78 Admin. L. Rev. (2025) (Forthcoming)

Evaluation as Due Process: Civil Service in an Automated Age

Daniel E. Ho, Olivia Martin, Amy Perez & Kit Rodolfa±

Abstract

Modern government is built on the civil service. The twentieth century consensus, emerging out of the New Deal, was based on a social safety net administered by civil servants to ensure fair administration and due process. Yet today, decades of denying agencies the legal flexibility and technical resources to modernize have created a crisis that fuels political backlash. This crisis is now driving hasty attempts at technological fixes: the "Department of Government Efficiency" seeks to dismantle the civil service through mass layoffs and rapid AI deployment. At the same time, understaffed state agencies are experimenting with AI for benefits administration, with one claiming its system "taught itself all of the [state's] eligibility rules" and eliminated a massive claims backlog in one month. Both approaches can rush toward AI solutions without adequate safeguards, presenting a false choice between dysfunctional status quo and reckless technological disruption, when thoughtful modernization could preserve administrative due process while meeting modern challenges.

This Article charts a path beyond this impasse by examining how administrative law should govern AI's evolving role in benefits adjudication. We trace how merit staffing requirements, designed to ensure procedural fairness and insulate decisions from politics, have also impeded responsible innovation. These regulations produce contradictory guidance that leaves states uncertain about AI's legal boundaries, forcing them to choose between violating timeliness mandates or merit staffing rules.

The pandemic intensified this dilemma, spurring widespread state experimentation—from fraud detection to chatbots to "auto-adjudication." While AI tools can drastically reduce backlogs, increase outreach, and improve efficiency, they also introduce risks of error, opacity, and bias. Through a simple yet novel audit of state systems, we expose vulnerabilities in these largely unexamined experiments. Yet we also show, through the IRS's evaluation of its phone help line that integrated machine assistance, that rigorous assessment of human-AI systems is both feasible and essential.

We argue that the path forward to vindicating constitutional and administrative law values lies in evaluation. While implicit in *Mathews v. Eldridge*, the failure to make this explicit has led to neglect. Evaluation as due process is not merely a bureaucratic exercise but a legal imperative—a means of operationalizing due process at scale and with flexibility that preserves the dignity, fairness, and accountability of administrative systems.

± Daniel E. Ho is the William Benjamin Scott and Luna M. Scott Professor of Law, Professor of Political Science, Professor of Computer Science (by courtesy), Senior Fellow at the Stanford Institute for Economic Policy Research (SIEPR) and the Stanford Institute for Human-Centered AI, and Faculty Director of the Regulation, Evaluation, and Governance Lab (RegLab). Olivia Martin earned her J.D. at Stanford Law School and is a Ph.D. student in Economics at Stanford University and a Graduate Student Fellow at RegLab. Amy Perez led unemployment insurance modernization work at the U.S. Department of Labor after years working at Colorado's unemployment agency. Kit Rodolfa is a Principal Research Scientist at RegLab. We are grateful to leadership in various state unemployment offices for the conversations that helped inspire this article, to Ellen Kwon for excellent research assistance, as well as to seminar participants at Stanford's RegLab for helpful feedback and comments.

Introduction

One of the central roles of government is the provision of the social safety net. As Justice Brennan wrote in *Goldberg v. Kelly*, "[p]ublic assistance . . . is not mere charity, but a means to 'promote the general Welfare, and secure the Blessings of Liberty to ourselves and our Posterity.'"[1] These significant constitutional interests counsel "uninterrupted provision" to those eligible to receive assistance; failure to do so risks violating citizens' due process rights.[2] The twentieth century consensus, emerging out of the New Deal, was that the safety net would be administered by a neutral civil service to ensure fair administration and due process. Yet the future role of safety net programs and the law governing them is at a critical juncture, shaped by three pressing challenges.

First, the infrastructure to administer these programs has become frozen in outdated frameworks due to both chronic resource constraints and rigid legal restrictions. This institutional paralysis—agencies operating with 1970s technology and 1930s bureaucratic constraints while processing 21st century caseloads[3]—has predictably fueled political backlash.[4] The "Department of Government Efficiency" has seized on these failures to justify wholesale elimination, lambasting the civil service as an "antidemocratic" institution staffed by "millions of unelected, unappointed" officials.[5] The solution is as sweeping as it is concerning: mass reductions in head count across federal agencies, with certain agencies "deleted outright," accompanied by untested deployment of AI to fill the gaps.[6] President Trump has personally acted on these critiques, making the civil service a focal point of the first days of his second term.[7] This false choice between dysfunction and demolition obscures a third path: equipping government with modern tools and appropriate oversight to fulfill its constitutional obligations.

Second, these systems are under tremendous duress precisely when citizens need them most, revealing how understaffing and outdated technology create artificial scarcity in public services. For

---

1 397 U.S. 254, 265 (1970).
2 *Id.*
3 *See generally* JENNIFER PAHLKA, RECODING AMERICA (2024); JENNIFER PAHLKA & ANDREW GREENWAY, THE HOW WE NEED NOW: A CAPACITY AGENDA FOR 2025 AND BEYOND, NISKANEN CENTER (DEC. 2024), https://www.niskanencenter.org/wp-content/uploads/2024/12/Niskanen-State-Capacity-Paper_-Jen-Pahlka-and-Andrew-Greenway-2.pdf.
4 *See* Jennifer Pahlka, *This Is How Democrats Can Counter Elon Musk*, N.Y. TIMES (Feb. 7, 2025), https://www.nytimes.com/2025/02/07/opinion/democrats-elon-musk-doge.html.
5 Elon Musk & Vivek Ramaswamy, *The DOGE Plan to Reform Government*, WALL ST. J. (Nov. 20, 2024), https://www.wsj.com/opinion/musk-and-ramaswamy-the-doge-plan-to-reform-government-supreme-court-guidance-end-executive-power-grab-fa51c020. DOGE is not (yet) a department, which would require an act of Congress. U.S. CONST. art. II, § 2, cl. 2.
6 *Id.*; Andrea Hsu, *Officially, 59,000 Federal Jobs Are Gone under Trump. There's More to the Picture*, NPR (June 6, 2025), https://www.npr.org/2025/06/04/nx-s1-5421277/trump-federal-workers-layoffs-doge; Kyle Chayka, *Elon Musk's A.I. Fueled War on Human Agency*, NEW YORKER (Feb. 12, 2025), https://www.newyorker.com/culture/infinite-scroll/elon-musks-ai-fuelled-war-on-human-agency (referring to DOGE's implementation of AI, such as AI filters scanning Treasury grant proposals for "forbidden terms," as "techno-fascism by chatbot"); Coral Davenport, *Inside Trump's Plan to Halt Hundreds of Regulations*, N.Y. TIMES (Apr. 15, 2025), https://www.nytimes.com/2025/04/15/us/politics/trump-doge-regulations.html (describing Elon Musk's development of an AI tool "intended to comb through the 100,000-pluys pages of the Code of Federal Regulations and identify rules that are outdated or legally vulnerable in the wake of the two Supreme Court decisions").
7 *See* Restoring Accountability to Policy-Influencing Positions Within the Federal Workforce, Exec. Order No. 14171 (Jan. 21, 2025); Reforming the Federal Hiring Process and Restoring Merit to Government Service, Exec. Order. No. 14170 (Jan. 21, 2025); Restoring Accountability for Career Senior Executives, Presidential Memorandum (Jan. 20, 2025); Hiring Freeze, Presidential Memorandum (Jan. 21, 2025); Return to In-Person Work, Presidential Memorandum (Jan. 21, 2025)https://www.whitehouse.gov/presidential-actions/2025/01/return-to-in-person-work/.

example, during the pandemic, state unemployment agencies faced claim volumes as much as 33 times higher than normal.[8] One state estimated that it would need at least three times as much staff to handle the increased workload; in actuality, staffing grew only 42% from 2020 to 2021.[9] As a result of these issues, over a third of applicants to unemployment insurance waited over a month for benefits for much of 2020.[10] Further, an estimated 11% of total unemployment insurance benefits paid during the pandemic—or $100 billion—were identified as fraudulent.[11] And these challenges extend beyond the pandemic: in 2023, nearly half of denials of food assistance claims were adjudicated improperly or without adequate notice.[12] Such operational shortcomings don't just harm individual claimants; they fuel the narrative that government cannot deliver abundance and prosperity to its citizens.[13]

Third, many see AI and automation as solutions to these shortcomings. However, the government's track record with modernization efforts has been fraught, displaying either outright incompetence or design entrenched in scarcity thinking, using technology to deny rather than deliver benefits.[14] The automated fraud detection system deployed by Michigan's labor department accused nearly 40,000 applicants of fraud with a staggering error rate of 93%, turning a tool meant to protect public resources into one that wrongfully denied them.[15] Meanwhile, New York City's generative AI chatbot provided inaccurate legal advice, incorrectly informing landlords that they could reject tenants with Section 8 vouchers or that they could not evict tenants for refusing to pay rent.[16] Without rigorous evaluation and human oversight, AI systems can entrench rather than eliminate the barriers that create artificial scarcity in public services.[17]

This Article addresses these central challenges to government by proposing a path forward that rejects both the dysfunctional status quo and the dismantling proposed by DOGE. Properly implemented, AI can expand access to benefits, reduce wait times from months to days, and ensure that eligible citizens receive the support they're entitled to. The choice is not between human judgment and machine efficiency, but rather how to combine both to achieve what neither can accomplish alone: a government that delivers on its constitutional promise of promoting the general welfare through

---

8 U.S. GOV'T ACCOUNTABILITY OFF., GAO-22-104251, UNEMPLOYMENT INSURANCE: PANDEMIC PROGRAMS POSED CHALLENGES, AND DOL COULD BETTER ADDRESS CUSTOMER SERVICE AND EMERGENCY PLANNING 13 (2022), https://www.gao.gov/assets/gao-22-104251.pdf.

9 *Id.* at 14; MINNESOTA MANAGEMENT AND BUDGET, STATE OF MINNESOTA WORKFORCE PLANNING REPORT: FY 2021 8, 9, 31 (2021), https://mn.gov/mmb-stat/workforce-reports/2021.pdf.

10 *Benefits: Timeliness and Quality Reports*, U.S. DEP'T OF LABOR, https://oui.doleta.gov/unemploy/btq.asp.

11 U.S. GOV'T ACCOUNTABILITY OFF., GAO-23-106696, UNEMPLOYMENT INSURANCE: ESTIMATED AMOUNT OF FRAUD DURING PANDEMIC LIKELY BETWEEN $100 BILLION AND $135 BILLION 17 (2023), https://www.gao.gov/products/gao-23-106696.

12 *SNAP Case and Procedural Error Rates*, U.S. FOOD & NUTRITION SERV., https://www.fns.usda.gov/snap/qc/caper. This is unlikely to be due to increased strain during the pandemic, as applications to SNAP only increased by approximately 10% during the pandemic. *Supplemental Nutrition Assistance Program Participation and Costs*, U.S. FOOD & NUTRITION SERV. https://fns-prod.azureedge.us/sites/default/files/resource-files/snap-annualsummary-11.pdf.

13 *See generally* EZRA KLEIN & DEREK THOMPSON, ABUNDANCE (2025).

14 *See generally* JENNIFER PAHLKA, RECODING AMERICA: WHY GOVERNMENT IS FAILING IN THE DIGITAL AGE AND HOW WE CAN DO BETTER (2023).

15 Alejandro de la Garza, S*tates' Automated Systems Are Trapping Citizens in Bureaucratic Nightmares With Their Lives on the Line*, TIME (May 28, 2020), https://time.com/5840609/algorithm-unemployment/; Rachel Kohl, *Automated Stategraft: Faulty Programming and Improper Collections in Michigan's Unemployment Insurance Program*, 2024 WISC. L. REV. (2024).

16 Kyle Orland, *NYC's Government Chatbot Is Lying About City Laws and Regulations*, ARSTECHNICA (Mar. 29, 2024), https://arstechnica.com/ai/2024/03/nycs-government-chatbot-is-lying-about-city-laws-and-regulations/.

17 Bryce Covert, *States Are Turning Their Public Benefits Systems over to AI. The Results Have Often led to 'Immense Suffering*, FAST COMPANY (Jan. 23, 2025), https://www.fastcompany.com/91265363/states-are-turning-their-public-benefits-systems-over-to-ai-the-results-have-often-led-to-immense-suffering.

accessible, accurate, and timely services. Our six key contributions advance debates about the future of government programs, the civil service, administrative law, and the digital state.

First, we provide an in-depth characterization of the operational reality of mass adjudication systems. Critiques of automation often fail to consider this empirical reality: state agencies face an acute staffing shortage and an onslaught of cases that cannot be handled by paper alone. We show this through case studies of two of the country's largest social safety net programs: unemployment insurance (UI), which provides transition payments for workers losing a job through no fault, and the Supplemental Nutrition Assistance Program (SNAP), which provides food benefits for low-income individuals. We explore how labor-intensive adjudication has struggled to deliver timely and accurate outcomes and how modernization efforts have remained piecemeal and constrained by legal and financial limitations.

Second, we trace the origins of state civil service or "merit staffing" requirements to the emergence of these benefits programs. We examine ambiguities and contradictions in federal policies that govern the boundaries between civil service and automation. Conflicting regulations mandate both that benefits determinations be made by merit staff and that technology be outsourced, creating profound uncertainty when automation is introduced into benefits adjudication.

Third, in light of legal uncertainty, we document a staggering diversity of state experimentation with AI-based tools in benefits adjudication. We categorize these efforts into three distinct waves of AI implementation—fraud detection, chatbots, and adjudication assistance, culminating in the use of Generative AI.

