Artificial Intelligence for Adjudication: The Social Security Administration and AI Governance[1]

Kurt Glaze
Social Security Administration

Daniel E. Ho
Stanford University

Gerald K. Ray
Social Security Administration (retired)

Christine Tsang
Stanford University

Abstract

Despite widespread skepticism of data analytics and artificial intelligence (AI) in adjudication, the Social Security Administration (SSA) pioneered path breaking AI tools that became embedded in multiple levels of its adjudicatory process. How did this happen? What lessons can we draw from the SSA experience for AI in government?

We first discuss how early strategic investments by the SSA in data infrastructure, policy, and personnel laid the groundwork for AI. Second, we document how SSA overcame a wide range of organizational barriers to develop some of the most advanced use cases in adjudication. Third, we spell out important lessons for AI innovation and governance in the public sector. We highlight the importance of leadership to overcome organizational barriers, “blended expertise” spanning technical and domain knowledge, operational data, early piloting, and continuous evaluation. AI should not be conceived of as a one-off IT product, but rather as part of continuous improvement. AI governance is quality assurance.

Keywords: artificial intelligence, social security, adjudication, innovation, administrative law

[1] Affiliations are for identification only. This research does not represent any official views or opinions of the SSA. This article draws on prior work by two of the coauthors (Bajandas & Ray, 2018; Engstrom et al., 2020; Engstrom & Ho, 2020; Ray & Lubbers, 2014; Ray & Sklar, 2019). Ray was an Administrative Appeals Judge and Deputy Executive Director of the Office of Appellate Operations at SSA and spearheaded numerous initiatives documented herein. Glaze is a Program Analyst and the Creator and Product Owner of the Insight Software at SSA. We thank Nikita Aggarwal, Justin Bullock, Johannes Himmelreich, Sonoo Thadaney Israni, Michael Matheny, and Padmashree Gehl Sampath for useful comments.

Table of Contents

I. Introduction

The use of Artificial Intelligence (AI) in adjudication is controversial. Popular press accounts are chock full of alarmist accounts of "robo-judges" replacing humans (e.g., Sayer, 2016). How can the adjudicatory process, which is fundamentally concerned with tailoring law to circumstance, rely on *automated* decisions? France has gone as far as to criminalize the use of judicial analytics (Tashea, 2019). The skepticism of the use of analytics in adjudication echoes an earlier wave of critiques of quantitative approaches to judicial behavior (Edwards and Livermore 2008; Tushnet 1980).

Yet there is one adjudicative system in the United States that has been able to use AI to help its judges and attorneys make core adjudicative decisions: the Social Security Administration (SSA) Disability Program. In its most ambitious form, SSA has developed and deployed an automated AI system that enables adjudicators to check draft decisions for roughly 30 quality issues, addressing long-standing questions about the accuracy, consistency, and speed of case processing (Ames et al., 2020).

How did this come to be? How did "the largest adjudication agency in the western world" (Barnhart v. Thomas, 540 U.S. 20, 28–29 (2003)) overcome well-known structural challenges in the public sector to climb this "Agency Everest" to become a poster child for AI innovation in government? In this chapter, we tell the story of how SSA overcame significant roadblocks to develop and implement AI use cases and draw policy implications for AI innovation. This account contributes to three central questions in administrative adjudication, AI governance, and public administration. First, the story is important for understanding how to develop AI in large organizations and specifically in the challenging context of mass adjudication, such as immigration adjudication, Medicare appeals, veterans benefits determinations, and patent examination (Ho, 2017). The SSA case study illustrates the process by which AI can be deployed to advance, not undermine, due process goals in adjudication.

Second, the SSA case study has broad lessons for AI governance in the public sector. A recent executive order establishes guidance for federal agencies regarding the adoption of AI and its use in the delivery of services.[2] It commits federal agencies to accelerate the adoption of AI in ways that will modernize government and cultivate public trust in AI. Likewise, the U.S. National AI Initiative anchors itself around principles for trustworthy AI. The SSA case holds important lessons for turning AI governance principles into practice in one of the most contested terrains. Most importantly, it shows the importance of what we call "blended expertise" -- i.e., expertise at the intersection of

[2] Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government, Exec. Order No. 13,960, 85 Fed. Reg. 78,939 (2020).

domain and technical knowledge -- to identify, develop, and test new AI-powered innovations in a way consistent with governing law and policy (Engstrom et al., 2020; Engstrom & Ho, 2020).

Third, our case study also informs our understanding of public administration and innovation (Bovens & Zouridis, 2002; Busch & Henriksen, 2018; Cinar et al., 2019). This literature has examined conditions for innovation, the technological transformation of street-level bureaucrats to "screen-level" or "system-level" bureaucrats (Bovens & Zouridis, 2002; Bullock et al., 2020), and the impact of AI systems on public administration (Criado et al., 2020; Young et al., 2019). Our case study confirms the importance of where the core organization responsible for innovation is located (Moldogaziev & Resh, 2016) and the involvement of the end users of a system in development (Criado et al., 2020). While some have theorized that AI is most impactful for tasks with low levels of discretion (Bullock et al., 2020; Young et al., 2019), we examine the use of AI to improve a highly complex, discretion-laden area of public administration: adjudication.

This chapter proceeds as follows. Section II will provide background on the SSA Disability Claims System. Section III discusses the historical challenges in the accuracy of decision making. Section IV discusses the groundwork of electronic case management, case analytics, and policy development that enabled the agency to pilot AI use cases. Section V discusses how the SSA grappled with technology governance and innovation barriers when first piloting these use cases. Section VI discusses AI applications that SSA has developed for their adjudication program. Section VII draws lessons for AI governance from SSA's experience, focusing on data infrastructure, leadership support, blended expertise that spans domain and technical knowledge, the software environment, and continuous iteration. Section VIII concludes.

II. The Social Security Disability Claims System

Under the Social Security Act, SSA provides disability benefits to individuals between the ages of 18 and 65 who meet the insurance requirements of the program and are unable to work because of a disability. The system for making disability determinations at SSA comprises the largest adjudicatory system in the United States (Ames et al., 2020), which paid over $200B to 18M Americans in Fiscal Year 2020. A successful claimant will typically receive around $270,000 in lifetime benefits, plus Medicare coverage (Gelbach & Marcus, 2016).

