny & Crawford, *supra* note 48, at 976.

63 EU General Data Protection Regulation (GDPR): Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ 2016 L 119/1.

64 15 U.S.C. § 1681 (2018) (containing guidance on topics such as permissible purposes of such credit reports (1681b), disclosure rules (1681d/f-g/u-w), liability for noncompliance (1681n-o), administrative enforcement (1681s), etc.).

65 Andrew Tutt, *An FDA for Algorithms*, 69 Admin. L. Rev. 83 (2017) (wherein the author proposes an independent government agency designed specifically to ensure the safety and efficacy of algorithms distributed in the US, much like the FDA's mandate to promote safety and efficacy in drugs and medical devices).

66 Desai & Kroll, *supra* note 48, at 46.

67 The classic formulation is that society can regulate "entry" or "results." *See* Sam Issacharoff, *Regulating After the Fact*, 56 DePaul L. Rev. 375 (2007); Steven Shavell, *Liability for Harm Versus Regulation of Safety,* 3 J. Legal Stud. 357 (1984).

68 L. Jason Anastasopoulos & Andrew B. Whitford, *Machine Learning for Public Administration Research, With Application to Organizational Reputation*, 29 J. Pub. Admin. Res. & Theory 491, 506 (2019); Selbst & Barocas, *supra* note 54, at 32.

69 *Id.*

70 *See*, *e.g.,* Citron, *supra* note 49; Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. Rev. 93 (2014); Andrew Guthrie Ferguson, *Big Data and Predictive Reasonable Suspicion*, 163 U. Pa. L. Rev. 327, 329-30 (2015).

71 *See*, *e.g.,* Citron, *supra* note 49; Cary Coglianese & David Lehr, *Regulating by Robot: Administrative Decision Making in the Machine-Learning Era*, 105 Geo. L.J. 1147, 1198 (2017); 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. Parrilo ed., 2017).

72 *See* Kaminski, *supra* note 48, at 15.

73 *See* Heckler v. Chaney, 470 U.S. 821 (1985) (holding that agency decisions not to enforce are not subject to review); Fed. Trade Comm'n v. Standard Oil Co., 449 U.S. 232, 242 (1980) (holding that an agency's decision to proceed with an enforcement action is not immediately challengeable).

74 Kroll et al., *supra* note 48, at 699.

75 Jacob Gersen & Adrian Vermeule, *Thin Rationality Review*, 114 Mich. L. Rev. 1355 (2016) (advocating for lighter touch "thin" rationality review).

76 Catherine M. Sharkey, *State Farm "With Teeth": Heightened Judicial Review in the Absence of Executive Oversight*, 89 N.Y.U. L. Rev. 1589, 1592 (2014) (advocating for heightened scrutiny to independent agencies not subject to executive oversight, including the SEC).

77 5 U.S.C. § 552a (2018).

78 5 U.S.C. § 552(b)(4) & (7).

79 Sonia K. Katyal, *The Paradox of Source Code Secrecy*, 104 Cornell L. Rev. 101, 109 (2019); Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System,* 70 Stan. L. Rev. 1343 (2018).

80 48 C.F.R. 12.212 (2018).

81 470 U.S. 821 (1985).

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

83 The natural analogy here is Inspectors General (IG) offices or "offices of goodness" and 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); Neal Kumar Katyal, *Internal Separation of Powers: Checking Today's Most Dangerous Branch from Within*, 115 Yale L.J. 2314 (2006).

84 *See* David Freeman Engstrom & Daniel E. Ho, *Algorithmic Accountability in the Administrative State,* 37 Yale J. on Reg. (forthcoming 2020).