Fourth, to show the critical need for evaluation, we design and carry out a simple audit of state-level chatbot implementations, evaluating their capabilities, accuracy, and vulnerabilities in providing critical guidance to claimants. We show how these tools challenge the conventional line between general guidance and claimant-specific advice, with the latter conventionally seen as reserved for the civil service. Rigorous evaluation of the risks and benefits of such systems is sorely lacking.

Fifth, while evaluation, auditing, and assessing AI systems in benefits systems may seem like a daunting challenge, we show that this is eminently feasible. Indeed, we show that functionally, the federal government has already done so for the same kind of service that chatbots purport to offer: citizen assistance through call lines. The Internal Revenue Service's modernization efforts to establish taxpayer help lines, which over time integrated machine assistance, highlight how a large, resource-constrained federal agency has implemented evaluation frameworks to navigate similar challenges, while balancing efficiency and accountability.

Finally, we argue that evaluation is the linchpin for reconciling automation with constitutional and administrative law values, notably due process. The current arsenal of legal frameworks—such as merit staffing and ex-post judicial review for due process—is inadequate to guide states in managing AI's complex challenges. Modern benefits systems cannot rely alone on a select few plaintiffs, like those in *Goldberg* and *Mathews*, to ensure that they are meeting the demands of due process at the scale and speed inherent to AI. While scholars considering automation in the abstract have called for transparency,[18] improved notice and re-designed hearings,[19] and designing technical tools to ensure procedural regularity,[20] evaluation uniquely accounts for the institutional realities of mass adjudication. Moreover, our proposal merges the dominant focus in AI and Machine Learning (ML) on evaluation

---

[18] *See, e.g.,* Danielle Keats Citron, *Open Code Governance*, 2008 U. CHI. L. LEGAL. F. 355, 356 (2008); Deirdre K. Mulligan & Kenneth A. Bamberger, *Saving Governance-By-Design,* 106 CAL. L. REV. 697, 770-783 (2018); Margaret Mitchell et al., *Model Cards for Model Reporting*, Proceedings of the Conference on Fairness, Accountability and Transparency – FAT * '19, at 220 (2019).

[19] *See, e.g.,* Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1303-1313 (2008); Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93 (2014).

[20] Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 662-672 (2017).

standards with the day-to-day realities of public administration, demonstrating how to adapt these standards beyond mere lab benchmarks. Agencies must have clear benchmarks to determine whether a human or an automated system poses a greater "risk of erroneous deprivation," particularly as automation and AI reshape benefits systems. Put differently, evaluation is due process.

The Article proceeds as follows. In Part I, we analyze the operational demands of benefits adjudication. Part II examines the history and law of merit staffing requirements. Part III examines the three waves of state experimentation with automation in benefits systems, offers our survey and audit results, and discusses the counterexample of evaluation of IRS help lines. Part IV assesses existing legal sources of accountability and makes the case for evaluation as due process. Technology must assist and augment the civil service, and evaluation will be critical to designing government technology systems and modernizing government programs to complement and vindicate the values we have long expected of merit staff.

I. The Status Quo: A Strained Benefits Adjudication System

Benefits administration—programs like Unemployment Insurance (UI) and the Supplemental Administration Program (SNAP)—represents one of the most frequent and consequential touchpoints between the public and government. Millions of Americans interact with these systems each year, relying on them during periods of economic hardship or personal crisis. As such, these programs play an outsized role in shaping public trust: when they function smoothly, they reinforce confidence in government's capacity to deliver essential services. But when they fail—whether through delayed payments, administrative backlogs, or wrongful denials—they not only undermine the livelihoods of vulnerable claimants but also fuel broader frustrations with public institutions.

In 2020, roughly one in eight American adults received unemployment benefits.[21] That same year, nearly 40 million people—or nearly one eighth of the overall population, received food assistance through SNAP.[22] Both programs are emblematic examples of cooperative federalism[23]: the federal government provides the majority of administrative funding and overarching guidelines, while state agencies handle administration and adjudication shaped by the specifics of state law. This structure allows for considerable flexibility in how state legislatures determine eligibility for benefits, but often leaves little flexibility in how state agencies administer their programs on an operational level.[24]

A. The Adjudication Process

---

[21] It is estimated that somewhere between 23.6 million (according to Census records) and 45.4 million (according to IRS records) unique individuals received unemployment insurance in 2020. Jeff Larrimore, Jacob Mortenson & David Splinter, *Unemployment Insurance in Survey and Administrative Data*, BD. OF GOVERNORS OF THE FED. RSRV. SYS. (July 5, 2022), https://www.federalreserve.gov/econres/notes/feds-notes/unemployment-insurance-in-survey-and-administrative-data-20220705.html. The adult population was approximately 258 million in 2020. Stella U. Ogunwole et al., *Population Under Age 18 Declined Last Decade*, CENSUS (Aug. 12, 2021), https://www.census.gov/library/stories/2021/08/united-states-adult-population-grew-faster-than-nations-total-population-from-2010-to-2020.html.

[22] *Characteristics of SNAP Households: FY 2020*, U.S. FOOD & NUTRITION SERV., https://www.fns.usda.gov/research/snap/characteristics-households-fy20-early-covid19-characteristics/.

[23] *See* Abbe R. Gluck, *Nationalism as the New Federalism (and Federalism as the New Nationalism): A Complementary Account (and Some Challenges) to the Nationalist School*, 59 ST. LOUIS U. L.J 1045, 1045 (2014); Michael C. Dorf & Charles F. Sabel, *A Constitution of Democratic Experimentalism*, 98 COL. L. REV. 267, 340 (1998).

[24] *See infra* Part II.

Adjudicating benefits eligibility is a labor-intensive process for both agencies and applicants. In the case of unemployment insurance, the process begins with a person submitting a claim to a state agency explaining why and how they separated from their job.[25] An adjudicator then decides their eligibility under state and federal law, generally evaluating the claimant on two criteria: whether they earned enough money during the applicable period and whether they left their job due to no fault of their own (e.g., laid off or needed to care for a sick family member).[26] The employer is also notified of the claim, asked to provide details about the worker's separation, and may dispute the claim by alleging that the worker was terminated for cause or quit without reason. If there are discrepancies between the claims, the adjudicator must conduct additional fact finding with the claimant, the employer, or both. Fact-finding processes can take weeks, depending on the time taken to respond. Workers and employers can appeal the initial determination, resulting in a hearing before an administrative law judge.

If all goes smoothly, it takes two to three weeks from filing a claim to receiving the initial check.[27] Some claims are straightforward—for example, a layoff undisputed by an employer—but many require careful review, such as quitting a job for specific reasons allowed by law, like attending certain types of training.[28] The eligibility determination process is a significant filter: in 2022, only one quarter of unemployed individuals applied for UI, and only half of applicants received benefits due to eligibility issues.[29] Most of the unemployed who did not apply reported that they did not believe they were eligible to receive benefits.[30]

SNAP applications begin similarly, with individuals completing an application either online or in person at a local welfare office. However, federal regulations mandate that all applicants also be interviewed.[31] During the interview, the eligibility worker must verify the applicants' information and review documents confirming their residence and income. To continue receiving benefits, recipients must re-certify their income and eligibility at a frequency determined by the state (ranging from every 6 months to two years or more).[32] The interview, often available only during working hours, is commonly cited as the largest barrier to SNAP participation.[33]

---

[25] *See How Do I File for Unemployment Insurance?*, U.S. DEP'T OF LABOR, https://www.dol.gov/general/topic/unemployment-insurance.
[26] *Id.*
[27] *Id.*
[28] *See, e.g.*, Cal. Code Regs. tit. 22, § 1256-5 (2023).
[29] *Characteristics of Unemployment Insurance Applicants and Benefit Recipients Summary: March 2023,* U.S. BUREAU OF LABOR STAT., https://www.bls.gov/news.release/uisup.nr0.htm.
[30] The UI take-up rate is around 77%, which is roughly in line with other social benefits programs. *See, e.g.,* Stéphane Auray, David L. Fuller, & Damba Lkhagvasuren, *Unemployment Insurance Take-up Rates in an Equilibrium Search Model,* 112 EUR. ECON. REV. 1 (2019); Avraham Ebenstein & Kevin Stange, *Does Inconvenience Explain Low Take-Up? Evidence from Unemployment Insurance*, 29 J. POL. ANALYSIS & MGMT. 111 (2009).
[31] 7 C.F.R § 273.2.
[32] U.S. FOOD AND NUTRITION SERV., STATE OPTIONS REPORT 18 (2017), https://fns-prod.azureedge.us/sites/default/files/snap/14-State-Options.pdf. During the pandemic, however, many states were able to waive the initial eligibility interview requirement for new applicants. *See SNAP – Adjusting Interview Requirements Due to Novel Coronavirus (COVID-19),* U.S. FOOD & NUTRITION SERV., https://www.fns.usda.gov/snap/adjusting-interview-requirements-covid-19. However, the possibility for such waivers ended with the end of the official federal public health emergency. *COVID-19 Public Health Emergency*, DEP'T OF HEALTH & HUMAN SERVS., https://www.hhs.gov/coronavirus/covid-19-public-health-emergency/index.html.
[33] *See* ALLIANCE TO TRANSFORM CALFRESH, ENROLLING MEDI-CAL PARTICIPANTS IN CALFRESH: WHAT WORKS? LESSONS FROM COUNTY-LEVEL EXPERIMENTATION IN CALIFORNIA AND NATIONAL RESEARCH (2019), https://nourishca.org/CalFresh/CFPAPublications/ATC-Enrolling-Medi-Cal-Participants-in-CalFresh-What-Works-2019.pdf.

For both programs, federal agencies oversee state performance through statutorily mandated evaluation systems, though these systems only focus on the select areas emphasized in the law. The Social Security Act instructs the Secretary of Labor to only certify unemployment insurance payments to states if their "methods of administration" are "reasonably calculated to insure full payment of unemployment compensation when due."[34] The Department of Labor (DOL) aims to meet this requirement by regularly evaluating the "Benefits Timeliness and Quality" (BTQ) of state agencies' eligibility determinations.[35] For timeliness, DOL calculates the percentage of first payments made within two or three weeks, with the "acceptable" level of performance being 87%.[36] For quality, DOL randomly audits 60-100 of a state's eligibility determinations each quarter.[37] Each issue is scored out of 100, with points deducted for inadequate fact collection, improper application of law and policy, and insufficient reasoning in the written explanation.[38] A determination fails the audit if it scores below 95, and the "acceptable" level of performance for a state is for 75% of determinations to pass.[39] From 2010 to 2019, states met this bar just 52% of the time; some states, like New Jersey, Florida, and Pennsylvania met this level less than 5% of the time.[40] States are also evaluated based on the extent of their improper payments—benefits both overpaid and underpaid.[41]

SNAP similarly has a quality control system; however, it focuses almost exclusively on payment error rates—both over and underpayments.[42] However, the Food and Nutrition Service (FNS) also measures the case and procedural error rate ("CAPER") at which households are improperly denied benefits due to an inaccurate or procedurally incorrect decision.[43] In 2019, 34% of denials were improperly made.[44] In 2023, this rate stood at 45%. States are not penalized for persistently high CAPERs.[45]

There is a limited range of options for what the federal agency can do if states fail to meet performance benchmarks. In the case of UI, states are merely placed in corrective action plans if their annual performance is deemed unacceptable; in 2022, states submitted a total of 874 such action plans,

---

[34] 42 U.S.C. § 503(a).

[35] EMP. & TRAINING ADMIN., EMPLOYMENT AND TRAINING HANDBOOK NO. 301 (5th ed.), https://oui.doleta.gov/dmstree/handbooks/301/5th/hb301_5.pdf [hereinafter ET Handbook 301].

[36] *UI Performs Core Measures*, DEP'T OF LABOR, https://oui.doleta.gov/unemploy/pdf/Core_Measures.pdf. The applicable standard is 14 days for states with a "waiting week"—a non-compensable period that must be served before benefits commenced. This standard applies to most large states, including California, Texas, New York, and Florida. The 21-day standard applies to states without a waiting week. 20 C.F.R. § 640.

[37] Large states draw a sample of 100 determinations, and small states draw a minimum sample of 60 determinations. The sample must be equally composed of separation (i.e., whether the claimant left the job due to no fault of their own) and non-separation issues (e.g., issues such as failing to actively search for work). ET Handbook 301, *supra* note 35.

[38] EMP. & TRAINING ADMIN., EMPLOYMENT AND TRAINING HANDBOOK NO. 401 (5th ed.), https://www.dol.gov/sites/dolgov/files/ETA/handbooks/2017/ETHand401_5th.pdf.

[39] *UI Performs*, *supra* note 36.

[40] Analysis of data from *Benefits: Timeliness and Quality Reports*, *supra* note 10. For these calculations, meeting the acceptable level of performance means scoring 75% or greater on both the non-separation and separations audits.

[41] *Id.*

[42] *SNAP Quality Control*, FOOD & NUTRITION SERV., fns.usda.gov/snap/qc. This is consistent with the language of the Food and Nutrition Act, which describes the quality control mandate as follows: "In carrying out the supplemental nutrition assistance program, the Secretary shall carry out a system that enhances payment accuracy and improves administration by establishing fiscal incentives that require State agencies with high payment error rates to share in the cost of payment error." 7 U.S.C. § 2025(c).

[43] *SNAP Case and Procedural Error Rates,* FOOD & NUTRITION SERV., https://www.fns.usda.gov/snap/qc/caper.