The Social Security Act defines "disability" as the "inability to engage in any substantial gainful activity by reason of any medically determinable physical or mental impairment which can be expected to result in death or . . . last for [at least] twelve months" (42 U.S.C. § 423(d)(1)(A)). SSA uses a five-step sequential evaluation process for ascertaining disability:

1. Is the claimant engaged in substantial gainful activity? If yes, the claim is denied.

2. Does the claimant have "a severe and medically determinable physical or mental impairment" or combination of impairments lasting of sufficient duration? If no, the claim is denied.

3. Does the impairment meet the severity of roughly two hundred listed impairments? If yes, the claimant is disabled.

4. Does the claimant retain "residual functional capacity" to perform any past relevant work? If no, the claimant is disabled.

5. Is the claimant able to perform any other work existing in significant numbers in the national economy, based on the claimant's residual functional capacity, age, education, and work experience? If no, the claimant is disabled.

The adjudicatory process of the disability claims system works in four stages. First, the State Disability Determination Service ("DDS") processes the application and makes an initial determination about whether the applicant is disabled (20 C.F.R. §§ 404.900, 902). Second, a claimant can request reconsideration of this initial determination and submit additional evidence (20 C.F.R. §§ 404.904, 913). Third, after receiving the results of the reconsideration, a claimant can request a hearing with an Administrative Law Judge (ALJ) for de novo review of their claim and submit additional evidence not available at the time of prior proceedings (20 C.F.R. §§ 404.915, 916, 935). Roughly 1,500 ALJs across 162 hearing offices preside in these due process hearings. Medical and vocational experts are sometimes called upon to provide opinions (HALLEX I-2-5-34, I-2-6-74), and the claimant has the right to appear at a hearing (20 C.F.R. § 404.930). Last, the Appeals Council, which together with its staff comprises the Office of Appellate Operations (OAO), considers appeals of ALJ decisions and represents the final level of appeal within the agency. Although claimants may further pursue their claims in federal district court, the vast majority of decisions are resolved internal to SSA. Most of due process, in that sense, functionally plays out inside the agency (Mashaw 1985; 1973).

III. Historic Challenges

Because determinations are both factually and legally complicated, the disability claims system has faced serious challenges. A seminal study found that wide disparities between ALJs appeared driven by subjective factors (Mashaw et al., 1978), and such disparities continue to present day (Engstrom & Ho, 2020). Between 2008-2019, most claimants waited over a year for an appeal to be resolved (Government Accountability Office, 2020). Nearly 110k applicants died during that period prior to receiving a final disability decision (id.). The high volume of cases has made quality improvement and the incorporation of federal court opinions very challenging (Ames et al., 2020; Gelbach & Marcus, 2017).

While procedural due process mandates "accuracy" of decisions, the accuracy and quality of decision-making is hard to verify. Employees are trained to follow agency policy, but many policies are open to interpretation. This problem can be compounded for claims involving medical evidence, which is also subject to the interpretation of medical service providers. The volume of casework and the need to adjudicate quickly can further affect quality, as individual adjudicators develop heuristic shortcuts that they employ in case processing, which may lead to incorrect outcomes in some cases.

Heuristics, training, lack of a clear quality standard, and gaps in or loosely written policy guidance all contribute to the high variability in decision-making. Quality is often measured with a bottom-line approach, meaning whether the ultimate outcome is *acceptable* within the statutory and regulatory scheme (Ames et al., 2020). For example, SSA's hearings process uses the term "legally sufficient" to describe a decision of adequate quality, with the hearing level procedural manual specifying that decisions should be accurate and legally sufficient (HALLEX I-2-8-1). Other markers of quality, such as remand and grant review rates of hearing level decisions, can themselves be impacted by the variability of the reviewers of the decisions, and may do little to address the variability in hearing decisions (Gelbach & Marcus, 2017; Mashaw, 1973).

IV. Foundational Infrastructure

In light of these challenges, how did SSA become the pioneer for pathbreaking use cases of AI in adjudication? We document here some of the foundational steps -- data infrastructure, policy clarification, and analytics personnel -- that facilitated the adoption of AI tools. Many of the steps were taken long before SSA considered AI applications. While these steps gave SSA a leg up when considering AI use cases, we emphasize that these are not necessary predicates for considering all types of AI. Many pilots can be conducted in parallel to IT modernization, but taking these steps with downstream AI use cases in mind can be highly beneficial.

A. Digitizing and Systematizing Workflow

The efforts of the SSA to systematize and digitize its core workflows through electronic systems created highly valuable data and data infrastructure for later AI applications. Between the 1990s and 2000s, SSA implemented several electronic case management systems (eCMS) to organize its case activities and developed electronic folders to store digitized copies of claim evidence related to each claim. It also built out case analysis tools to structure and record staff notes and analysis about a claim's merits.

In the mid-2000s, early efforts to build tools to improve the quality and consistency of adjudication emerged. SSA created an electronic case analysis tool (eCAT), an electronic questionnaire that guided adjudicators at the DDS through policy compliant pathing to reach a disability determination, while capturing structured data. At the same time, SSA built a similar tool directly into the eCMS for the Appeals Council. Known as the Appeals Council Analysis Tool (ACAT), the tool used a similar electronic form with questions to guide adjudication while capturing structured data about disability decisions. Appeals Council Members used the existing structure of SSA policy to develop a decision tree which mapped the policy compliant pathing to approximately 2,000 types of decisions possible under the sequential evaluation process. This decision tree was used in creating ACAT (Ray & Lubbers, 2014).