Bias, Disparate Treatment, and Disparate Impact

85 James Zou & Londa Schiebinger, *AI Can Be Sexist and Racist—It's Time to Make It Fair,* 559 Nature 324 (2018).

86 Julia Angwin, Jeff Larson, Surya Mattu & Lauren Kirchner, *Machine Bias*, ProPublica, May 23, 2016.

87 Jeff Dastin, *Amazon Scraps Secret AI Recruiting Tool that Showed Bias against Women*, Reuters, Oct. 9, 2018.

88 *See* Alexandra Chouldechova, F*air Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments*, 5 Big Data 153 (2017); Sorelle A. Friedler, Carlos Scheidegger & Suresh Venkatasubramanian, *On the (Im)possibility of Fairness*, Cornell Univ. (2016), https://arxiv.org/abs/1609.07236.

89 Sandra G. Mayson, *Bias In, Bias Out*, 128 Yale L.J. 1-5 (2019, forthcoming); Sam Corbett-Davies & Sharad Goel, *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning*, Cornell Univ. (2018), https://arxiv.org/abs/1808.00023.

90 Joy Boulamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 Proc. Machine Learning Res. 77 (2018).

91 U.S. Dep't of the Treasury, Annual Privacy, Data Mining, and Section 803 Report 22 (2017).

92 Alexandra Chouldechova et al., *A Case Study of Algorithm-Assisted Decision Making in Child Maltreatment Hotline Screening Decisions*, 81 Proc. Machine Learning Res. 1 (2018); Virginia Eubanks, *A Child Abuse Prediction Model Fails Poor Families,* Wired, Jan. 15, 2018.

93 Michael A. Livermore, Vladimir Eidelman & Brian Grom, *Computationally Assisted Regulatory Participation*, 93 Notre Dame L. Rev. 977 (2017).

94 Mark A. Lemley & Bhaven Sampat, *Examiner Characteristics and Patent Office Outcomes*, 94 Rev. Econ. & Stat. 817 (2012).

95 Solon Barocas, Moritz Hardt & Arvind Narayanan, Fairness and Machine Learning: Limitations and Opportunities 41-42 (2018).

96 Cynthia Dwork et al., "Fairness through Awareness." *Proceedings of the 3d Innovations in Theoretical Computer Science Conference*, ACM (2012); Jon Kleinberg et al., "Algorithmic Fairness," *AEA Papers and Proceedings*, vol. 108 (2018).

97 Grutter v. Bollinger, 539 U.S. 306 (2003).

98 Gratz v. Bollinger, 539 U.S. 244 (2003).

99 Grutter, 539 U.S. at 334.

100 L.A. Water & Power v. Manhart, 435 U.S. 702 (1978).

101 Wisconsin v. Loomis, 881 N.W.2d 749 (2016).

102 Sonja B. Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 Stan. L. Rev. 803 (2014).

103 Women's Equity Action League v. Cavazos, 906 F.2d 742, 750-51 (D.C. Cir. 1990).

104 Cristina Ceballos, David Freeman Engstrom, Daniel E. Ho, & Derin McLeod, *Disparate Limbo* (Working Paper, 2020).

105 For an illustration of how reliance on reported information can exacerbate bias, see Kristen M. Altenburger & Daniel E. Ho, *When Algorithms Import Private Bias into Public Enforcement: The Promise and Limitations of Statistical Debiasing Solutions*, 175 Journal of Institutional and Theoretical Economics 98-122 (2018).

106 If this has a substantial effect on small start-up businesses, the tool could come into conflict with the Regulatory Flexibility Act.

107 *See* Engstrom & Ho, *Algorithmic Accountability*, *supra* note 84.

Hearing Rights and Algorithmic Governance

108 *See, e.g.,* Goss v. Lopez, 419 U.S. 565 (1975) (requiring minimal due process before suspending a student).