[44] *Id.*

[45] Dottie Rosenbaum & Katie Bergh, *SNAP Includes Extensive Payment Accuracy System,* CTR. ON BUDGET & POL'Y PRIORITIES (June 21, 2024), https://www.cbpp.org/research/food-assistance/snap-includes-extensive-payment-accuracy-system.

one for each deficiency, such as failing to meet timeliness, quality, or appeals performance standards.[46] The only more severe corrective response available is the complete denial of administrative funding, the so-called "nuclear option."[47]

B. The Demands of the Pandemic

Applications to both UI and SNAP rose during the pandemic, as is expected for income stabilizing benefits programs. For UI, this strain was exacerbated by the introduction of three UI expansion programs through the 2020 Coronavirus Aid, Relief, and Economic Security (CARES) Act.[48] Not even including the new CARES Act programs, regular initial UI claims submitted early in the pandemic ranged from 11 to 33 times higher than volume in the preceding three months.[49] States were particularly unprepared because administrative funding from the federal government is tied to claims-related caseloads, which were at historic lows immediately prior to the pandemic.[50]

As a result, states struggled to adequately increase staff to accommodate this surge in claims; some agencies borrowed staff from other state departments or hired contractors to take on certain roles in call centers.[51] While some agencies were able to increase their capacity in these ways—for example, Florida increased staff by about 93%—the staffing increases paled in comparison to the increase in claim volume. Minnesota officials reported that based on the model that DOL uses to allocate funding to states, the Minnesota UI agency would have needed about 4,000 total staff—nearly three times the

---

46 U.S. Dep't of Labor, Emp. & Training Admin., Unemployment Insurance Program Letter No. 17-22 (Aug. 12, 2022), https://www.dol.gov/sites/dolgov/files/ETA/advisories/UIPL/2022/UIPL%2017-22/UIPL%2017-22%20-%20Blue%20.pdf; Agency Information Collection Activities; Comment Request; Unemployment Insurance (UI) State Quality Service Plan (SQSP), 88 Fed. Reg. 36505 (June 5, 2023), https://www.federalregister.gov/documents/2023/06/05/2023-11835/agency-information-collection-activities-comment-request-unemployment-insurance-ui-state-quality. The corrective action plans may be quite brief; for example, many of Washington's are just a few sentences each. *Unemployment Insurance Program*, WASHINGTON WORKFORCE INNOVATION AND OPPORTUNITY ACT STATE PLAN, htttps://wioaplans.ed.gov/node/81486.

47 Social Security Act § 303, Pub. L. No. 74-271, 49 Stat. 620 (1935). Such extreme sanctions appear to never have been used, though some Senate Republicans urged their use against California during the pandemic in response to high levels of fraud in the state. Press Release, U.S. Senate Comm. on Health, Educ., Lab. & Pensions, Ranking Members Cassidy & Crapo Demand Transparency on DOL's New Policy Forgiving Julie Su for Losing $32 Billion to Unemployment Insurance Fraud (Sept. 12, 2023), https://www.help.senate.gov/ranking/newsroom/press/ranking-members-cassidy-crapo-demand-transparency-on-dols-new-policy-forgiving-julie-su-for-losing-32-billion-to-unemployment-insurance-fraud. The inadequacy of this "nuclear option" is also lamented in Project 2025's chapter on DOL, which calls on Congress to "provide DOL with more reasonable enforcement tools for the UI system." Joanthan Berry, *Department of Labor and Related Agencies, in* PROJECT 2025 598 (Paul Dans & Steven Groves ed., 2024).

48 OFF. OF INSPECTOR GEN., U.S. DEP'T OF LABOR, COVID-19: STATES STRUGGLED TO IMPLEMENT CARES ACT UNEMPLOYMENT INSURANCE PROGRAMS (2021), https://www.oig.dol.gov/public/reports/oa/2021/19-21-004-03-315.pdf. These extensions, however, were not COVID-19 specific. The UI program was also expanded during the Great Recession. *See* CONG. BUDGET OFF., UNEMPLOYMENT INSURANCE IN THE WAKE OF THE RECENT RECESSION (2012).

49 2022 GAO REPORT, *supra* note 8 at 12.

50 *Id.* at 7. Funding available for state administration declined by about 32% in real terms from 2010 to 2019. *Id.* at 7-8.

51 *Id.* at 13. YOLANDA RICHARDSON & JENNIFER PAHLKA, EMPLOYMENT DEVELOPMENT DEPARTMENT STRIKE TEAM DETAILED ASSESSMENT AND RECOMMENDATIONS (2020), https://www.govops.ca.gov/wp-content/uploads/sites/11/2020/09/Assessment.pdf; Letter from Julie A. Su, Sec'y, Cal. Lab. & Workforce Dev. Agency, to Assemb. David Chiu (Aug. 19, 2020), https://abgt.assembly.ca.gov/sites/abgt.assembly.ca.gov/files/Sec.%20Su%20Response%20to%20Asm.%20Chiu%20Letter%208.19.20%20and%20Referenced%20Letter.pdf (California used over 500 "surge staff" from Deloitte to staff call centers during the pandemic).

level of their 1,500 employees[52]-- a level they characterized as "impractical."[53] Staffing up quickly was further complicated by the fact that claim adjudication demands attaining a high level of statutory understanding and institutional knowledge. Official training periods are typically 6 to 8 weeks,[54] but many workers take far longer to feel fully "up to speed." In Former U.S. Deputy Chief Technology Officer Jennifer Pahlka's study of California's UI agency, an employee referred to himself as "the new guy" after 17 years on the job—his tenure dwarfed by colleagues with over 25 years of experience.[55] In such settings, adding staff can initially lower productivity as senior staff are diverted to training new hires; for example, after hiring additional staff early in the pandemic, California's agency required 2-5 times longer to complete certain tasks.[56]

The results of this staggering caseload coupled with inadequate staffing were predictable: increased wait times for applicants, a significant drop in decision quality, and heightened fraud rates. Put differently, the UI system failed precisely at a time when American workers needed it most. Immediately prior to the pandemic, roughly 90% of first UI payments were made within three weeks, but as depicted in the right panel of Figure 1, this rate plummeted to nearly 50% with COVID-19, and did not rise above 80% until 2023.[57] At the beginning of the pandemic, around 10% of initial applicants—or, around 250,000 people—waited over 70 days to receive their initial benefit, with some waiting as long as four months or more.[58] Quality measures also dropped sharply (left panel of Figure 1), from a pass rate of over 80% at baseline to below 60% in the first period after the pandemic. As staff were overwhelmed learning to operate new systems, many fraudulent claims were approved. In 2023, GAO estimated that states made somewhere between $100 and $135 billion—about 11 to 15% of total benefits paid—in fraudulent payments during the pandemic, sparking public and congressional outrage.[59]

---

52 MINNESOTA MANAGEMENT AND BUDGET, STATE OF MINNESOTA WORKFORCE PLANNING REPORT: FY 2021 8, 9, 31 (2021), https://mn.gov/mmb-stat/workforce-reports/2021.pdf.

53 2022 GAO REPORT, *supra* note 8 at 14.

54 *See, e.g.*, *id.* at 14; *Unemployment Benefit Specialist (Adjudicator) FAQs,* WISCONSIN DEP'T OF WORKFORCE DEV., https://dwd.wisconsin.gov/jobs/faqs/240108-ubs-adjudicator.htm.

55 PAHLKA, *supra* note 14 at 38.

56 RICHARDSON & PAHLKA, *supra* note 51 at 15.

57 *See Benefits: Timeliness and Quality Reports, supra* note 10; U.S. Dep't of Labor, Emp. & Training Admin., Unemployment Insurance Program Letter No. 11-26 (June 14, 1971), https://oui.doleta.gov/dmstree/uipl/uipl_pre75/uipl_1126.htm.

58 Greg Iacurci, *He Was Stuck in Unemployment 'Limbo' for Four Months. Then Came $23,000 in Benefits,* CNBC (Aug. 15, 2020), https://www.cnbc.com/2020/08/15/coronavirus-mans-23000-in-unemployment-pay-delayed-four-months.html (four months in New York); Kate Davidson, *Months Later, Thousands Still Waiting for Unemployment Benefits in Oregon*, OPB (Aug. 5, 2020), https://www.opb.org/article/2020/08/05/months-later-still-waiting-for-unemployment-benefits-in-oregon/ (five months in Oregon).

59 U.S. GOV'T ACCOUNTABILITY OFF., GAO-23-106696, UNEMPLOYMENT INSURANCE: ESTIMATED AMOUNT OF FRAUD DURING PANDEMIC LIKELY BETWEEN $100 BILLION AND $135 BILLION (2022), https://www.gao.gov/assets/gao-23-106696.pdf; Richard Lardner, Jennifer McDermott & Aaron Kessler, *How Billions in COVID-19 Pandemic Relief Aid Was Stolen or Wasted,* PBS (June 14, 2023), https://www.pbs.org/newshour/politics/how-billions-in-covid-19-pandemic-relief-aid-was-stolen-or-wasted; Press Release, House Comm. on Oversight & Accountability, Oversight Committee Releases Report on Rampant Waste, Fraud, and Abuse in Pandemic Unemployment Relief Programs (Sep. 10, 2024) https://oversight.house.gov/release/oversight-committee-releases-report-on-rampant-waste-fraud-and-abuse-in-pandemic-unemployment-relief-programs/.

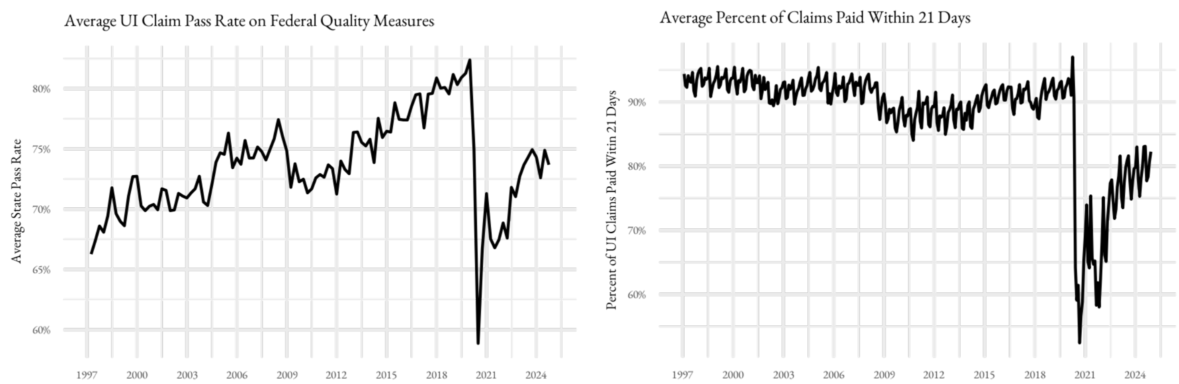

Figure 1: Federal measures of state UI program performance. The left chart shows the average percent of claims that pass the quality audit (making the correct decision and providing adequate explanation) across states and types of claims. To be considered "acceptable," 75% of claims must pass the audit. The right chart shows the average percent of claims paid within 21 days, which is the more lenient of two standards (14 and 21 days).

The strain on SNAP did not increase as dramatically during the pandemic as it did for UI, partly due to its more stringent eligibility requirements:[60] the number of SNAP recipients only increased by about 10% from 2019 to 2020.[61] Yet states still struggled, then and now, to deliver benefits within the 30 days mandated by regulation.[62] In 2023, twenty states were meeting this 30-day requirement less than 80% of the time, with D.C. and Alaska only meeting the requirement 48% and 39% of the time, respectively.[63] As with UI, state officials often pointed to staffing shortages as a cause for backlogs; for example, Alaska's state officials reported having 142 adjudicators processing SNAP applications, but that it needed nearly 200 to clear its backlog.[64]

State failures in administering SNAP benefits have not gone unnoticed. In February 2024, the USDA issued warning letters to 32 states for delays in food stamp distribution.[65] Civil society groups, including one in Alaska, have sued state agencies for failing to comply with federal SNAP requirements.[66] One such suit against the Missouri Department of Social Services was successful, with a federal judge ordering the agency to propose a compliance plan and to file monthly updates with the court.[67]

---

[60] Generally, net income must be around the poverty line to qualify for SNAP. *A Quick Guide to SNAP Eligibility and Benefits*, CTR. ON BUDGET & POL'Y PRIORITIES, https://www.cbpp.org/research/food-assistance/a-quick-guide-to-snap-eligibility-and-benefits.

[61] *Supplemental Nutrition Assistance Program Participation*, U.S. FOOD & NUTRITION SERV., https://fns-prod.azureedge.us/sites/default/files/resource-files/snap-annualsummary-11.pdf.

[62] 7 C.F.R. § 274.2.

[63] *FY 2023 Reported SNAP Application Processing Timeliness*, U.S. FOOD & NUTRITION SERV., https://www.fns.usda.gov/snap/qc/timeliness/fy23.

[64] Alex Brown, *As Millions Wait on Food Stamp Approvals, Feds Tell States to Speed It Up,* S.C. DAILY GAZETTE (Feb. 26, 2024), https://scdailygazette.com/2024/02/26/as-millions-wait-on-food-stamp-approvals-feds-tell-states-to-speed-it-up/.

[65] *Id.*

[66] Press Release, Nat'l Center for Law and Econ. Justice, Groups Sue Over Alaska's Failure to Provide SNAP Benefits to Vulnerable Communities (Jan. 20, 2023), https://nclej.org/news/groups-sue-over-alaskas-failure-to-provide-snap-benefits-to-vulnerable-communities; Press Release, Nat'l Center for Law and Econ. Justice, Federal Court Rules Missouri Violated SNAP Law and the ADA (May 10, 2024), https://nclej.org/snap-highlights/federal-court-rules-missouri-violated-snap-law-and-the-ada.