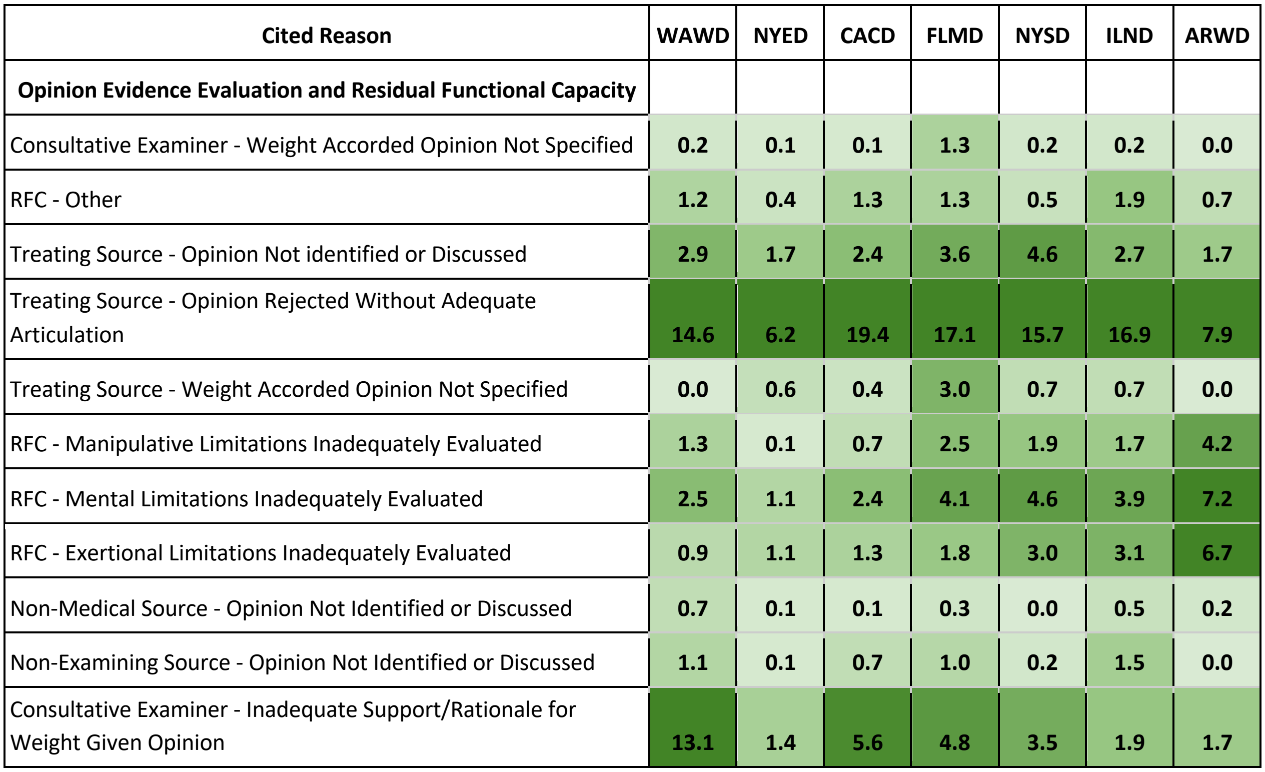
Prior to ACAT, analysts assisting the Appeals Council Members used a wide variety of forms for disability case analysis. The bulk of the analysis consisted of a written description of potential remandable issues and a recommended course of action for the Appeals Council Members to take. The free-form nature of these analyses may have contributed to variability in adjudication.

The Appeals Council attempted to address this variability issue by more narrowly defining a quality decision as one that is factually accurate, procedurally adequate, supported by the record, and policy compliant. The Appeals Council used ACAT to bring a more uniform structure to this analysis and, importantly, was able to obtain access to the data captured in ACAT and the new electronic case management system. The Council began analyzing this data to provide feedback to Appeals Council Members to encourage greater consistency in adjudication. Ultimately this effort provided a clear baseline for evaluating the quality of disability analysis and later proved important for the development of the quality flags used by the Insight tool described below.

B. Structured Policies and Procedures

Over the years, SSA developed an increasingly structured process for evaluating disability claims. For example, court decisions on the use of vocational expert evidence led the agency to develop a series of Medical-Vocational guidelines that directed a conclusion of “disabled” or “not disabled” based on factors such as the claimant’s residual functional capacity, age, education and prior work experience. These guidelines, known colloquially as “the Grid Rules,” take administrative notice of information contained in the Dictionary of Occupational Titles (20 C.F.R. § 404.1569, and Appendix 2 to Subpart P of Part 404), reducing dependence on vocational experts and improving consistency of adjudication. SSA also developed a sequential evaluation process for disability claims described in Section II. Class action litigation in the 1980s led to other refinements in policy structure, particularly related to the evaluation of subjective complaints and medical opinions.

To further refine SSA’s policies and procedures, OAO began taking steps to leverage their access to data and reliable data infrastructure to improve the quality and consistency of adjudication. For instance, ACAT included data about why cases were remanded by the Appeals Council and federal courts. OAO generated data visualizations, such as heat maps that applied color-coding to identify trends and easily observe the frequency of error types across different hearing offices (see Table 1). This data-backed approach allowed OAO to quickly identify the policies generating the most substantial adjudicatory errors. Executives also addressed several circuit and district court judicial conferences and described the differences in district court behaviors to the judges, most of whom had never seen detailed data about their adjudications.



Cited Reason	WAWD	NYED	CACD	FLMD	NYSD	ILND	ARWD
Opinion Evidence Evaluation and Residual Functional Capacity							
Consultative Examiner - Weight Accorded Opinion Not Specified	0.2	0.1	0.1	1.3	0.2	0.2	0.0
RFC - Other	1.2	0.4	1.3	1.3	0.5	1.9	0.7
Treating Source - Opinion Not identified or Discussed	2.9	1.7	2.4	3.6	4.6	2.7	1.7
Treating Source - Opinion Rejected Without Adequate Articulation	14.6	6.2	19.4	17.1	15.7	16.9	7.9
Treating Source - Weight Accorded Opinion Not Specified	0.0	0.6	0.4	3.0	0.7	0.7	0.0
RFC - Manipulative Limitations Inadequately Evaluated	1.3	0.1	0.7	2.5	1.9	1.7	4.2
RFC - Mental Limitations Inadequately Evaluated	2.5	1.1	2.4	4.1	4.6	3.9	7.2
RFC - Exertional Limitations Inadequately Evaluated	0.9	1.1	1.3	1.8	3.0	3.1	6.7
Non-Medical Source - Opinion Not Identified or Discussed	0.7	0.1	0.1	0.3	0.0	0.5	0.2
Non-Examining Source - Opinion Not Identified or Discussed	1.1	0.1	0.7	1.0	0.2	1.5	0.0
Consultative Examiner - Inadequate Support/Rationale for Weight Given Opinion	13.1	1.4	5.6	4.8	3.5	1.9	1.7

Table 1: Sample heat map indicating the cited reasons for remands (in rows) of cases across district courts (in columns). The numbers represent the percentage of cases remanded by each district court based on the reason cited. Darker shading indicates a more commonly cited reason for remand. The acronym “RFC” is used in place of the term ‘residual functional capacity,” an SSA-used term describing the capabilities of a disability applicant after consideration of the limitations caused by their medical impairments.