109 As noted in Part II's case studies of the SSA and PTO, the APA creates two types of adjudication: those subject to §§ 554-557 of the APA, and those subject to § 555 alone. *See* 5 U.S.C. § 554-557 (2012). The former cluster of APA provisions imposes three main requirements that apply in so-called "formal" adjudications: agency notice to affected persons; the opportunity for those persons to submit facts and arguments at a "hearing" before the agency, whether live or on papers; and, presiding at any live hearing an administrative law judge, or ALJ. We previously noted that these hearings are often referred to as "Type A" hearings. "Informal" adjudication is governed by the minimum requirements of Due Process and § 555, which requires that an agency allow affected persons to be represented by counsel and issue subpoenas for a party. We previously noted that these hearings are often referred to as "Type C" hearings. A third category of hearings, referred to previously as "Type B" hearings, are typically defined in an agency's organic statute and can vary significantly in their procedural requirements, from something approaching Type A's trial-like proceeding to something far more "informal" in procedural make-up.

110 Tom R. Tyler, *What Is Procedural Justice?: Criteria Used by Citizens to Assess the Fairness of Legal Procedures* 22 L. & Soc. Rev. 103, 128 (1988). *See generally* Tom R. Tyler, Whey People Obey the Law (2006).

111 Jerry Mashaw *Administrative Due Process: The Quest for a Dignitary Value*, 61 Bos. L. Rev. 885 (1981).

112 Mireille Hildebrandt, *Law as Computation in the Era of Artificial Legal Intelligence: Speaking Law to the Power of Statistics*, 68 U. Toronto L.J. 12, 21-22 (2018). *See also* Daniel Marovits, A Modern Legal Ethics: Adversary Advocacy in a Democratic Age (2008).

113 R. Parasuraman & D.H. Manzey, *Complacency and Bias in Human Use of Automation: An Attentional Integration*, 52 Hum. Factors 381, 391 (2010); Linda J. Skitka, Kathleen L. Mosier, & Mark Burdick, *Does Automation Bias Decision-Making?*, 51 Int'l. J. Hum.-Computer Stud. 991 (1991) Citron, *supra* note 49, at 1271-72. Conversely, one could discuss algorithm appreciation, which wanes when subjects are experts and when poised against one's own judgment. *See* Jennifer M. Logg, Julia A. Minson & Don A. Moore, *Algorithm appreciation: People Prefer Algorithmic to Human Judgment*, 151 Organizational Behavior and Human Decision Processes 90–103 (2019).

114 Berkeley J. Dietvorst, Joseph P. Simmons, & Cade Massey, *Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err*, 144 Journal of Experimental Psychology: 114–126 (2015); Berkeley J. Dietvorst, Joseph P. Simmons & Cade Massey, *Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them*, 64 Management Science 1155–1170 (2018); Dilek Onkal et al., *The Relative Influence of Advice From Human Experts and Statistical Methods on Forecast Adjustments*, 22 Journal of Behavioral Decision Making 390–409 (2009); Ben Green & Yiling Chen, *The Principles and Limits of Algorithm-in-the-Loop Decision Making*, 3 Proc. ACM Hum.-Comput. Interact. 1, 3 (2019); RM Dawes, D Faust & PE Meehl, *Clinical Versus Actuarial Judgment*, 243 Science 1668 (1989).

115 Silvia Bonaccio & Reeshad S. Dalal, *Advice Taking and Decision-Making: An Integrative Literature Review, and Implications for the Organizational Sciences*, 101 Organizational Behavior and Human Decision Processes 129-130 (2006).

116 Jerry L. Mashaw, Bureaucratic Justice: Managing Social Security Disability Claims (1983).

117 Ames, et al., *supra* note 48.

118 Citron, *supra* note 49, at 1249; Mathews v. Eldridge, 424 U.S. 319 (1976).

119 Londoner v. City and County of Denver, 210 U.S. 373 (1908).

120 Bi-Metallic Investment Co. v. State Board of Equalization, 239 U.S. 441 (1915).

121 461 U.S. 458 (1983) (upholding the adoption of medical-vocational guidelines that classified whether jobs exist in the national economy based on physical ability, age, education, and work experience of the claimant).

122 *Id.* at 467.