[67] Holmes v. Knodell, No. 2:22-CV-04026, at 1 (W.D. Mo. May 9, 2024), https://nclej.org/wp-content/uploads/2024/05/161-Order-Granting-Summary-Judgment.pdf.

C. Modernization

Improving operational efficiency through modernization could enhance the resilience of benefits agencies. State IT systems are notoriously outdated, with many depending on outdated hardware and programming languages such as COBOL (a language invented in 1959) for decades after their development in the 1970s and 1980s.[68] As recently as 2021, several states still used paper-based procedures for claims and notifications,[69] and only half had migrated UI applications to the cloud.[70] Many attributed state agencies' pandemic-era failures to dysfunctional technology systems combined with staffing shortages.[71]

However, modernization is easier said than done. Much of the challenge lies in getting technology procurement right—something government has long struggled to achieve. Steven Kelman attributes many failures of performance-based technology initiatives to poor contract design, where procurement officials lack the expertise to evaluate services and are constrained by competition rules that discourage using critical information[72] Jennifer Pahlka highlights similar issues in recent technology challenges, such as launching HealthCare.gov or connecting data systems and processes to follow through on voters' call to expunge criminal records.[73] Pahlka emphasizes the mismatch between government's outdated, top-down approach to technology and the iterative, agile practices of the private sector. Institutional risk aversion and policy complexity—like the 3,600 unique requirements needed to assess UI eligibility—make it difficult, if not impossible, to build scalable systems without years of development and budgets in the hundreds of millions.

UI and SNAP suffer from the same issues of policy complexity and a history of contracting out technology systems that has led to continued use of fragile and outdated systems. Modernization is further hindered by inconsistent funding. While administrative funding is formula-based on claim volume,[74] IT modernization funding is sporadic, complicating long-term planning. For example, in 2017, DOL awarded short-term IT modernization grants totaling $50 million, though a single state's overhaul was estimated to cost nearly this amount.[75] DOL took larger strides in 2021 when the American Rescue Plan Act (ARPA) allocated $2 billion to improve "fraud prevention, equitable access, and timely payment to eligible workers," though half was later rescinded by the Fiscal Responsibility

---

68 *See* U.S. GOV'T ACCOUNTABILITY OFF., GAO-12-957, INFORMATION TECHNOLOGY: DEPARTMENT OF LABOR COULD FURTHER FACILITATE MODERNIZATION OF STATES' UNEMPLOYMENT INSURANCE SYSTEMS 8 (2012), https://www.gao.gov/assets/gao-12-957.pdf at 8; PAHLKA, *supra* note 14.

69 GAO 2022 report at 58, 60, 66, 67-68 (discussing such processes in Delaware, Maine, Pennsylvania, and Tennessee)

70 *Id.* at 25.

71 2022 GAO REPORT, *supra* note 8; Josephine Nesbit, *Food Stamps: This Is the Average Wait Time to Have Your SNAP Application Approved — But Thousands Are Left Waiting Much Longer*, NASDAQ (Dec. 7, 2023), https://www.nasdaq.com/articles/food-stamps:-this-is-the-average-wait-time-to-have-your-snap-application-approved-but; Elena Botella, *Why New Jersey's Unemployment Insurance System Uses a 60-Year-Old Programming Language*, SLATE (Apr. 9, 2020), https://slate.com/technology/2020/04/new-jersey-unemployment-cobol-coronavirus.html.

72 *See* Steven Kelman, *Procurement and Public Management: The Fear of Discretion and the Quality of Government Performance* (1990). For example, Kelman provides an example in which the IRS contracted for the replacement of computers with a vendor whose equipment would meet the IRS's published test specifications, but which the vendor knew were insufficient to meet the agency's long term needs. Kelman's preferred solution is increased discretion for procurement officers to be more aligned with the discretion that private sector procurement officers have.

73 PAHLKA, *supra* note 14.

74 *See* 2022 GAO REPORT, *supra* note 8 at 8.

75 *See* U.S. Dep't of Labor, Emp. & Training Admin., Unemployment Insurance Program Letter No. 22-17 (Sept. 8, 2017), https://oui.doleta.gov/dmstree/uipl/uipl2k17/uipl_2217.pdf.

Act of 2023.[76] With these funds, DOL created a temporary Office of Unemployment Insurance Modernization[77] and deployed 'Tiger Teams"-- multi-disciplinary technical assistance experts—to 36 states.[78] By September 2023, $783 million in grants had been issued, focusing on fraud detection ($227 million), IT modernization ($204 million), and equity, defined as eliminating administrative barriers to applications, reducing backlogs, and improving timeliness ($219 million).[79] The Department also identified investments in Robotic Process Automation (RPA) of "certain non-discretionary tasks," such as mailing out appeals decisions or requesting information from employers, as an effective means of reducing claims backlogs.[80]

SNAP similarly received about $1 billion in additional funding from ARPA to support modernization efforts.[81] State SNAP agencies largely focused on projects that improved the customer experience (e.g., by application re-design or self-service kiosks in community centers), improving application recertification processes, and initiatives to improve hiring and training of additional staff.[82] A few projects also implemented RPA, such as a project in Ohio that worked with Deloitte to design a bot to "automate processes such as discontinuing benefits."[83]

Yet just like previous grants made during economic downturns, the ARPA funds are one-time and thus best suited for short-term projects, not supporting continuous and iterative development and evaluation. That said, stable funding is not the only barrier to the adoption of more advanced approaches to administration. As the next section will show, the very way the federal government currently conceives of staffing is in tension with any significant adoption of AI into benefits administration.

II. Governing by Merit Staffing

A. The Emergence of Merit Staffing Requirements

In the depths of the Great Depression, Congress enacted the Social Security Act of 1935.[84] While this landmark piece of legislation established federal-state partnerships to provide economic security for the elderly, unemployed, and other vulnerable populations, it initially imposed few requirements on state administration. Programs such as unemployment insurance were largely left to state control, and the Act specifically exempted the "selection, tenure of office, and compensation of personnel" from federal oversight.[85]

---

[76] American Rescue Plan Act of 2021 § 9032, Pub. L. 117-2, 135 Stat. 4 (2021); U.S. DEP'T OF LABOR, INSIGHTS AND SUCCESSES: AMERICAN RESCUE PLAN INVESTMENTS IN UNEMPLOYMENT INSURANCE MODERNIZATION, https://www.dol.gov/sites/dolgov/files/ETA/ui-modernization/ARPA%20Investments%20in%20Unemployment%20Insurance%20Modernization.pdf [hereinafter DOL ARPA Report].
[77] 2022 GAO REPORT, *supra* note 8 at 38 n. 80
[78] DOL ARPA REPORT, *supra* note 76 at 14; *Tiger Team Updates*, U.S. DEP'T OF LABOR (June 2023), https://www.dol.gov/agencies/eta/ui-modernization/tiger-teams.
[79] DOL ARPA REPORT, *supra* note 76 at 23.
[80] *Id.* at 53-54.
[81] *Exploring States' SNAP Modernization Projects,* URBAN INST., https://www.urban.org/projects/exploring-states-snap-modernization-projects.
[82] *Id.*
[83] *Id.*
[84] Social Security Act, 49 Stat. 620 (1935).
[85] *Id.* § 303, 49 Stat. at 626.

Reports of weeks-long delays in benefit payments soon surfaced, raising concerns about the appointment of state personnel through political patronage rather than merit.[86] In 1939, the President called attention to the fact that "in some states incompetent and politically dominated personnel has been distinctly harmful," and called on Congress to require that states "establish and maintain a merit system for the selection of personnel."[87] Later that year, the Act was revised to require states to establish and maintain "personnel standards on a merit basis."[88]

This was no small request. Few states had anything that looked like a general civil service system for personnel.[89] The Executive Director of the Social Security Board at the time recounted having to "point out again and again" to states the need for developing strong merit systems and the importance of "selecting people with a high sense of public duty and with courage and integrity."[90] To support states in meeting these new standards, the Board offered states extensive assistance in developing their civil service programs.[91] The Director concluded his 1945 retrospective on the early years of imposing merit requirements with the following:[92]

> There has been too great waste in this country of time, money and effectiveness in the public service through inefficient selection and political turn-over. Despite its inadequacies and the difficulties in its application, the merit principle has made one of the most valuable and lasting contributions to our political economy. Despite its shortcomings, I have heard no suggestion of a substitute that would seem to serve as well.

As both President Roosevelt's remark and the Social Security Board's concerns suggest, two key factors drove the federal government to require states to implement merit staffing systems. First, the sheer complexity of administering such large and impactful programs demanded quality staff, and merit-based hiring was seen as the best way to obtain top talent. Second, merit systems were designed to minimize political turnover in state agencies following elections, ensuring continuity of services. To achieve these ends, the federal government's merit staffing mandate played a transformative role in spurring the development of civil service systems across the states.[93]

---

[86] 75 CONG. REC. 3724 (1937) (noting that "incompetent clerical help is being appointed to responsible positions" in the states and that such appointments were "hampering effective administration").
[87] Franklin D. Roosevelt, Transmittal to Congress of a Report of the Social Security Board (Jan. 16, 1939), https://www.presidency.ucsb.edu/documents/transmittal-congress-report-the-social-security-board. President Roosevelt also noted that the requirement of state merit programs would reduce federal encroachment on the states by "promot[ing] efficiency and eliminat[ing] the necessity for minute Federal scrutiny of state operations." *See also* Lawrence D. Greene, *Federal Merit Requirements: A Retrospective Look*, 11 PUB. PERSONNEL. MGMT. (1982).
[88] 42 U.S.C. §503(a)(1); 53 Stat. 1378 (1939).
[89] Greene, *supra* note 87. For example, as of 1936, only four states—California, Massachusetts, New York, and Wisconsin—required unemployment compensation appointments to be made under civil service regulations. Robert N. Cook, *The Bodies Administering Unemployment Compensation Laws*, 3 LAW & CONTEMPORARY PROBLEMS 95 (1936) (citing Cal. §88; Mass. §9 K; N. Y. §518 (I); Wis. STAT. (1933) §14). The timing presented other difficulties as well. Also in 1939, President Roosevelt released an Executive Order requiring Personnel Departments in all federal agencies, which swept up may personnel professionals into the federal government, living "[o]nly a limited crop" to the States. Oscar M. Powell, *Merit Systems in the Social Security Program*, 8 SOC. SEC. BULL. (1945). Moreover, the United States' entry into World War II just two years later in 1941 further strained states' burgeoning civil service systems and caused the Social Security Board to relax a number of minimum requirements for states as they grappled with the challenges of finding an adequate workforce during wartime. *Id.*
[90] Powell, *supra* note 89.
[91] *Id.*
[92] *Id.*
[93] Greene, *supra* note 87.

B. The Inherently Governmental Boundary to Merit Staffing

Such sweeping requirements inevitably led to questions and pushback from the states, particularly regarding the definition of a "merit-based system."[94] The Office of Personnel Management (OPM) was eventually assigned interpretive authority for the SSA's merit requirements[95] and promulgated standards requiring (1) a selection and advancement system based on ability; (2) merit-based and equitable training, compensation and retention policies; and (3) insulation of employees from "coercion for partisan political purposes."[96] Despite these standards, states continued to debate specific requirements, such as whether merit-based hiring necessitated examination-based selection systems.

Beyond definitional concerns, states also grappled with the scope of merit staffing requirements. Which roles, exactly, required merit-based hiring? Did positions such as paper suppliers or cafeteria staff fall under these mandates? At the core of these discussions was a foundational civil service question: what qualifies as "inherently governmental" and thus necessitates merit system protections?

While states worked to develop their civil service programs, the federal government was confronting similar issues amid a post-World War II contracting boom.[97] The Eisenhower Administration made it explicit policy "that the Federal Government will not start or carry on any commercial activity to provide a service or a product for its own use if such a product or service can be procured from private enterprise through ordinary business channels."[98] While some raised concerns about this new reliance on contracting,[99] the executive branch continued to build policy cementing the practice of contracting. These efforts culminated in the 1966 release of Circular A-76, which affirmed the Eisenhower-era policy of contracting out commercial activity.[100]

The Office of Management and Budget (OMB) has revised A-76 a number of times over the years, with the most recent substantial revision in 2003, but many main components remain the same.[101] A-76 requires agencies to classify all activities they perform as either "commercial" or "inherently governmental."[102] Inherently governmental activities must be undertaken by government personnel,

---

[94] *Id.*
[95] Intergovernmental Personnel Act of 1970, Pub. L. No. 91-648, § 1, 84 Stat. 1909 (1971).
[96] 5 C.F.R. § 900.603.
[97] Mohab Tarek Khattab, *Revised Circular A-76: Embracing Flawed Methodologies*, 34 PUB. CONTRACT L. J. 469 (2005).
[98] Executive Office of the President, Bureau of the Budget, Budget Bulletin 55-4 (Jan. 15, 1955), https://www.governmentcompetition.org/wp-content/uploads/2018/11/Bureauof_the_Budget_Bulletin_55-4_January_15_1955.pdf.
[99] For example, in 1962, cabinet leaders prepared a report to President Kennedy, expressing concern that the administration's reliance on contracting has "blurred the traditional dividing lines between the public and private sector," risks draining off administrators to private industry, and endangers the "axiomatic" understanding that policy decisions "must be made by full-time Government officials clearly responsible to the President and to the Congress." DAVID E. BELL, REPORT TO THE PRESIDENT ON GOVERNMENT CONTRACTING FOR RESEARCH AND DEVELOPMENT, S. Doc. No. 94, 87th CONG., 2d Sess. (1962). *See also* Dan Guttman, *Governance by Contract: Constitutional Visions; Time for Reflection and Choice*, 33 PUB. CONTRACT L. J. 321, 323 (2004).
[100] *See* Khattab, *supra* note 97.
[101] JOHN R. LUCKEY, CONG. RSCH. SVC., OMB CIRCULAR A-76: EXPLANATION AND DISCUSSION OF THE RECENTLY REVISED FEDERAL OUTSOURCING POLICY, RS21489 (2003), https://www.everycrsreport.com/files/20030910_RS21489_c9ac20e42626f78c8c3a3e978c7666fc123f842b.pdf.
[102] OFF. OF MGMT. & BUDGET, EXEC. OFF. OF THE PRESIDENT, CIRCULAR NO. A-76 (REVISED) (2003), https://www.whitehouse.gov/wp-content/uploads/legacy_drupal_files/omb/circulars/A76/a76_incl_tech_correction.pdf.