The availability of this type of granular data enabled “focused reviews” of critical problem areas and informed training programs for adjudicators. Focused reviews provided specific information about how an ALJ evaluated the evidentiary record and applied agency policies and procedural guidance, and helped OAO staff identify training issues related to the misinterpretation or misapplication of policy guidance. OAO found that nearly all adjudicative errors were inadvertent. The errors were likely caused by heuristics that adjudicators adopted as shortcuts to skip aspects of policy-compliant pathing and generally still reach a policy-compliant result. Occasionally, however, the shortcuts resulted in noncompliant decisions. Training materials on reasons for remand were made available online and enabled adjudicators to self-study and close gaps in their knowledge.

OAO staff also addressed issues prone to misinterpretation where policy guidance was either unclear or insufficiently precise. OAO staff proposed policy and procedural changes that would aid adjudicators in understanding and correctly applying the policy. Staff researched the history and legislative intent of laws, background information and historical changes to regulations, as well as memoranda, legal opinions and procedural guidance. OAO also undertook hundreds of changes to procedural guidance for hearings and appeals operations (in a manual known as the HALLEX). Clarifying its policies and procedures was important to SSA’s development of rule-based AI that required such specificity for its structure and user adoption. Combined, these policy efforts laid the groundwork for some of the AI features SSA later developed.

V. Overcoming Organizational and Personnel Barriers in Technology Governance

The path of AI innovation seen at the SSA’s Office of Appellate Operations was possible not because of SSA's existing institutional structure, but rather in spite of it: OAO embarked on a campaign of “stealth innovation.”

As documented in public sector innovation scholarship, agency culture and climate can often impede innovation (Cinar et al., 2019). Securing support for new ideas and projects can be challenging. At the highest levels, there are many competing requests for limited resources, and most resources are allocated to maintain existing processes.

Resource constraints cause new project ideas to be deferred or dropped altogether unless executives are persistent across multiple budget cycles. Even if an executive can get a project off the ground, such as by explaining how the project might improve staff productivity or the timeliness or quality of work, an immediate focus on measurable and reportable results can limit the development and exploration of new and longer-term opportunities that might flow from the initial concept.

After early requests for staffing were rebuffed, OAO decided to neither advertise the efforts it was undertaking nor seek resources for these projects. This approach provided OAO with the latitude to fully explore a range of ideas without being worried about the results, provided OAO met business objectives.

Without budget or additional headcount, OAO first set about freeing up resources for their work by improving the productivity of its existing team. OAO established numeric productivity performance standards for the professional staff, publishing internal branch goals for productivity and timeliness, and reorganizing case flow. An Appeals Council Member identified performance measures for twelve categories of work activities. The performance measures were based on the types of actions taken and hours worked over a two year period, and baselines were set for successful and outstanding performance levels based on these measures. OAO also developed and implemented continuing learning techniques, reducing the time for trainees to become fully productive from 18 months to 5 months. These efforts yielded a significant rise in the productivity of the staff, from 94 case dispositions per staff member in Fiscal Year (FY) 2009 to 146 in FY 2013, to 161 dispositions per staff member in FY 2017.

As productivity rose, OAO gained some latitude to branch out and move existing resources into more ambitious projects. OAO started slowly by borrowing the services of one data scientist from another SSA component, actively recruited lawyers with data science backgrounds, and developed a summer intern program to temporarily hire law students to assist with the development of training material. OAO then reprogrammed staff to address policies and procedures and expand training, and eventually added more staff to data analytics efforts.

It was only after demonstrating some success -- by improving productivity and timeliness and developing compelling data visualizations -- that OAO began to describe externally what it had done. OAO then sought permission for a small budgetary allocation to expand its quality assurance efforts and supplemented the allocation by also reprogramming much of its attrition hiring into quality assurance. The quality assurance staff assisted with data analytics, potential fraud investigation and analysis, and policy analysis and formulation projects. Ultimately, OAO deployed roughly 6% of its staff to long-range, exploratory analytics projects not previously undertaken by the agency. They also performed the critical, labor intensive work required to develop modern AI tools for adjudication, like data labeling, which required high-caliber subject matter experts well-versed in the law. Repurposing trained analytical staff in this way was risky, particularly in light of increasing caseloads and staff turnover (due to a large retirement wave), but OAO was able to achieve its operational goals and continually reduce the size and age of its pending workload during this time.

OAO also sought to develop internal capacity that traversed technical and domain specific knowledge, an approach somewhat inconsistent with agency norms. Agencies tend to organize around function, with component parts of the agencies specializing in performing certain tasks, which can lead to a siloing effect that can be detrimental to agency functioning and innovation in particular (Cinar et al., 2019). Performance plans, promotion paths, and bargaining unit agreements all pressure individuals to stay in their lane to succeed and advance. This dynamic of narrow focus operates not only at the component office level but also at the person-level. Agency employees are provided with a specific position description outlining the duties of their job. Work outside of this scope generally can only be performed for short periods of time, as part of an official detail to perform those duties. Often such details must be announced and employees selected through an open competitive process in accordance with bargaining unit agreements.

OAO took a few important steps. First, OAO focused on hiring and cultivating individuals with blended expertise. To overcome restrictions that limited OAO to hiring attorneys, OAO identified existing and new attorneys with backgrounds in statistics, mathematics, econometrics, computer science, and adult education. OAO assembled these individuals into teams to address questions of policy, training, data analytics, and the innovative use of technology. These cross-cutting teams acquired domain expertise by adjudicating cases, while at the same time developing ideas on how to deploy analytics more effectively. In addition, OAO borrowed the services of several SSA data scientists and operational research specialists to assist in cleaning, summarizing, classifying and analyzing the data captured by analysts and adjudicators using ACAT and other data sets as they became available.