123 *See* Office of the Inspector Gen., Soc. Sec. Admin., A-12-18-50353, Audit Report: The Social Security Administration's Use of Insight Software to Identify Potential Anomalies in Hearing Decisions D1-2 (2019).

124 Eugene Volokh, *Chief Justice Robots,* 68 Duke L.J. 1135 (2019); Michael A. Livermore and Daniel N. Rockmore, *Introduction: From Analogue to Digital Legal Scholarship*, in Law as Data: Computation, Text, & the Future of Legal Analysis at xiv (Livermore & Rockmore, eds., 2019).

125 *See* Benjamin Alarie, *The Path of the Law: Toward Legal Singularity*, 66 U. Toronto L.J. 443 (2016).

126 *See* Anthony J. Casey & Anthony Niblett, *The Death of Rules and Standards*, 92 Ind. L. Rev. 1401 (2017); Anthony J. Casey & Anthony Niblett, *Self-Driving Laws*, 66 U. Toronto L.J. 429, 431 (2016).

127 *See* Brian Sheppard, *Warming Up to Inscrutability: How Technology Could Challenge Our Concept of Law*, 68 U. Toronto L.J. 36, 40 (2018).

128 *See, e.g.,* Aziz Z. Huq, *A Right to a Human Decision*, 105 Va. L. Rev. (forthcoming 2020).

129 Hildebrandt, *supra* note 112, at 21.

130 *Id.* at 22.

131 Casey & Niblett, Death, *supra* note 127; Mireille Hildebrandt, *Law as Information in the Era of Data-Driven Agency*, 79 Mod. L. Rev. 1,1 (2016) (decrying "mindless agency" of machine learning algorithms); *Joshua P. Davis, Law Without Mind: AI, Ethics, and Jurisprudence* (working paper 2018), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3187513.

Gaming and Adversarial Learning

132 *See, e.g.,* Anthony D. Joseph, Blain Nelson, Benjamin I.P. Rubinstein, & J.D. Tygar, Adversarial Machine Learning (2019); Daniel Lowd & Christopher Meek, *Adversarial Learning*, Proceedings of the 11th ACM SIGKDD Int'l Conf. on Knowledge Discovery in Data Mining 641, 641 (2005), https://dl.acm.org/citation.cfm?id=1081950.

133 *See* Jane R. Bambauer & Tal Zarsky, *The Algorithm Game*, 94 Notre Dame L. Rev. 10 (2018).

134 *Id.*; Julie E. Cohen, Configuring the Networked Self: Law, Code, and the Play of Everyday Practice 256 (2012).

135 *See* Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy, *Explaining and Harnessing Adversarial Examples*, Cornell Univ. (2014), https://arxiv.org/abs/1412.6572; Vikas Sehwag et al., *Not All Pixels Are Born Equal: An Analysis of Evasion Attacks under Locality Constraints*, ACM SIGSAC Conf. on Computer and Comms. Sec. 2285, 2285 (2018). *See generally* Engstrom & Ho, *Algorithmic Accountability*, *supra* note 84.

136 *See* Brian Christian & Tom Griffiths, Algorithms to Live By: The Computer Science of Human Decisions 157-58 (2016).

137 Bambauer & Zarsky, *supra* note 134, at 11.

138 Cf. Edward K. Cheng, *Structural Laws and the Puzzle of Regulating Behavior*, 100 Nw. U.L. Rev. 655, 671 (2006).

139 *See* Art Unit Analysis, Serco, https://sercopatentsearch.com/art-unit-analysis.

140 It is important to note that distributive effects are not limited to the gaming or the enforcement context. Regulatory mechanisms designed to achieve transparency and accountability can likewise have a distributive cast. As just one example, erasure rights—that is, the "right to be forgotten" at the heart of the GDPR—are not costless to invoke and may be far more likely exercised by well-heeled individuals with an economic incentive to expunge negative information. Put another way, some citizens may be better equipped than others to take advantage of the epistemic benefits of AI technologies. John Danaher, *The Threat of Algocracy: Reality, Resistance and Accommodation*, 29 Phil. & Tech. 235, 262 (2016).