roughly tracking Constitutional nondelegation principles,[103] and commercial activities should largely be outsourced through a competitive bidding process.[104]

An inherently governmental activity is one "so intimately related to the public interest as to mandate performance by government personnel."[105] Such activities typically require "the exercise of substantial discretion," including (1) binding the government by contract; (2) military, diplomatic, judicial and contract management decisions; (3) decisions "[s]ignificantly affecting the life, liberty or property of private persons" or (4) exerting ultimate control over government property, including funds.[106] But OMB notes that "not every exercise of discretion is evidence that an activity is inherently governmental."[107]

Over the years, OMB has attempted to provide additional guidance by providing examples of inherently governmental tasks, such as criminal investigations and prosecutions, the "determination of agency policy," the "selection" and "control of Federal employees," and the "selection of grant and cooperative agreement recipients."[108] When an agency considers a function outside of these examples, OMB counsels them to make a "case-by-case" assessment of whether the function: (1) "involve[s] the exercise of sovereign powers of the United States" and "thus is governmental by [its] very nature"; and (2) includes discretion to "commit[] the government to a course of action" absent oversight by agency officials.[109]

This test, however, remains ambiguous. Agencies have long struggled to distinguish inherently governmental functions from their routine tasks.[110] As one commentator put it, defining the term is "like trying to nail Jell-O to the wall; only nailing Jell-O is easier."[111]

But it is this wobbly, Jell-O-like line that often defines the scope of agencies' merit staffing requirements under federal grant programs. For example, the Department of Labor explicitly directs states to consult Circular A-76 in determining "what functions a State may outsource under a program where a Federal merit-staffing requirement applies."[112] Following these guidelines can particularly challenging when thinking about modern-day technology services that would be unrecognizable at the time of A-76's creation.

---

103 A.L.A. Schechter Poultry Corp. v. United States, 295 U.S. 495 (1935); Carter v. Carter Coal Co., 298 U.S. 238 (1936). For the relationship between the private nondelegation doctrine and the classification of what is "inherently governmental," see KATE M. MANUEL, CONG. RSCH. SERV., DEFINITIONS OF 'INHERENTLY GOVERNMENTAL FUNCTION' IN FEDERAL PROCUREMENT LAW AND GUIDANCE 20 (2014) (noting that "a judicial declaration that a function is inherently governmental under a constitutional test would not necessarily preclude the executive branch from contracting out this function"), https://sgp.fas.org/crs/misc/R42325.pdf.
104 OMB CIRCULAR A-76, *supra* note 102.
105 *Id.*
106 *Id.*
107 *Id.*
108 Policy Letter 11–01, Performance of Inherently Governmental and Critical Functions, 76 Fed. Reg. 56,227, 56,236 (Sept. 12, 2011).
109 *Id.* at 56,237.
110 *See* BRIDGET C.E. DOOLING & RACHEL AUGUSTINE POTTER, CONTRACTORS IN RULEMAKING 44 (Rep. for the Admin. Conf. of the U.S. 2022), https://www.acus.gov/sites/default/files/documents/Contractors%20in%20Rulemaking%20Final%20Report.pdf (describing agency officials as having a "widespread but incomplete awareness of the existence of an inherently governmental function line with respect to rulemaking").
111 David Isenberg, *To Be, or Not to Be, Inherent: That Is the Question*, HUFF POST: THE BLOG (June 15, 2010), https://www.huffpost.com/entry/to-be-or-not-to-be-inhere_b_539933.
112 U.S. Dep't of Labor, Emp. & Training Admin., Unemployment Insurance Program Letter No. 12-01 (Dec. 28, 2000), https://www.dol.gov/sites/dolgov/files/ETA/advisories/UIPL/2001/UIPL12-01.html. *See also* 78 C.F.R. § 42188 (a final rule from the Centers for Medicare & Medicaid Services asserting that "determining Medicaid eligibility is an inherently governmental function that must be performed by governmental agencies").

C. Interpretative Uncertainty in Defining Automation's Role in Merit Staffing

Modern technology services complicate the inherently governmental inquiry, but in an era of simpler technology, the government's stance was clearer. Before the release of the first Circular A-76, Congress enacted the Automatic Data Processing Act—more familiarly known as the Brooks Act—tasking the General Services Administration with ensuring the "economic and efficient purchase, lease, and maintenance of automatic data processing equipment by Federal agencies."[113] The 1983 iteration of Circular A-76 reinforced this policy by explicitly classifying automatic data processing as a commercial activity rather than an inherently governmental function.[114] Under the capacious heading of automatic data processing fell activities ranging from the clearly outdated tasks of "batch processing" and "key punching" to more timeless practices such as "programming," "design," and "systems engineering."[115]

Jennifer Pahlka argues that the combined effect of Circular A-76 and the Brooks Act marked a pivotal shift in government technology policy, effectively mandating the outsourcing of what has since become the technical and digital backbone of modern organizations.[116] She contends that this outsourcing imperative is responsible for the federal government's persistent technology failures in the 21st century, from the Healthcare.gov debacle to unemployment insurance programs dependent on decades old programming languages.[117]

Pahlka may well be right. But the technology at the center of current debates is enormously different than the room-size computers that were the setting of the Brooks Act, or even the heroic modernization initiatives that Pahlka describes. As AI systems grow more sophisticated, they no longer merely enhance government operations but increasingly shape the fundamental nature of governance itself.[118] Rather than being a way to improve how a government website operates or how information is routed through agencies, AI-enabled automation probes the foundation of how government decisionmaking works and what the civil service means.

Despite the transformative potential of AI, federal agencies overseeing benefits administration have been inconsistent in delineating what forms of automation are permissible under merit staffing requirements. In an Unemployment Insurance Program Letter issued in January 2021, DOL identified the following activities as inherently governmental and thus necessitating merit staffing: (1) "[a]dvising a claimant regarding . . . eligibility for benefits"; (2) "[a]nalysis of facts so as to actually make a determination of benefit eligibility or tax liability"; (3) "[a]ctually making a determination of benefit eligibility or employer tax liability;" and (4) "[d]irect supervision of individuals carrying out the activities described in numbers 1 – 3."[119]

Under this guidance, most forms of automation—and even AI-assisted decision-making—appear to be precluded. Fully automated review would obviously be in violation of (3), AI-powered chatbots

---

113 Pub. L. 89-306, 79 Stat. 1127 (1965). .

114 OFF. OF MGMT. & BUDGET, EXEC. OFF. OF THE PRESIDENT, CIRCULAR NO. A-76 (REVISED) (1983), https://www.whitehouse.gov/wp-content/uploads/legacy_drupal_files/omb/circulars/A76/a076.pdf.

115 *Id.*

116 PAHLKA, *supra* note 14at 105 (2023) ("Since [Circular A-76] . . . government has thought of digital technology much like the pens and paper clips that GSA buys for government offices: something government would be crazy to produce for itself."); *see also id.* at 118 (describing the Brooks Act and Circular A-76 as having "consigned digital to the purview of contracting").

117 *Id.* at 28, 125.

118 *See* DAVID FREEMAN ENGSTROM, DANIEL E. HO, CATHERINE M. SHARKEY & MARIANO-FLORENTINO CUELLAR, GOVERNMENT BY ALGORITHM: ARTIFICIAL INTELLIGENCE IN FEDERAL ADMINISTRATIVE AGENCIES (2020).

119 U.S. Dep't of Labor, Emp. & Training Admin., Unemployment Insurance Program Letter No. 12-01 Change 2 (Jan. 8, 2021), https://www.dol.gov/sites/dolgov/files/ETA/advisories/UIPL/2021/UIPL%2012-01_Change_2.pdf.

providing eligibility advice would be in violation of (1), and AI tools designed to assist adjudicators by analyzing cases could be in violation of (2). Likely in response to severe criticism of states' failed attempts with fraud detection automation,[120] the guidance also expressly clarifies that "[d]eterminations of overpayments or fraud may not be made using automated systems; they must be made by merit-staffed employees."[121] The guidance does, however, allow for staffing "flexibility" in activities such as "[n]otating answers to fact-finding questionnaires," "[d]ata entry where no discretion is required," "[r]outine data processing of instances of failure to report," and "[c]omputer programming and other activities associated with maintaining state UI IT programs."[122]

Puzzlingly, other DOL guidance seems to contradict these prohibitions, suggesting more meaningful flexibility for states to experiment with automation. For example, the most recent edition of the Employment and Training Administration Handbook on UI from 2012 explicitly acknowledges that some states have implemented automated systems making determinations:[123]

> In an effort to be more efficient, some states have implemented systems that issue nonmonetary determinations on certain limited issues solely on the basis of claimants' responses about their eligibility into an automated system without adjudicator intervention. Issues concerning a claimant's availability for work, or search for work, are often adjudicated in this manner in those states. Automated nonmonetary determinations must meet all quality guidelines outlined in Chapter V. Most importantly, facts must lead to only one conclusion on the issue; an adjudicator must intervene if they do not.

This guidance expressly allows for some forms of automation within adjudication, while maintaining legal defensibility by requiring such automation to occur only in areas of narrow discretion—i.e., when facts "lead to only one conclusion on the issue"—though this would seem to be a very difficult line to police.[124] Resolving these internal contradictions is of increasing importance given the rise of AI systems for states at the frontlines facing crushing caseloads.

These challenges surrounding automation and merit staffing in UI programs are not unique. Unlike UI, which has had a long-standing merit staffing requirement, SNAP had a later and more gradual legislative development. Despite an initial four-year run from 1939-1943, formal authorizing legislation did not come until 1959, and the program did not become permanent until the 1964 Food Stamp Act.[125] This Act provided states with considerable flexible in program administration, allowing

---

120 *See* discussion of Michigan's failed system in *infra* Part III.B.i and IV.A.i

121 UIPL 12-01 Change 2, *supra* note 119.

122 *Id.*

123 U.S. Dep't of Labor, Emp. & Training Admin., Unemployment Insurance Program Letter No. 01-13 (Oct. 1, 2012), https://www.dol.gov/sites/dolgov/files/ETA/advisories/UIPL/2012/UIPL_1-13.pdf (including the new ETA Handbook as an attachment).

124 This language seems modeled after A-76's guidance on the level of discretion required to categorize an activity as inherently governmental: "the use of discretion shall be deemed inherently governmental if it commits the government to a course of action *when two or more alternative courses of action exist and decision making is not already limited or guided by existing policies*." OMB CIRCULAR A-76, *supra* note 102.

125 *A Short History of SNAP*, FOOD & NUTRITION SERV., https://www.fns.usda.gov/snap/short-history-snap; Pub. L. 86-341, 73 Stat. at 608 ("The Secretary shall issue, to each welfare department or equivalent agency of a State . . . food stamps for each kind of surplus food to be distributed."). This 1959 Act defined "needy persons" as the eligible recipients for such support, and defined a need person as "anyone receiving welfare assistance" or anyone "who is, in the opinion of such agency or agencies, in need of welfare assistance but is ineligible to receive it because of State or local law." *Id.* at 608-09. These concepts were more fleshed on the 1964 Act, which defines eligibility based on an income level that is "a substantial limiting factor in the attainment of a nutritionally adequate diet." Pub. L. 88-525, 78

them to set their own plans of operation, eligibility standards, certification processes, and privacy safeguards—all subject to federal approval.[126] Notably, however, the 1964 Act makes no mention of merit staff.

The first mention of merit staffing appeared in the 1977 update to the Act, which required that state agency personnel "utilized in undertaking [applicant] certification" be employed in accordance with merit system standards.[127] Such a change was likely a natural outflow of the passage of the Intergovernmental Personnel Act of 1970, which sought to improve the quality of public service in state and local governments, in part, by merit system requirements.[128] The most recent legislative authorization for the program, the Food and Nutrition Act of 2008, maintains the same mandate: state agencies must certify applicants in accordance with "general procedures" set out by USDA and the personnel "utilized in undertaking such certification" must be employed in accordance with merit standards.[129]

This is a very differently scoped legislative merit staffing mandate than DOL's, which required merit personnel across the board and was only later narrowed to roles involving inherently governmental activities. SNAP's implementing regulations are even more specific, specifying that only merit staff—not "[v]olunteers and other non-State agency employees"—may "conduct certification interviews or certify SNAP applicants."[130]

Despite the seemingly narrower focus on certification, state SNAP agencies have still expressed confusion and frustration with the boundaries of this merit staffing requirement, particularly during high-volume times like economic downturns. For example, states' use of private firms to help households complete applications, gather information, and enter information into state automated data systems resulted in a 2009 rebuke from FNS asserting that such outsourcing results in a "more complex and cumbersome" enrollment process.[131] FNS then further restricted tasks that "involve any

---

Stat. 703, 704. The Act directs states to set eligibility standards consistent with those used in administration of other "federally aided public assistance programs," subject to Secretary approval, and mandates these eligibility evaluations to consider income as well as "resources." *Id.*

126 *Id.*

127 Pub. L. 95-113, 91 Stat. 913, 971.

128 Pub. L. 91-648, 84 Stat. 1909. As part of this goal, OPM established an advisory council from which it mandated a report on the "feasibility and desirability of extending merit policies and standards to additional Federal-State grant-in-aid programs." *Id.* at 1911. The Act also defined a set of merit principles that it believed should be central to personnel administration systems. *Id.* at 1909. For a fuller discussion of the wide-ranging impacts of the Intergovernmental Act of 1970 and its mobility program, see Isaac Cui, Daniel E. Ho, Olivia Martin, & Anne Joseph O'Connell, *Governing by Assignment*, 173 U. PENN. L. REV 157 (2024).