Second, personnel were given substantial space to explore a range of use cases. Such leeway ultimately enabled the group to strategically sequence use cases that would best align with the mission of SSA, namely to improve caseload production and the quality of decisions issued. Federal agencies like SSA often view IT projects as largely finite in development effort and cost: a large, heavily resourced development team executes a project roadmap to deliver a feature-complete product. After delivery, a much smaller team maintains the system, capable of making necessary changes but not building significant new features. Federal budgetary, contracting, and IT reporting requirements may reinforce this approach (Rubenstein, 2021). Projects with less clear end states may be viewed as running counter to these norms. And AI projects that support or make decisions are much more likely to require substantial and continuous attention and evolution as policies, business processes, and even operational norms evolve. Put differently, the AI system may require continuous resources akin to a human staff that require continuous training during their tenure (Casado & Bornstein, 2020).

Third, OAO developed some more open-ended position descriptions that provided managers with more flexibility for assigning duties, particularly among the clerical and support staff positions. The combination of improving technology and more efficient support staff enabled the redeployment of some positions into analytical jobs, which also improved productivity and performance. The flexibility in some of the position descriptions also provided the opportunity to deploy employees into policy analysis work, which included analysis of data and information. Additionally, executives and managers worked with bargaining unit representatives to creatively extend details by keeping the employees active in normal job duties part of the time and when performing overtime work.

The result of these efforts was to embed data scientists and attorneys with technical facility in an operational environment. The initiative was not without risk -- as the use of staff time could be seen as a drain on case production -- but OAO made a bet that data analytics would, in the long run, improve the quality, consistency, and timeliness of disability adjudication, reducing errors and the reworking of cases.

Despite OAO’s success, there is still no clear promotional path to encompass the types of activities performed by blended teams like those it developed. The success of such teams may be poorly recognized outside of the highest executive channels, while line managers struggle with managing employees who may have been responsible for large business value in an area outside of the duties normally performed in that position. Promotional paths available to an employee working within specific position descriptions may lead the employee away from the very work they so successfully performed, while work more closely aligned with what they accomplished may be out of their reach or fall within another siloed component over which the employee’s manager has little leverage in the promotional process. In short, inventing around barriers is not ideal.

VI. AI Use Cases

A. Structured Learning for Workload Management

With a strong foundation of data infrastructure, policy, and personnel, OAO made a number of focused bets designed to use data science to improve the efficiency and quality of adjudication. One prototype was based on the notion that it would be easier for adjudicators to consider cases involving similar issues together. If similar cases could be assigned in batches, adjudicators might recognize the similarities in issues and require less time researching the relevant regulations and policies. The question was therefore whether a simple reordering of the assignment of work could lead to significant gains in efficiency.

To identify cases with similar characteristics without first reviewing the case files, Appeals Council Members and OAO staff worked closely with data scientists to develop a clustering analysis of the pending workload. This was accomplished by using structured data from ACAT and other hearing office level data to train algorithms that sorted cases into small batches with similar characteristics. Since the agency has interpreted the Administrative Procedure Act to preclude specialized units from processing cases by case type, the cases were worked in the usual manner by the same employees who otherwise would have worked them, just not in the same order. This project, though implemented for a limited time, appeared to reduce processing time and the need to rework erroneous cases.

OAO also worked to develop a naive Bayes supervised learning model of pending hearing level workloads. The analysis was designed to estimate the probability of an award of benefits based solely on certain characteristics found in the metadata captured in the hearing level eCMS and ACAT. The project was later extended to identify claims dismissed for procedural reasons that otherwise would have resulted in an award of benefits. Probability of outcomes were predicted but were not shared with adjudicators so as to not prejudice outcomes. Cases with higher probabilities of allowance generally were worked before other cases to speed processing for claimants most likely to be found disabled. SSA officials reported that the model overestimated the number of cases likely to be allowed (10% of cases as compared with the average fully favorable rate of 2.5-3%), but was useful in moving likely allowances ahead in the pending workload queue.

Other parts of SSA have similarly begun to integrate machine learning. SSA created a lab using Hadoop to store and analyze data quickly. It also developed the Quick Disability Determination (QDD) process which uses a predictive model to identify cases involving one or more impairments that usually result in disability. The QDD process enabled the agency to skip resource-intensive hearings when cases were likely to result in an award.

B. AI to Support Adjudication: Insight Software

Entrepreneurs within the agency with both subject matter and technical expertise were critical in spurring AI innovation by devising an ambitious suite of decision support tools known as Insight.[3] Adjudicative staff at the hearings and appeals levels of adjudication work with written decision documents. Staff generate the basic structure of the decision using a template system and then manually add substantive findings and rationale. For decades, the well-established practice was for staff to work on these decisions independently in assembly line fashion. For instance, an ALJ would prepare instructions directing the content of a decision, which would then be handed off to a decision writer to independently transform the document into a draft decision, who then handed off the completed draft back to the ALJ for independent review.

SSA leveraged Insight to improve this siloed approach: With a click, staff can now analyze their decision document and receive alerts on potential quality issues as they work, plus receive a variety of case-specific reference information and tools enriched by what Insight found in the decision's content. At the hearing level, staff use Insight to analyze draft decisions, enabling them to evaluate and react to Insight's quality feedback prior to issuance. At the appeals level, staff use Insight to analyze issued hearing decisions under their review, helping to ensure they identify and evaluate all potential quality issues prior to making a recommendation to appellate judges. Importantly, Insight is explicitly designed only as an assistive tool: It does not decide any element of a decision nor advise any specific remedy to potential quality issues. Rather, staff are trained and even explicitly reminded in the interface that Insight's content is to serve only as a jumping off point for further analysis.

Insight's features require several AI technologies to function. First, Insight applies natural language processing (NLP) to extract information from the written decision, such as details of its findings and rationale. Insight then retrieves existing structured data about the case and claimant from workload systems (e.g., claimant claim history and biographical data, etc.). Using this more complete picture, Insight applies both rule-based and probabilistic machine learning algorithms to identify potential quality issues.

[3] Co-author Kurt Glaze, an SSA attorney with an interest in computer science, devised and pitched Insight to OAO in 2015.

Insight has been fully deployed to adjudicative staff at the appeals level since late 2017 and the hearings level since late 2018. Internal studies by SSA of Insight's effect on adjudication have found that its use is associated with improved work quality (e.g., improved rates of quality issue remediation during drafting, improved quality issue recognition on appeal, etc.) and more efficient case processing.