141 *See* Peter Loewen, Algorithmic Government (University of Toronto 2019) (powerpoint presentation), https://static1.squarespace.com/static/58ecfd18893fc019a1f246b4/t/5ce388f929c4e80001f25bd3/1558415625030/Halbert_Loewen.pdf.

142 *See* Bambauer & Tvarsky, *supra* note 134, at 14-15.

143 Kroll, et al., *supra* note 48, at 654.

144 *See* Ian Goodfellow et al., *Generative Advesarial Networks*, Proceedings of the Int'l Conf. on Neural Information Processing Sys. (2014), 2672–2680; 37 C.F.R. § 1.56 (2018).

The External Sourcing Challenge: Contractors and Competitions

145 Oliver E. Williamson, *Public & Private Bureaucracies: A Transaction Cost Perspective*, 15 J.L. Econ. & Org. 306, 319 (1999).

146 For general literature on contracting out, including historical perspective on its evolution, *see* John Donohue, The Privatization Decision: Public Ends, Private Means (1991); Jon Michaels, Constitutional Coup: Privatization's Threat to the American Republic (2017); Verkuil, *supra* note 3.

147 Interview with Kurt Glaze, *supra* note 46.

148 James Ridgeway, Presentation at "A Roundtable Discussion on the Use of Artificial Intelligence in the Federal Administrative Process," NYU School of Law (Feb. 25, 2019). Ridgeway's comments seem an apt description of the Patent and Trademark Office's Sigma tool which was developed internally, but was never fully deployed after it was found to only improve efficiency for examiners with a computer science background. Arti Rai, *Machine Learning at the Patent Office: Lessons for Patents and Administrative Law*, 104 Iowa L. Rev. 2626-37 (2019); Thomas A. Beach, *Japan Patent Information Organization Presentation: USPTO Bulk Data,* U.S. Patent & Trademark Off. 22-25, http://www.japio.or.jp/english/fair/files/2016/2016e09uspto.pdf (last visited Feb. 23, 2019).

149 As noted elsewhere in Part III, security researchers have proved capable of defeating complex systems in short timeframes, as exemplified by the strategic placement of stickers in the view of Tesla's autopilot system causing the vehicle to unexpectedly change lanes. *See* Nicholas Carlini & David Wagner, *Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods*, Cornell U. (Nov. 1, 2017), https://arxiv.org/abs/1705.07263; Evan Ackerman, *Three Small Stickers in Intersection Can Cause Tesla Autopilot to Swerve into Wrong Lane*, IEEE Spectrum (Apr. 1, 2019), https://spectrum.ieee.org/cars-that-think/transportation/self-driving/three-small-stickers-on-road-can-steer-tesla-autopilot-into-oncoming-lane.

150 Reinhold, *supra* note 32.

151 *See* Verkuil, *supra* note 3, at 140-50; Levin & Tadelis, *supra* note 4, at 508.

152 Kroll et al., *supra* note 48, at 701.

153 *See* John D. Donahue, *The Transformation of Government Work: Causes, Consequences, and Distortions,* in GOVERNMENT BY CONTRACT: OUTSOURCING AND AMERICAN DEMOCRACY 41, 49 (Jody Freeman & Martha Minow eds., 2009); Jody Freeman & Martha Minow, *Introduction* in *id*., 2. Creaming occurs when a service provider strategically privileges subjects for interventions who generate the greatest increase within agreed-upon metrics. Shirking (or, in economics, shading) is where a contractor takes advantage of fuzzy contract terms by reducing quality in ways that violate the spirit but not the letter of the contract. Note that this discussion maintains generality to yield simplicity. However, the public management literature further distinguishes between different types of contracting and outputs—for instance, "public-private contracting" (defined as private provision of governance services) and "direct service provision" (defined as private provision of services to third-party beneficiaries). *See, e.g.,* Jody Freeman, *The Contracting State*, 28 FLA. ST. L. REV. 155, 165 (2000).