129 Pub. L. 110-246, 122 Stat. 1652.

130 7 C.F.R. § 272.4. The regulations do, however, explicitly exempt the certification process for SSI households, rural Alaskan households, and disaster victims from merit staffing requirements. Moreover, the regulations explicitly encourage the use of volunteers or "outreach, prescreening, assisting applicants in the application and certification process, and in securing needed verification."

131 Memo from U.S. Dep't of Agric., Food & Nutrition Serv., to Regional Directors, Federal Enrollment Support for Supplemental Nutrition Assistance Program (Oct. 1, 2021), https://www.fns.usda.gov/snap/admin/federal-enrollment-support. *Merit Staff Questions and Answers*, FOOD & NUTRITION SERV. (Jun. 30, 2010), https://fns-prod.azureedge.us/sites/default/files/snap/Merit-Staff-Questions-Answers.pdf. FNS did not see increased caseloads as adequate justification to increase merit staffing flexibility:

> Question 11: What does FNS suggest State agencies do to manage case-loads during increased volume when States have time-limited funds thus inhibiting their ability to hire permanent merit staff employees?
> Answer 11: State Agencies may take full advantage of hiring non-merit temporary staff to perform activities that do not involve client contact, such as data entry, typing, data matching, and document scanning, to free up the time of merit staff who handle all tasks involving client contact. States can seek prior approval from FNS to

client contact" to merit staff. The FNS later relaxed this requirement in in 2017 to allow non-merit staff in call centers to provide case information to individuals (but not to alter case status) and in 2020 allowed such personnel to screen for eligibility, provide application assistance, answer client question about missing information, and provide verification guidance (albeit with case-by-case FNS approval).[132]

In recent years, debates over merit staffing in SNAP have increasingly focused on automation in the eligibility determination process. While FNS has encouraged states to experiment with robotic process automation (RPA), the precise boundaries of automation remain unclear.[133] An FNS-commissioned study examining several states' RPA implementation interpreted 7 C.F.R. § 272.4 as "requir[ing] that a [merit staff] worker make the final decision on every case," though the regulation does not use the language "final decision."[134] The study concluded that "RPA cannot authorize, deny, or change benefit status but can be used to organize and edit information to 'stage the case' for an eligibility worker."[135] FNS's summary of this report also interprets SNAP regulations as requiring merit workers to "make the final decision on every case," and notes that "[b]ecause a worker still needs to review any updates made to a SNAP case by an RPA, the number of tasks assigned to a worker does not diminish, though they may not need to spend as much time on each task."[136]

Yet despite these apparent limitations, FNS's most recent guidance on advanced AI-based automation contains little discussion of the merit staffing requirements of § 272.4.[137] Instead, the agency has situated another rule—one that requires states to report "major changes in their operation of SNAP" to FNS—as the regulatory boundary on state automation efforts.[138] The rule explicitly identifies "[s]ubstantial increased reliance on automated systems for the performance of responsibilities previously performed by State merit system personnel" and "[a]ny reduction or change of the functions or responsibilities currently assigned to SNAP merit system personnel" as major changes.[139] Notably, however, this rule imposes only a reporting requirement, rather than an

---

allow non-merit temporary staff to perform limited client contact tasks, such as application assistance, that do not involve interview and certification activities.

132 *Id.;* Memo from U.S. Dep't of Agric., Food & Nutrition Serv., to Regional Directors, Revised Non-Merit System Personnel Guidance for Call Centers (Dec. 19, 2017), http://www.transparency.ri.gov/uhip/documents/legislative-reports/01-05-2018/12-19-17%20Revised%20Non-Merit%20System%20Personnel%20Guidance%20December%202017.pdf. Memo from U.S. Dep't of Agric., Food & Nutrition Serv., to Regional Directors, Revised Guidance for Use of Vendor/Private Staff in Call Centers: 2020 Update (Feb. 6, 2020), https://fns-prod.azureedge.us/sites/default/files/resource-files/Non-merit%20call%20center%20guidance%20and%20revised%20policy%20February%202020%20FINAL.pdf.

133 U.S. FOOD & NUTRITION SERV., ANALYSIS OF ROBOTIC PROCESS AUTOMATION IN SUPPLEMENTAL NUTRITION ASSISTANCE PROGRAM: THREE CASE STUDIES: FINAL REPORT, https://fns-prod.azureedge.us/sites/default/files/resource-files/snap-bots-rpa-final-report.pdf.

134 *Id.*

135 *Id.* The report also notes that "RPA must comply with policy on what action the State takes when the information is questionable, unclear or considered known to the State agency from another program administered by the same state agency."

136 U.S. FOOD & NUTRITION SERV., ANALYSIS OF ROBOTIC PROCESS AUTOMATION IN SUPPLEMENTAL NUTRITION ASSISTANCE PROGRAM: THREE CASE STUDIES: RESEARCH SUMMARY (2023), https://fns-prod.azureedge.us/sites/default/files/resource-files/snap-bots-rpa-summary.pdf.

137 Memo from U.S. Dep't of Agric., Food & Nutrition Serv., to Regional Directors, Use of Advanced Automation in SNAP (Jan. 10, 2024), https://www.fns.usda.gov/snap/advanced-automation.

138 7 C.F.R. § 272.15.

139 *Id.* One example of such reliance on automation is "adding an overlay on an existing legacy automated system used by eligibility workers." Unless the change will impact less than five percent of the state's SNAP applicants or participants, it is considered a major change.

operational limitation. FNS has approved a number of demonstration waivers from states to use automation to "stage the case" for eligibility workers.[140]

Even in its response to the Biden AI Executive Order,[141] which expressly directed agencies to issue guidance on the use of automated systems "to ensure that programs using those systems . . . employ automated or algorithmic systems in a manner consistent with any requirements for using merit systems personnel in public-benefits programs,"[142] FNS largely avoided discussing merit staffing obligations.[143] In conclusory fashion, the response simply stated: "All AI must be used in compliance with program requirements for the use of merit systems personnel, such as those applicable to SNAP."[144]

* * * *

In this Part, we have shown that the emergence of the civil service was intimately intertwined with the rise of the social safety net. The legal lines drawn by Congress and the White House to protect what is "inherently governmental," however, have grown increasingly untenable. Technology is central to effective administration of large-scale governmental programs, and federal agency guidance manifests deep conflicts, contradictions, and uncertainties about the boundary between human and automated decisionmaking in these programs. Attempts to clarify, such as the FNS's conclusory response to President Biden's AI Executive Order, have achieved little but to cloud the picture. In the next Part, we show that this legal uncertainty has led to a cacophony of state experimentation.

III. Experimentation Under Legal Uncertainty

Murky and internally contradictory federal guidance has not stopped states from experimenting with AI. To the contrary, state agencies facing unprecedented onslaughts of claims during the pandemic had few alternatives. Workforces cannot easily be tripled; nor can new employees be quickly trained in byzantine rules of eligibility.[145] Faced with the Hobson's choice of either complying with (a) federal timeliness and quality standards or (b) uncertain merit staffing requirements, many states chose the former, experimenting with integrating AI-based systems into benefits adjudication. This section situates these state experiments in the broader theory of democratic experimentalism, catalogues three distinct waves of experimentation with AI across states, and contrasts this approach to modernization to a series of automation and reform efforts at IRS that centered evidence and evaluation.

A. Unfacilitated Experimentalism

The absence of meaningful federal involvement, oversight, and supervision of state efforts is significant. States' exploration of AI tools may exemplify the role of states as "laboratories of

---

140 Advanced Automation in SNAP Memo, *supra* note 137.
141 Exec. Order No. 14110, *Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, 88 Fed. Reg. 75191 (Nov. 1, 2023), https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence. This Executive Order was rescinded by President Trump in Exec. Order No. 14145, *Initial Rescissions of Harmful Executive Orders and Actions*, https://www.whitehouse.gov/presidential-actions/2025/01/initial-rescissions-of-harmful-executive-orders-and-actions/.
142 *Id.*
143 *Framework for State, Local, Tribal, and Territorial Use of Artificial Intelligence for Public Benefit Administration*, FOOD & NUTRITION SERV. (Apr. 29, 2024), https://www.fns.usda.gov/framework-artificial-intelligence-public-benefit.
144 *Id.*
145 Richardson & Pahlka, *supra* note 51.

democracy,"[146] an idea refined through theories of democratic experimentalism. In their seminal article, *A Constitution of Democratic Experimentalism*, Michael Dorf and Charles Sabel propose democratic experimentalism as a new model of deliberative democratic governance positioned to grapple with the complexity and volatility of the modern administrative state.[147] Decentralized decision-making fosters innovation through policy experiments, like states testing AI-based process improvements.

But the success of democratic experimentalism hinges on effective information pooling mechanisms that allow disparate entities to learn from different models. According to Dorf and Sabel, one of the "chief purposes" of federal agencies is to facilitate this information sharing by "creating the infrastructure of decentralized learning" and assisting state and local governments in "benchmarking and experimentalism generally."[148] The federal government can additionally provide necessary political support—or permission structures—to embolden risk-averse agencies by setting clear expectations for experimentation and offering safe harbors in order to foster genuine innovation.

It is nigh impossible for federal agencies to simultaneously demand compliance with merit staffing regulations that, at best, seriously circumscribe the use of AI in benefits adjudication while also facilitating meaningful information exchange across states.[149] And yet, some innovative groups within agencies have tried to navigate this thicket of departmental mandates, such as the multi-disciplinary Tiger Teams US DOL launched in 2021 to assist state UI agencies in identifying systems improvement opportunities.[150] A number of the Tiger Teams' recommendations centered around automation to improve timeliness and address pandemic-related case backlogs.[151] However, the Tiger Team's funding was cut when the Fiscal Responsibility Act of 2023 clawed back $1B of modernization funding that had been allocated to UI.[152]

Because of this clawback and the legal limbo of automation, little systematic documentation and information pooling of states efforts exists. We hence offer a survey of the diversity of innovation waves, culminating in the use of generative AI.

B. The Three Waves of AI Tools

We document three distinct waves of state experimentation with AI in the benefit adjudication space. The first is the use of machine learning to detect fraudulent applications. The second is the use of chatbots that directly engage with applicants to answer common questions. The third, newly emerging, is the use of AI as a decision tool to assist adjudicators in deciding claims.

---

[146] New State Ice Co. v. Liebmann, 285 U.S. 262, 311 (1932) (Brandeis, J., dissenting).
[147] Dorf & Sabel, *supra* note 23; *see also* Daniel E. Ho, *Peer Review*, 69 STAN. L. REV. 1, 17 (2017).
[148] Dorf & Sabel, *supra* note 23.
[149] This compliance orientation can be exacerbated by its officers. For example, E.O. 14110 and the OMB's M-Memo on agency use of AI requires that agencies designate a Chief AI Officer (CAIOs). Safe, Secure, and Trustworthy Artificial Intelligence, Exec. Order No. 14,110, 88 Fed. Reg. 210 (Oct. 30, 2023). The intention of the CAIO proposal is to create officials that balance innovation and compliance. However, the role for such CAIOs in state experimentation is unclear, due to the fact that the M-Memo's compliance requirements (discussed *infra* Part IV) do not reach state innovation and the likely reluctance of states to collaborate with a federal officer that might reveal their merit staffing missteps.
[150] U.S. Dep't of Labor, Emp. & Training Admin., Unemployment Insurance Program Letter No. 2-22 (Nov. 2, 2021), https://www.dol.gov/sites/dolgov/files/ETA/advisories/UIPL/2021/UIPL_02-22.pdf; *see also Tiger Team Updates*, *supra* note 78.
[151] Tiger Team Updates, *supra* note 78.
[152] *Department of Labor Announces $377M in Available Grants to States to Strengthen Unemployment Insurance Programs, Modernize Systems,* DEP'T OF LABOR (July 13, 2023), https://www.dol.gov/newsroom/releases/eta/eta20230713-1; Joint Economic Committee Democrats, *Modernizing Benefit Systems Can Improve Lives While Also Saving Time and Money* (Mar. 15, 2024), https://www.jec.senate.gov/public/index.cfm/democrats/2024/3/modernizing-benefit-systems-can-improve-lives-while-also-saving-time-and-money.

i. Wave One: Fraud Detection

Fraud has long been a concern for agencies administering benefits,[153] and machine learning has a long history of fraud detection, with core motivating applications in consumer finance, spam filtering, and telecommunications.[154] However, early implementations of AI fraud detection systems had major issues, sometimes to disastrous results.