The foundations described in Section IV were essential to Insight's development and operational success. First, SSA's existing case processing system provided data that enabled Insight developers to target specific pools of historic hearing decisions to streamline the assembly of labeled training data for machine learning features. The data infrastructure also enabled quality checks that rely on access to 'ground truth' outside of decisional text to function. For example, a quality check of whether an age-related regulation cited in the decision is in fact applicable to the claimant requires the claimant's date of birth as stored in the case processing system.

Second, policies that rigidly structure the findings and content of disability decisions have been essential to the success of Insight's NLP and logic. For instance, the sequential evaluation process spelled out in Section II explicitly defines the findings that must appear in a decision and their sequence. This structure is manifested in decision templates used by staff (e.g., the template reliably outputs a 'Step 1' finding prompt followed by a 'Step 2' prompt, etc.). The predictability created by these policies greatly simplified the development of high precision information extraction.

Third, personnel flexibilities championed by OAO leadership enabled an adjudicative staffer to 'pitch' Insight to SSA leadership and then have the flexibility outside regular legal duties to pursue its full development and release.

Finally, and critically, SSA ultimately invested millions of dollars over multiple years to transform Insight from essentially an 'under the desk' proof of concept to a much more full featured, enterprise-class system capable of supporting thousands of users. This enabled Insight to access highly skilled software development staff and contractors, as well as secure additional time from numerous highly knowledgeable business staff to help further guide the project and complete necessary data annotation tasks.

While SSA reports that Insight has improved quality and productivity, formal evaluations of the impact of the Insight system on accuracy and remand rates have been limited (Ames et al., 2020; Office of the Inspector General, Social Security Administration, 2019). There is great potential for more rigorous evaluation and harnessing of more recent advances in AI. Many Insight quality flags, for instance, rely on more simple forms of machine learning and do not yet take advantage of the most important developments in deep learning with natural language processing.

VII. Lessons

The SSA case study illustrates broad organizational and personnel challenges with AI innovation in the public sector. While OAO maneuvered around these bureaucratic impediments, AI innovation will require leveling these barriers more systematically. We hence spell out more general lessons to foster an improved ecosystem for AI innovation, accountability, and governance in the public sector.

A. Leadership Support and Blended Expertise

A chief lesson from SSA is the critical role played by leadership at OAO to drive forward agency capacity to learn from its own data. Many of the key moves --- capturing data, formalizing policy into an adjudicatory decision tree, leveraging individuals with blended expertise, making the strategic choice to invest in early analytics projects --- would not have occurred in the absence of strategic leadership at the top.

SSA's experience also shows the deep value of 'nexus' resources -- those with both business and technical expertise -- in driving AI innovation. First, nexus resources accelerate the speed of AI innovation. Leveraging their deep understanding of business operations, they are able to rapidly evaluate the potential value of an innovation along with its policy and cultural acceptability. Leveraging their technical expertise, they can often take major steps toward building and prototyping the innovation. By breaking the need to coordinate to devise, build, and validate, nexus resources tighten iterative development cycles, resulting in projects that fail or succeed much faster.

Second, nexus resources can increase the likelihood innovations will succeed. At federal agencies like SSA, significant funding for an innovation project requires a successful presentation of a well-researched business case to an investment review board. However, many ideas for innovation are complex to substantiate. While large organizations often have technical teams with the remit to pursue test cases, they are generally reliant on contractors or staff who are unfamiliar with the day-to-day business functions to which the innovation relates. The result is a knowledge gap that can delay or even stifle the substantiation of AI innovations. Nexus resources can bridge this gap by using their blended in-house expertise to brainstorm across functional teams, discover operational insights, and immediately build out a prototype.

Yet organizational structure often stands in the way. Many federal government positions are designed around specific sets of skills, with career tracks and promotions driven largely by work activity within those set areas. Blended expertise fits uncomfortably within these hired duties.

Leaders of organizations interested in AI innovation should consider how they can proactively structure their human resources to secure and foster nexus resources. First, more open-ended duties in position descriptions could provide the flexibility needed to leverage staff with nexus resources by retasking them to technical tasks under those duties. Getting position descriptions right will be critical for AI innovation (Engstrom et al., 2020; National Security Commission on Artificial Intelligence, 2021). Second, many agencies have drawn on partnerships with academic institutions -- through vehicles like the Intergovernmental Personnel Act or Cooperative Research and Development Agreements -- to bring in technical and nexus resources (Engstrom et al., 2020). One of the core challenges lies in the fact that AI innovation is occurring at a furious pace, and such partnerships, sabbaticals, and exchanges can leverage the core AI talent at research universities and the domain expertise of agencies.

B. The Value of Operational Data

Data fuels AI innovation. SSA's Insight software is proof of the significant value earned from data describing the Disability Program's workflows and decision making generated in applications such as the eCMS. Even this data is only a foothold on a longer climb: If decision-facing AI innovations like Insight are ever to meet or exceed the accuracy and breadth of human counterparts, they need access to the same level of information as them. Moreover, robust data on workflows and decision making is often necessary to meaningfully achieve several principles of AI governance under Executive Order 13,690, such as accuracy, effectiveness, and traceability. It is difficult to evaluate if an AI system outperforms existing processes or is 'accurate' without robust data cataloguing the operations it targets.

To achieve this, organizations can take steps to digitize operational activity and decision making through systems akin to SSA's eCMS. They also should strongly consider the passive collection of data, such as logging what is done and for how long in public and staff websites and other software. Passive data collection is highly cost effective because it requires no direct action to generate. Passive data also enables operational analyses -- including analyses of the effectiveness of AI innovations[4] -- whose probative strength and granularity are not possible without it. Indeed, leading private companies

[4] For example, if an organization implemented an AI system to automatically update customer mailing addresses based on the free text within various inbound mail, they may well want to benchmark how human staff perform that task. Without scaled passive data recording, benchmarking would likely consist of observing a small number of examples or simply asking staff about the task, both of which may miss critical insights into as-is performance.

not only passively collect numerous forms of data, they act on it in real time.[5] While highly valuable, organizations should be cognizant that its collection is likely to raise ethical and even legal questions for those from whom it collects.[6] For example, staff-facing efforts could be introduced with commitments from management that the data would not be used for performance evaluations.