154 *See* Wendy Netter Epstein, *Contract Theory and the Failures of Public-Private Contracting*, 34 CARDOZO L. REV. 2211, 2222 (2013).

155 There has been some work on extracting figures and charts from PDF documents, as well as *classifying* those figures (for example, to distinguish bar-charts from pie-charts). However, current AI techniques are not equipped to understand what a figure represents. *See* Yan Liu et al., *Review of Chart Recognition in Document Images*, VISUALIZATION AND DATA ANALYSIS SPIE 1 (2013).

156 *See The Serco IP Difference*, SERCO, https://sercopatentsearch.com/why-us (last visited Dec. 15, 2019).

157 TEST SUMMARY REPORT, *supra* note 38, at 7.

158 Levin & Tadelis, *supra* note 4, at 529 (finding politicians and public manager preference for in-house production where quality matters).

159 The Intergovernmental Personnel Act Mobility Program, for example, "provides for the temporary assignment of personnel between the Federal Government and state and local governments, colleges and universities, Indian tribal governments, federally funded research and development centers, and other eligible organizations." *See Hiring Information: Intergovernment Personnel Act*, U.S. OFFICE OF PERS. MGMT., https://www.opm.gov/policy-data-oversight/hiring-information/intergovernment-personnel-act/ (last visited Dec. 15, 2019). Academic collaborations include the FDA's "Entrepreneur-in-Residence" program in 2017 as part of its Digital Health Innovation Action Plan, and the EPA's partnership with the University of Chicago Energy and Environment Lab to supply research fellows. *See* Food and Drug Admin., Digital Health Action Plan 7 (2017), https://www.fda.gov/downloads/MedicalDevices/DigitalHealth/UCM568735.pdf; *Our People*, UNIV. CHI. URB. LABS, https://urbanlabs.uchicago.edu/people/sarah-armstrong. Examples of academic partnerships include a collaboration between the University of Michigan and the United States Postal Service on automated mail delivery, and work between Stanford University, the University of Chicago, and EPA to develop environmental enforcement tools.

160 *See generally*, Suzanne B. Schwartz, Assoc. Dir. for Sci. & Strategic Partnerships, , Ctr. for Devices & Radiological Health, U.S. Food & Drug Admin., *The Medical Device Ecosystem and Cybersecurity––Building Capabilities and Advancing Contributions* (Nov. 1, 2019), https://www.fda.gov/NewsEvents/Newsroom/FDAVoices/ucm624749.htm.

161 *See Examining the GM Recall and NHTSA's Defect Investigation Process: Hearing before the Subcomm. on Consumer Prot., Prod. Safety, & Ins. of the S. Comm on Commerce, Sci. & Transp.,* 113th Cong. 65 (2014) (testimony of David J. Friedman, Acting Administrator, National Highway Traffic Safety Administration).

162 *Environmental Protection Agency, Agenda for EPA-State SNC National Compliance Initiative, Symposium II*, Jan. 16, 2020, https://www.acwa-us.org/wp-content/uploads/2020/01/Final-Agenda-for-SNC-Conference-II-01162020.pdf (noting "EPA-Stanford University collaboration").

163 For instance, the Department of Veterans Affairs developed a strategy in its health care partnership with Alphabet's DeepMind that uses cryptographic hashes to obscure veterans' sensitive personal information and thus permit data-sharing. Simonite, *supra* note 25.

164 U.S. Food & Drug Admin., MOU 225-12-0010, Memorandum of Understanding between the United States Food and Drug Administration and Massachusetts Institute of Technology (2012), http://www.fda.gov/AboutFDA/PartnershipsCollaborations/MemorandaofUnderstandingMOUs/AcademiaMOUs/ucm318476.htm [https://perma.cc/DUA8-9KM4].