Perhaps the most notable example is the 2013 introduction of the Michigan Integrated Data Automated System (MiDAS), a $47 million system developed by private vendors to detect fraud in unemployment applications.[155] Between 2013 and 2015, the state sent accusations of fraud to approximately 40,000 Michigan residents and seized millions of dollars in wages and tax returns refunds as a result.[156] Later analysis found a 93% error rate in the system's fraud determinations and that most MiDAS cases received no human review before going into effect.[157] After nearly a decade of litigation, over 8,000 affected claimants received a $20 million settlement for the erroneous determinations.[158] The debacle has since been upheld by legal scholars as an example of "stategraft", a term coined by Bernadette Atuahene to describe the transfer of property from persons to the state in violation of state law or basic human rights.[159] It seems somewhat unlikely that Michigan intentionally designed the tool to "improperly seize[]" citizens' funds to "grow state coffers";[160] moreover, the more likely transfer would seem to be from citizens to unaccountable contractors, who designed the $47 million system and ultimately paid out a mere $180,000 in a federal settlement.[161]

Although Michigan's example stoked fear in state unemployment administrators—and continues to cast a long shadow over automation to present day—many states continued to cautiously experiment with AI-based fraud detection tools. The need for this experimentation was heightened by the COVID-19 pandemic, when states received an unprecedented number of fraudulent

---

153 *See* Fiscal Responsibility Act of 2023, Pub. L. 118-5, 137 Stat. 11 (2023) (shifts focus of modernization funds to fraud detection).

154 Emmanuel Gbenga Dada et al., *Machine Learning for Email Spam filtering: Review, Approaches and Open Research Problems*, 5 HELIYON 1 (2019; Dejan Varmedja et al., *Credit Card Fraud Detection – Machine Learning Methods*, 2019 18TH INT'L SYMPOSIUM INFOTEH-JAHORINA.

155 de la Garza, *supra* note 15.

156 *Id.*; Alex Eberg, *Triple Payouts Approved for Jobless Claims Stripped by Faulty AI*, BLOOMBERG LAW (Jan. 29, 2024), https://news.bloomberglaw.com/daily-labor-report/triple-payouts-approved-for-jobless-claims-stripped-by-faulty-ai.

157 *See* Kohl, note 15; Jonathan Oosting, *Michigan Settles Federal Unemployment Fraud Case*, THE DETROIT NEWS (Feb. 2, 2017), https://www.detroitnews.com/story/news/local/michigan/2017/02/02/michigan-settles-federal-unemployment-fraud-case/97395906/; Stephanie Wykstra, *Government's Use of Algorithm Serves Up False Fraud Charges*, UNDARK (June 1, 2020), https://undark.org/2020/06/01/michigan-unemployment-fraud-algorithm; Sonia M. Gipson Rankin, *The Midas Touch: Atauhene's "Stategraft" and Unregulated Artificial Intelligence*, 98 NYU L. REV. ONLINE 225 (2024).

158 Press Release, Mich. Dep't of Att'y Gen., State of Michigan Announces Settlement of Civil Rights Class Action Alleging False Accusations of Unemployment Fraud (Oct. 20, 2022), https://www.michigan.gov/ag/news/press-releases/2022/10/20/som-settlement-of-civil-rightsclass-action-alleging-false-accusations-of-unemployment-fraud. The Sixth Circuit had previously rejected state officials' claims of qualified immunity on the basis that the MiDAS system—not defendants—caused the due process deprivations. The court emphasized that state employees "implemented and oversaw MiDAS," "prescribed its operation" and "enforce[d] [its] false fraud determinations." Cahoo v. SAS Analytics Inc., 912 F.3d 887, 904 (6th Cir. Mich. 2019).

159 *See* Bernadette Atuahene, *A Theory of Stategraft*, 98 NYU L. Rev. 1 (2023); Kohl, *supra* note 15; Rankin, *supra* note 157

160 Kohl, *supra* note 15 at 47.

161 *See infra* Part IV.A.i. The district court had denied summary judgment, finding that the State relied heavily on the expertise of its contractors, which designed and implemented the system, trained state employees, monitored its performance, and conducted cost-benefit analyses. Cahoo v. Fast Enterprises LLC, 528 F. Supp. 3d 719, 734-740 (E.D. Mich. Mar. 25, 2021).

unemployment claims.[162] States across the country turned to a set of private vendors—namely, Google,[163] Deloitte,[164] Pondera,[165] Maximus,[166], and FAST Enterprises[167]—to implement machine learning to filter out the fraudulent claims inundating their systems.

While no single system has generated as much controversy as Michigan's MiDAS, contracts with vendors are often large and riddled with issues. The Securities and Exchange Commission, for instance, recently filed charges against the one CEO for lying about the efficacy of the company's touted AI fraud detection tool.[168] And, a system built in December 2020 for California's UI agency by Pondera, a provider of algorithmic fraud detection owned by Thomson Reuters, to investigate fraudulent claims has shown similar unreliability. After the department stopped payment on 1.1 million claims Pondera identified as potentially fraudulent more than half of them—600,000—were found to be legitimate.[169] Other systems left gaping holes in security, such as the fraud detection system developed by Deloitte for Ohio which allowed a subcontractor to remove fraud flags on applications as part of an illicit side hustle that she advertised on Instagram.[170]

The federal Department of Labor acknowledged the crisis in fraudulent claims and allocated a significant proportion of the funds it received under ARPA to assist states in modernizing systems

---

162 GAO 2023 UI Fraud Report, *supra* note 11. GAO estimated that fraud accounted for approximately 11-15% of the total amount of unemployment benefits paid during the benefit. A central part of the fraud detection challenge is the role of identity verification systems; DOL explicitly encouraged the development of such systems during the pandemic, but such systems have also come under close scrutiny for civil rights concerns. https://www.dol.gov/sites/dolgov/files/ETA/advisories/UIPL/2021/UIPL_16-21.pdf. For example, the ACLU of New York sued the state UI agency for its implementation of ID verification using ID.me, alleging that it is an "error-prone, inaccessible, unreliable, and invasive automated facial recognition tool." https://www.aclu.org/press-releases/nyclu-aclu-sue-new-york-state-department-of-labor-for-withholding-records-on-automated-identity-verification-tools.

163 Melissa Adamson & Prabhu Palanisamy, *Using AI-Powered Machine Learning Models to Identify Fraudulent Unemployment Claims*, GOOGLE CLOUD BLOG (May 26, 2021), https://cloud.google.com/blog/topics/public-sector/using-ai-powered-machine-learning-models-identify-fraudulent-unemployment-claims

164 *Id;* de la Garza, *supra* note 15.

165 Lauren Helper, *Internal Documents Reveal the Story Behind California's Unemployment Crash*, CAL MATTERS (Nov. 7, 2023), https://calmatters.org/economy/2023/11/california-unemployment-covid/; Press Release, Thomson Reuters, Thomson Reuters Acquires Pondera Solutions (Mar. 19, 2020), https://www.thomsonreuters.com/en/press-releases/2020/march/thomson-reuters-acquires-pondera-solutions.html.

166 MAXIMUS, HELPING STATES WITH UNEMPLOYMENT INSURANCE SURGE – QUICK FACTS, https://www.naswa.org/system/files/2021-03/maximus-ui-servicesnaswa-2020.pdf.

167 Jennifer Lord, *Opinion: Artificial "Intelligence": Unemployment System Denied Legitimate COVID-19 Claims*, THE DETROIT NEWS (Nov. 18, 2020), https://www.detroitnews.com/story/opinion/2020/11/19/opinion-unemployment-system-denied-legitimate-covid-19-claims/6339115002/; de la Garza, *supra* note 15.

168 Comp. ¶ 89, United States v. Roberts, 1:24-cr-00547 (S.D.N.Y. Sep. 16, 2024), https://storage.courtlistener.com/recap/gov.uscourts.nysd.628366/gov.uscourts.nysd.628366.1.0.pdf. The trend for companies to make fraudulent claims about their AI capacity has been dubbed "AI Washing." *See* Emma Woollacott, *What Is 'AI Washing' and Why Is It a Problem?*, BBC, June 26, 2024. For a study of the exaggerated claims in legal AI, see Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho, *Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools*, J. EMPIRICAL LEGAL STUD. (2025, forthcoming).

169 CALIFORNIA LEGISLATIVE ANALYST'S OFFICE, THE 2022-23 BUDGET: ASSESSING PROPOSALS TO ADDRESS UNEMPLOYMENT INSURANCE FRAUD (2022), https://lao.ca.gov/Publications/Report/4542.

170 Thomas Brewster, *States Spent Millions on Deloitte's 'Anti-Fraud' Covid Unemployment Systems. They Suffered Billions In Fraud.*, FORBES (Nov. 3, 2022), https://www.forbes.com/sites/thomasbrewster/2022/10/31/covid-pandemic-fraud-hits-billions-despite-deloitte-contracts-worth-hundreds-of-millions/. Deloitte was familiar with such schemes: in June 2020, it had produced a report for Michigan that detailed a highly similar kickback scheme in that state's system.

and investing in fraud detection.[171] Although the Department of Labor's materials do not explicitly acknowledge the use of AI in fraud detection, oblique references signal tacit awareness: for example, its Tiger Team recommendations acknowledge that "[s]ophisticated risk analytics are available to assign a risk-based score to claims to help the state to detect suspect activity early in the claims process, as well as minimize the number of false positives which helps to protect innocent claimants from being flagged for fraud."[172]

ii. Wave Two: Chatbots

Chatbots are a second form of AI-based tool that have exploded across all levels of government, including benefits agencies.[173] For example, Mississippi has effectively replaced its state homepage with a chatbot interface.[174] Like fraud detection systems, chatbots don't have to be AI-based; they can be built on relatively simple technology, such as a decision tree (a series of if-then statements) with prepackaged topics and answers, or they can use generative AI to allow for more free-form, adaptive conversations.[175]

Many states began launching decision-tree-based chatbots during the pandemic when call centers were overwhelmed.[176] For example, according to a GAO survey, unemployment insurance claimants in four states reported call wait times of up to 1 to 3 hours during the pandemic, and claimants in Florida and Wyoming reported wait times of 8 hours or more.[177] By mid 2020, nearly three-quarters of states had launched chatbots to address common questions about COVID-19 and unemployment insurance.[178] While most states employed user-facing chatbots to alleviate burdens on call centers,

---

171 *Preventing Fraud*, U.S. DEP'T OF LABOR, https://www.dol.gov/agencies/eta/ui-modernization/fraud; U.S. Dep't of Labor, Emp. & Training Admin. Unemployment Insurance Program Letter No. 22-21, Change 2 (Apr. 27, 2023), https://www.dol.gov/sites/dolgov/files/ETA/advisories/UIPL/2021/UIPL%2022-21%20Change%202/UIPL%2022-21%20Change%202.pdf.
172 *Tiger Team Cohort Trends*, U.S. DEP'T OF LABOR, https://oui.doleta.gov/unemploy/pdf/TigerTeamCohortTrendsJune_2022.pdf.
173 *See* Keely Quinlan, *Chatbot Snapshot: How State, Local Government Websites Use AI Assistants*, STATESCOOP (July 17, 2024), https://statescoop.com/government-ai-chatbots-state-local-websites-2024/; Tzhuhao Chen, Mila Gascó-Hernandez, & Marc Esteve, *The Adoption and Implementation of Artificial Intelligence Chatbots in Public Organizations: Evidence from U.S. State Governments*, 54 AM. REV. OF PUB. ADMIN. 255 (2024). Governments are not the only entities creating benefits-minded chatbots. *See The Helping Hand*, UPRISE RI, https://upriseri.com/hh/ (designed by a political advocacy group).
174 *See Official Website of Mississippi,* ms.gov.
175 For example, New York City launched generative AI-based chatbot to answer common business questions in October 2023. However, within days of launching, users pointed out concerning misstatements of law, such as a message from the chatbot that "suggested it is legal for an employer to fire a worker who complains about sexual harassment." Jake Offenhartz, *NYC's AI Chatbot Was Caught Telling Businesses to Break the Law. The City Isn't Taking It Down*, AP NEWS (Apr. 3, 2024), https://apnews.com/article/new-york-city-chatbot-misinformation-6ebc71db5b770b9969c906a7ee4fae2.
176 Call systems continued to have extremely long wait times after the pandemic. For example, SNAP claimants recently brought a claim against the Acting Director of the Missouri Department of Social Services alleging that they had waited on the phone for hours to complete their SNAP interview without being connected with a representative and subsequently had their SNAP application denied for failure to complete the interview. Holmes, No. 2:22-CV-04026.
177 *See* 2022 GAO REPORT, *supra* note 8.
178 Colin Wood, *Nearly 75% of States Launched Chatbots to Aid Pandemic Response*, STATESCOOP (June 26, 2020), https://statescoop.com/nearly-75-percent-states-launched-chatbots-aid-pandemic-response/; NAT'L ASS'N OF STATE CHIEF INFO. OFFICERS, CHAT WITH US: HOW STATES ARE USING CHATBOTS TO RESPOND TO THE DEMANDS OF COVID-19 (2020), https://www.nascio.org/wp-content/uploads/2020/06/NASCIO_ChatbotsRespondtoCOVID-19.pdf.

some states also launched internal assistants to help staff learn more quickly and answer basic questions.[179]

To systematically evaluate the current landscape of automated assistance in unemployment insurance administration, we conducted a comprehensive audit of state-deployed chatbots. Through targeted Google searches of both general state websites and unemployment insurance-specific portals, we identified chatbots in twenty-six states.[180] Most chatbots did not appear highly reliant on advanced use of AI: instead, they offered pre-set topics and some ability to map natural language entries to predetermined answers.[181] Nearly every chatbot identified itself to users as a "chatbot" or some form of virtual agent.[182] Many tools offered to connect users with a human agent upon request, with some offering a live agent as soon as the user expressed dissatisfaction with an answer.[183] And a few chatbots—like those in Florida and California—even offered chat in multiple languages.[184]

1. Audit of All Identified Chatbots

Many states have touted the efficacy of these new chatbots, with Georgia going so far as to claim that its AI chatbot "accurately identified requests 99% of the time via chat."[185] However, these chatbots have received little to no independent evaluation. To evaluate chatbot effectiveness in assisting unemployment insurance claimants, we developed a three-tiered question framework that assessed performance across questions of increasing complexity:

---

[179] For example, New Jersey has launched an internal assistant that leverages generative AI to assist state employees—after they have taken an AI training course. *Generative AI in New Jersey State Government*, N.J. OFF. OF INNOVATION, https://innovation.nj.gov/skills/ai/; *New Chatbot Helps Answer IT Procurement Questions*, N.C. DEP'T OF INFORMATION TECH. (Apr. 4, 2024), https://it.nc.gov/blog/2024/04/04/new-chatbot-helps-answer-it-procurement-questions (a North Carolina chatbot that answers employee's IT procurement questions).