Rich operational data may be essential to identifying roadblocks to reaping value from AI innovations. For many AI applications whose outputs are engaged with by human staff, the AI's value will turn on whether it supports a task space that has a broadly agreed-upon structure and meaning by its users. For example, AI tools that classify animals in photos would have little global appeal if the rules of what constituted a cat versus a dog were subject to the whims of each user. Rich operational data enables organizations to meaningfully and efficiently evaluate how consistently staff perceive a task before investing in an AI application to support it.

For example, In SSA's case, data collected from its eCMS and case analysis systems enabled leaders to identify operational inconsistencies and target them through policy clarifications and training, ultimately improving adjudicative consistency. This process was incredibly valuable to Insight's acceptance by users, as they were more likely to share and agree with Insight's applications of policy in its quality checks. Organizations should be ready to consider pursuing some level of 'upstream' policy and process reforms to promote consistency in a business task as needed before injecting supportive AI into it.

C. Test Early and Often

SSA's experience shows that testing AI innovations through pilots or limited releases can be a valuable means to evaluate the value of the innovation and build organizational buy-in. For example, SSA released Insight to a subset of OAO adjudicative staff in early 2017 for voluntary use in their cases. SSA then studied how Insight impacted OAO operations, which in part showed that Insight use had no material impact on case processing efficiency -- a concern prior to its release. SSA also surveyed staff who tried Insight, and a large majority of respondents indicated they

[5] For example, Amazon introduced AI camera systems into their delivery truck fleet to monitor drivers for potential traffic violations and distraction, e.g. drowsiness. Drivers must sign consent forms acknowledging this monitoring. Amazon reports that the AI technology has contributed to massive improvements in driver performance: "[A]ccidents decreased 48 percent, stop sign violations decreased 20 percent, driving without a seatbelt decreased 60 percent, and distracted driving decreased 45 percent." https://www.theverge.com/2021/3/24/22347945/amazon-delivery-drivers-ai-surveillance-cameras-vans-consent-form

[6] For example, the European Union's General Data Protection Regulation gives individuals from whom businesses collect personal information (including web analytics data) numerous privacy rights. See https://gdpr.eu/tag/chapter-3/

found Insight feedback to be accurate and easy to interpret. These results were critical to OAO's decision to expand Insight's release and provided objective support for SSA leadership outside OAO to increase the Insight project’s funding to expand its features and scale to the hearing level. These sorts of evaluations will yield valuable insights into how AI systems operate in practice, including the potential for unanticipated effects (Engstrom & Ho, 2020).

Although there have been some efforts to define the word “pilot,” we caution that the term should not automatically trigger layers upon layers of agency or congressional review. Review should be calibrated based on the risk posed. AI systems to augment existing quality improvement programs, for instance, may be precisely the kind of internal organizational decisions for which management flexibility is warranted.

D. Continuous Iteration and Evaluation

The SSA experience also demonstrates that public sector AI innovation is a process, not a product, requiring continuous analysis, evaluation, and iteration.

Consider the use of supervised learning. Recent advances in deep learning are particularly appealing because logical AI systems --- models based on hard-coded conditional logic --- are challenging to scale, given the thousands of fact scenarios evaluated under complex and often vague policy rules. Supervised learning, instead, requires only labeled examples of a targeted decision (e.g., “disabled” vs. “not disabled”).

Yet precisely because of the historical challenges in decisional accuracy, securing enough high quality labels to train models will be an ongoing process. Many federal agencies may be facing an explosion in their capacity to collect and analyze operational data, due to technologies such as cloud infrastructure, handwriting recognition, and speech-to-text. Through these advances, agencies may be beginning to learn far more about the quality and consistency of their past decision making. For example, with the advancement of technologies that can extract detailed data from claimant medical records, SSA may identify disparities in outcomes among highly factually similar historic claims, such as may be caused by inconsistent heuristics used by staff. Any such anomalies in past actions may equate to a ceiling for the performance of supervised learning models based on them, and agencies may not be comfortable with that ceiling.

If systemic errors, biases, or inconsistencies are exposed, one might be tempted to simply filter them out. However, a more viable path may be to use what is uncovered to improve *future* decision making through a *combination* of analysis, early stage AI, and human judgment. For example, analysis could reveal inconsistencies among staff in applying a particular regulation. Targeted training or human-in-the-loop AI features could improve consistency in this area. With a combination of targeted improvements, the quality and consistency of human decisions can often be improved and thus offer a better position from which to train a supervised AI system. This approach acknowledges the clear reality that human decision making may include flaws and instances of bias, as well as many virtuous attributes. For example, SSA’s disability adjudicators may make observable mistakes, but they also make thousands of discrete decisions that deftly navigate complex policies and medical fact settings. After a period of AI-augmented, well-analyzed human performance, agencies may finally have a series of historic decisions that are of sufficient quality to train a responsible decision-*making* AI system.

However they approach their historic actions, organizations reading SSA's example and seeing AI as a one-time panacea are cautioned to consider it as part of continuous improvement. Even if an AI-based tool does no worse than an existing human baseline, such AI-based tools may themselves generate the impetus for further performance improvement. Indeed, the logic of *Mathews v. Eldridge* --- which would point to the lower administrative burden of reducing error with AI-based tools --- may demand it.

E. Development and Deployment Ecosystem

Organizations often develop software using multiple development environments. Prior to release to a ‘production’ environment (real world operations), a ‘validation’ environment is often used to test features using ‘mock’ data designed to be structurally equivalent to production data without corresponding to any real world entity. However, this paradigm can impose several significant disadvantages particularly harmful to data hungry AI innovations. First, the creation of ‘mock’ data can be a slow, request-based process whose size does not compare to production. Second, the ‘mock’ data may not be complete or faithfully represent complexities in the underlying data. For example, agencies may find it difficult to meaningfully mock unstructured data such as legal motions or customer service transcripts.

This infrastructure is not optimal for the rapid, realistic, and massive scale experimentation and testing of new AI features. AI development often requires significant computational resources, the use of various open source software packages, and the full feature space, enabling better modeling and error analysis. Agencies should pursue infrastructure that will allow teams to safely but easily prototype and test new innovations at scale against real-world production data. For instance, agencies could establish a ‘data warehouse’ containing a replica of production system data that development teams could ingest from to experiment and validate new AI features.