165 Press Release, Scott Gottlieb, Comm'r, U.S. Food & Drug Admin., *FDA's Comprehensive Effort to Advance New Innovations: Initiatives to Modernize for Innovation* (Aug. 29, 2018), https://www.fda.gov/NewsEvents/Newsroom/FDAVoices/ucm619119.htm. In offering guidance for third-party organizations in the context of 510(k) certification, the FDA (1) articulates factors it considers when determining whether a device is eligible for Third Party Review, U.S. Food & Drug Admin., 510(k) Third Party Review Program: Draft Guidance for Industry, Food and Drug Administration Staff, and Third Party Review Organizations 10–12 (2018), and (2) describes the Third Party Review process. *Id.* at 12–16. The Agency is working with stakeholders in "the field of radiogenomics, where AI algorithms can be taught to correlate features on a PET or MRI scan with the genomic features of tumors." *Id*.

166 At the federal level in the United States, competitions were recently collected together into a single website, www.challenge.gov. For discussion, see Kevin D. Desouza & Ines Mergel, *Implementing Open Innovation in the Public Sector: The Case of Challenge.gov*, 73 PUB. ADMIN. REV. 882 (2013). For local government use of competitions, see Edward Glaeser et al., *Crowdsourcing City Government: Using Tournaments to Improve Inspection Accuracy* (Nat'l Bureau Econ. Res., Working Paper No. 22124, 2016), https://www.nber.org/papers/w22124.

167 There could also be industry, or even cross-agency, competitions and perhaps the manufacturers that uploaded new data or solutions to new technical problems could be rewarded with reduced agency scrutiny or with access to an agency maintained dataset. *See, e.g.,* 83 Fed. Reg. 50872 (Oct. 10, 2018); *see also Docket ID: NHTSA-2018-0092, Regulations.Gov*, https://www.regulations.gov/docket?D=NHTSA-2018-0092 (last visited Dec. 15, 2019).

168 Mary E. Gallo, Cong. Res. Serv., Federal Prize Competitions (2018) https://fas.org/sgp/crs/misc/R45271.pdf. NASA's 2014 Disruption Tolerant Networking Challenges produced technology that will serve as a "basis for security in future space communications architectures," for a "remarkably low" cost. *See* Office of Sci. & Tech. Policy, Implementation of Federal Prize Authority: Fiscal Year 2014 Progress Report 12 (2015), https://www.challenge.gov/assets/document-library/FY2014-Implementation-Federal-Prize-Authority-Report.pdf.

169 *Budget*, CBO, https://www.cbo.gov/topics/budget (last visited Nov. 30, 2019).

170 *See* Press Release, U.S. Food & Drug Admin., *FDA and DHS Increase Coordination of Responses to Medical Device Cybersecurity Threats under New Partnership; A Part of the Two Agencies' Broader Effort to Protect Patient Safety* (Oct. 16, 2018), https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm623574.htm; Food & Drug Admin., Memorandum of Agreement between the Department of Homeland Security, National Protection and Programs Directorate and the Department of Health and Human Services, Food and Drug Administration, Relating to Medical Device Cybersecurity Collaboration (2018), https://www.fda.gov/AboutFDA/PartnershipsCollaborations/MemorandaofUnderstandingMOUs/DomesticMOUs/ucm623568.htm.

171 This team could then liaise with individual agency experts to build custom interfaces with specific datasets and models for each distinct use case. The pooling of agency budgets for this central team could also potentially allow them to offer much more competitive contracts as compared to the large software companies.

172 AI in Government Act, H.R. 2575, 116th Cong. (introduced May 8, 2019). At the FDA, the Precertification or "Pre-Cert" umbrella works on regulating and approving AI, while its Information Exchange and Data Transformation ("INFORMED") focuses primarily on the FDA's internal use of AI.