Additionally, organizations should consider how their broader software ecosystem enables and integrates with AI use cases. AI innovations do not work in isolation; their outputs must be presented to users or effectuated within existing systems to generate value integration into existing systems. But existing systems often are not designed with integration in mind, forcing AI teams to spend significant time and money on 'plumbing' issues unrelated to their core objectives. To help AI innovations scale without this structural friction, organizations should consider requiring core systems to be engineered to facilitate rapid extension. For example, a workload system could offer a secure API endpoint that would enable a calling system to programmatically make changes to individual work items or push custom content to a portion of the user interface for review by users.

VIII. Conclusion

An evolution toward integrating AI into the recurring workloads of large organizations may be inevitable. SSA has nearly 60,000 employees, with administrative costs exceeding $6 billion annually.[7] Every merits decision in SSA's Disability Program is made by human staff with naturally varying levels of expertise and perceptions of often vague policies. Insight's release has proven that AI partnered with human staff can help counter these drawbacks to generate significant improvements in performance over humans alone.

SSA's example teaches us that for any AI system to achieve good governance, organizational reforms to enable invention, reflection, and assessment will be critical.

[7] https://www.ssa.gov/oact/STATS/admin.html

References

Ames, D., Handan-Nader, C., Ho, D. E., & Marcus, D. (2020). Due Process and Mass Adjudication: Crisis and Reform. *Stan. L. Rev.*, *72*, 1.

Bajandas, F. F., & Ray, G. K. (2018). *Implementation and Use of Electronic Case Management Systems in Federal Agency Adjudication*.

Bovens, M., & Zouridis, S. (2002). From Street-Level to System-Level Bureaucracies: How Information and Communication Technology is Transforming Administrative Discretion and Constitutional Control. *Public Administration Review*, *62*(2), 174–184.

Bullock, J., Young, M. M., & Wang, Y.-F. (2020). Artificial intelligence, bureaucratic form, and discretion in public service. *Information Polity*, *25*(4), 491–506. https://doi.org/10.3233/IP-200223

Busch, P. A., & Henriksen, H. Z. (2018). Digital discretion: A systematic literature review of ICT and street-level discretion. *Information Polity: The International Journal of Government & Democracy in the Information Age*, *23*(1), 3–28. https://doi.org/10.3233/IP-170050

Casado, M., & Bornstein, M. (2020). The New Business of AI (and How It’s Different From Traditional Software). *Andreesen Horowitz*, 2019–2020.

Cinar, E., Trott, P., & Simms, C. (2019). A systematic review of barriers to public sector innovation process. *Public Management Review*, *21*(2), 264–290.

Criado, J. I., Valero, J., Villodre, J., Giest, S., & Grimmelikhuijsen, S. (2020). Algorithmic transparency and bureaucratic discretion: The case of SALER early warning system. *Information Polity: The International Journal of Government & Democracy in the Information Age*, *25*(4), 449–470. https://doi.org/10.3233/IP-200260

Edwards, H. T., & Livermore, M. A. (2008). Pitfalls of empirical studies that attempt to understand the factors affecting appellate decisionmaking. *Duke LJ*, *58*, 1895--1989.

Engstrom, D. F., & Ho, D. E. (2020). Algorithmic accountability in the administrative state. *Yale J. on Reg.*, *37*, 800.

Engstrom, D. F., Ho, D. E., Sharkey, C. M., & Cuéllar, M.-F. (2020). Government by algorithm: Artificial intelligence in federal administrative agencies. *Report to the Administrative Conference of the United States*.

Gelbach, J. B., & Marcus, D. (2017). Rethinking Judicial Review of High Volume Agency Adjudication. *Tex. L. Rev.*, *96*, 1097.

Gelbach, J. B., & Marcus, D. (2016). A Study of Social Security Disability Litigation in the Federal Courts. *Final Report to the Administrative Conference of the United States*, 16–23. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2821861

Government Accountability Office. (2020). *Social Security Disability: Information on Wait Times, Bankruptcies, and Deaths among Applicants Who Appealed Benefit Denials* (GAO-20-641R).

Ho, D. E. (2017). Does Peer Review Work: An Experiment of Experimentalism. *Stanford Law Review*, *69*, 1–120.

Mashaw, J. L. (1973). Management Side of Due Process: Some Theoretical and Litigation Notes on the Assurance of Accuracy Fairness and Timeliness in the Adjudication of Social Welfare Claims. *Cornell L. Rev.*, *59*, 772.

Mashaw, J. L. (1985). *Bureaucratic justice: Managing social security disability claims*. Yale University Press.

Mashaw, J. L., Goetz, C. J., & Carrow, M. M. (1978). *Social security hearings and appeals: A study of the Social Security Administration hearing system*. Lexington Books.

Moldogaziev, T. T., & Resh, W. G. (2016). A systems theory approach to innovation implementation: Why organizational location matters. *Journal of Public Administration Research and Theory*, *26*(4), 677–692.

National Security Commission on Artificial Intelligence. (2021). *Final Report*.

Office of the Inspector General, Social Security Administration. (2019). *The Social Security Administration's Use of Insight Software to Identify Potential Anomalies in Hearing*

Decisions (A-12-18-50353).

Ray, G. K., & Lubbers, J. S. (2014). 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. *Geo. Wash. L. Rev.*, *83*, 1575.

Ray, G. K., & Sklar, G. (2019). An Operational Approach to Eliminating Backlogs in the Social Security Disability Program. *McCrery-Pomeroy SSDI Solutions Initiative*.

Rubenstein, D. S. (2021). Acquiring Ethical AI. *Florida Law Review*, *73*.

Sayer, P. (2016, October 24). Not robocop, but robojudge? A.I. learns to rule in human rights cases. *Computerworld*.

Tashea, J. (2019, June 7). France bans publishing of judicial analytics and prompts criminal penalty. *ABA Journal*.

Tushnet, M. (1980). Post-realist legal scholarship. *Wis. L. Rev.*, 1383.

Young, M. M., Bullock, J. B., & Lecy, J. D. (2019). Artificial discretion as a tool of governance: A framework for understanding the impact of artificial intelligence on public administration. *Perspectives on Public Management and Governance*, *2*(4), 301–313.