[180] If a state had chatbots on both pages, we continued the audit on their more unemployment insurance-focused chatbot. The availability of a chatbot was not necessarily predicted by the relative population size or income of a state; for example, we were unable to identify chatbots for Minnesota, Oregon, or Pennsylvania; however, Arkansas, Delaware, Iowa, and South Dakota all offered chatbots. Moreover, chatbots are also not necessarily a persistent phenomenon. For example, when we conducted the first version of this audit in August 2024, Nevada's Department of Labor offered a chatbot, but this option was gone as of our second audit in July 2025.

[181] This is descriptive of chatbots like those in Georgia and Arizona, for example.

[182] Texas's chatbot immediately connects users with a live agent as soon as it cannot answer a question, something it did in response to the first question on how to file for unemployment insurance.

[183] Interestingly, Utah's chat service is only available during "regular business hours," suggesting that it might be largely supported by human staff.

[184] Florida's chatbot was offered in English, Spanish, and Créole. California's chatbot offered English, Armenian, Chinese, Korean, Spanish, Tagalog, and Vietnamese.

[185] Keely Quinlan, *Georgia Labor Dept Says Upgraded AI Chatbot Is Highly Accurate*, STATESCOOP (Feb. 7, 2023), https://statescoop.com/georgia-labor-upgrades-ai-chatbot/. In our audit, Georgia's chatbot was unable to answer the relatively basic question of "What documents do I need to file for unemployment insurance?". *See also* Michigan Dep't of Labor and Economic Opportunity, *UIA Launches Chatbot That Leverages AI to Provide Key Information for Michigan Workers, Employers* (July 30, 2025), https://www.michigan.gov/leo/news/2025/07/30/uia-launches-ai-chatbot-to-provide-information-for-workers-employers (stating that the chatbot will deliver "quick and accurate responses"); Julia Edinger, *South Dakota IT Looks Ahead, Prioritizing the Citizen Experience*, GOVTECH (Feb. 20, 2024), https://www.govtech.com/workforce/south-dakota-it-looks-ahead-prioritizing-the-citizen-experience (South Dakota's IT director claiming that the tool is "much more accurate and much more responsive to a citizen's inquiry" than existing service models). The chatbot on the New York State Department of Labor's home page won an IT innovation award, with the state commissioner claiming that the AI-powered service was "sure to help meet the needs of its clients." Nat'l Ass'n of State Workforce Agencies, *New York State Earns National Recognition for IT Innovation* (Feb. 16, 2024), https://www.naswa.org/news/new-york-state-earns-national-recognition-for-it-innovation. However, the NYS chatbot was unable to answer any of our audit's three questions.

Complexity	***Question***	***Expected Answer***
Low	How do I file for unemployment insurance?	A brief description and a link to the application.
Medium	What documents do I need to file for unemployment insurance?	A brief description and a link to an explainer page.
High	Am I eligible for unemployment insurance if I quit my job due to health issues?	A brief explanation that eligibility is determined on a case-by-case basis and a link to a page explaining eligibility. Many state laws do allow claims in this circumstance.

Table 1: Questions and Expected Answers for Twenty-Six State Audit

We classified responses as "meaningfully helpful" when they directly addressed the question or provided relevant links to comprehensive answers. Responses that deflected, redirected inappropriately, or refused to engage with the substantive inquiry were deemed unhelpful. A full classification of responses can be found in Appendix Table A.

The results paint a disappointing picture of current chatbot efficacy in providing basic unemployment insurance guidance. Even on the simplest inquiry—how to file for benefits—only fourteen of the twenty-six systems provided meaningfully helpful responses. Examples of answers and their classifications are below in Table 2. States that failed to answer the first question typically used button-driven chatbot interfaces, despite also claiming to accept typed queries.

Answer Classification	***Example Answer***
Answered Well: Provided Steps	*South Dakota* To file for unemployment insurance in South Dakota, you can follow these steps: • Visit the South Dakota Department of Labor and Regulation website. • Look for the unemployment insurance section and find the link to file a claim online. • Fill out the required information, including personal details, employment history, and reason for unemployment. • Make sure to have necessary documents ready, like your Social Security number, contact information, and proof of prior employment. • Submit your claim online and wait for further instructions from the Department of Labor and Regulation.
Answered: Provided Link to Direct Answer	*Georgia* See the Get Unemployment Assistance page for information about the unemployment insurance program including an Apply for Unemployment Insurance Benefits video tutorial. Once you file a claim, you may check your claim and payment status on your MyUI Claimant Portal. [Page linked to directly answers the question]

Not Answered: Redirected	*Maine* Did you mean? Unemployment Benefits Eligibility Before Filing Initial Claim After Filing Initial Claim
Not Answered	*New York* I'm sorry, I didn't quite get that. Please try rephrasing your question in a few short words and ask again. Please note: I can only answer questions related to NYS Unemployment Insurance.

Table 2: Example Answers to the First Question

Performance degraded substantially with increased question complexity. Only eight chatbots provided helpful responses to the documentation requirements question. The remaining systems either acknowledged their limitations or, more problematically, provided occasionally misleading redirections. Example answer classifications are below in Table 3.

Answer Classification	***Example Answer***
Answered Well	*Connecticut* Have your Social Security card and a separation packet if one was provided from your employer. If you are separating from the military, have separation form DD214, Member-4. Federal employees will need separation form SF-8 and a copy of their most recent pay stub. If you are not a US citizen, you must have proof that you are work authorized in the USA. Do not delay filing a claim if you do not have these documents. Your claim can be filed without them. However, there may be a delay in payment until the document(s) are received.
Not Answered	*Virginia* To clarify, did you mean: How do I file an initial claim for benefits? How do I file my UI tax report and pay my UI taxes? None of these

Table 3: Example Answers to the Second Question

Most chatbots that failed to answer the second question either acknowledged their inability or redirected users to different questions, suggesting limited natural language processing capabilities.[186]

---

[186] Ohio's redirection was somewhat strange, stating (not incorrectly):

> Anyone can apply for unemployment insurance benefits, but you must have the proper work authorization. The Department of Homeland Security will verify your work authorization before benefits will be paid.

Ohio's tool also defaulted to assuming that the chatbot user was an employer for the first question, despite the fact that the question is posed as "How do I apply for unemployment insurance?".

More problematically, several states redirected to irrelevant questions, potentially confusing users. Colorado's redirection was relatively harmless, interpreting the question as asking about eligibility qualifications. Massachusetts redirected to overpayment information, responding: "You can find an overpayment waiver application by logging in to your Unemployment Services for Workers account." More concerning, New Jersey provided outdated pandemic information: "Federal benefits created during the pandemic including PUA, PEUC, and FPUC were discontinued as of Sept. 4. For more information about other benefit extensions, click the link below." This response could mislead claimants into believing unemployment insurance is no longer available.

The third and highest complexity question yielded predictably poor results, with only six states providing directly relevant responses that acknowledged health-related work separations. Most systems defaulted to generic eligibility discussions, though they generally emphasized the case-by-case nature of determinations—a substantively correct, if incomplete, response.[187]

2. Audit of Generative AI Chatbots

Not all chatbots depended on pre-set topics and decision trees. We identified five states that appeared to directly use generative AI to power their chatbots: Florida, Indiana, Michigan, Mississippi, and South Dakota. This was discernible either by explicit disclaimer about the use of generative AI (e.g., Michigan and South Dakota),[188] a lack of pre-set topics (Indiana and Mississippi), or in all cases, an ability to answer more complicated questions with contextually relevant responses. These systems substantially outperformed traditional rule-based alternatives on our initial three-question assessment, providing detailed procedural guidance and appropriate resource links, as illustrated by the comparison below.

---

[187] The answer from Colorado, however, could be interpreted as dissuasive:

> An individual who leaves work voluntarily without a good reason attributable to the employer is generally disqualified from receiving unemployment benefits. The eligibility will depend on whether the case demonstrates a good reason for quitting that is attributable to the employer. Quitting work without good cause just to obtain unemployment benefits qualifies as fraud, and could result in criminal penalties.

[188] Michigan requires user to view and accept a disclaimer before use. Indiana offers terms and conditions via a small link at the top of the chat. South Dakota warns users, via chat message at the beginning of each chat, that it is "a machine learning model, and [is] trained on information that may not be up-to-date," and reminders users to "[a]lways consider verifying information through our website or a live agent." Florida's Reemployment Assistant simply describes itself as a "virtual agent," and offers users the ability to "upvote" or "downvote" the usefulness of chatbot responses.

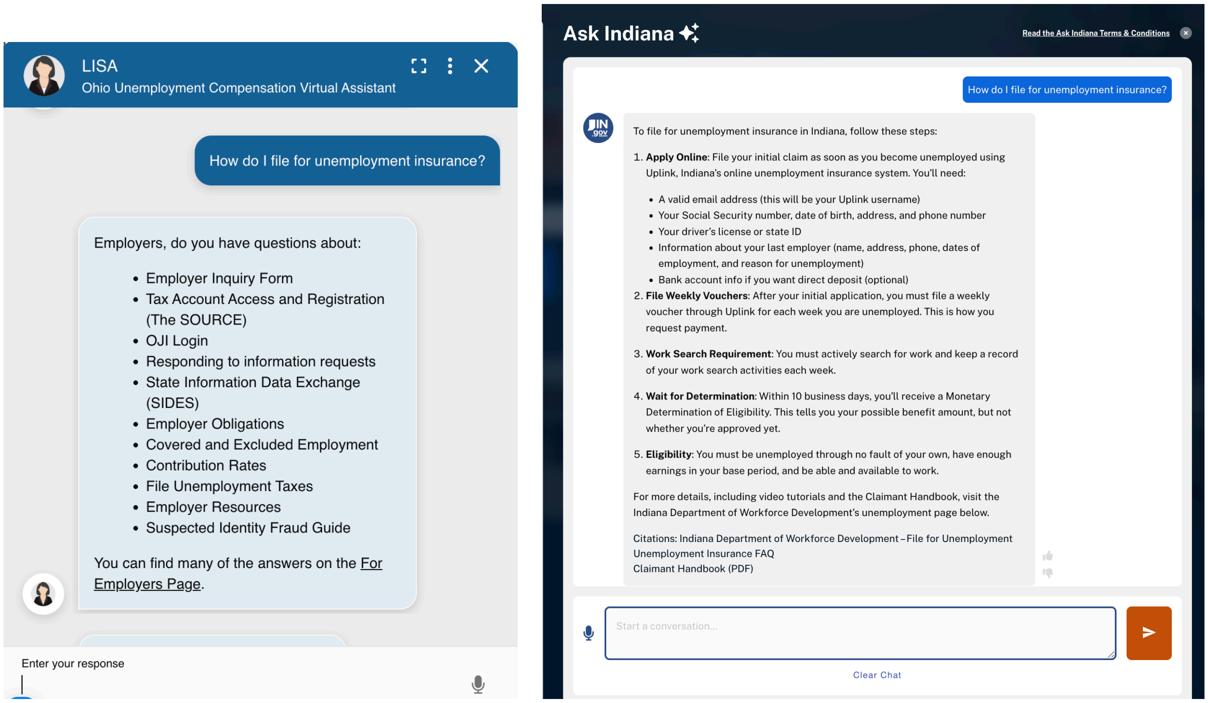

Figure 2: Comparing Ohio's Chatbot with Indiana's Generative AI Chatbot

To probe the boundaries of these advanced tools, we subjected the five systems to an expanded eleven-question evaluation featuring personalized scenarios and nuanced policy questions.[189] A full classification of responses can be found in Appendix Table B. The results revealed both the promise and peril of deploying large language models in government service delivery.

Indiana's chatbot performed most effectively, providing accurate responses aligned with state unemployment handbook guidance on ten of eleven questions. Indiana's sole error involved incorrectly stating that three work search activities were required weekly (the actual requirement is two).[190] Florida's chatbot achieved similar success rates, providing one incorrect response and failing to answer one question. Mississippi's chatbot performed next best, providing two partially incorrect answers. On one question, the chatbot incorrectly stated that claimants must have earned at least $780

[189] The evaluation was informed by the types of questions that have been found likely to provoke hallucinations in large language models. *See, e.g.,* Yanxu Zhu etl al., KG-FPQ: Evaluating Factuality Hallucination in LLMs with Knowledge Graph-based False Premise Questions, ARXIV 2407.05868 (2024), https://arxiv.org/abs/2407.05868 (false premise questions); Ben Snyder, Marius Moisescu & Muhammad Bilal Zafar, *On Early Detection of Hallucinations in Factual Question Answering*, 30TH ACM SIGKDD CONFERENCE ON KNOWLEDGE DISCOVERY & DATA MINING (KDD '24) (2024) (noting LLMs' difficulties with correctly retrieving facts). Questions also probed areas where states' unemployment law diverged from other states, following the logic used to develop a novel benchmark dataset to evaluate AI capabilities in simplifying state unemployment insurance laws. Emaan Hariri & Daniel E. Ho, *AI for Statutory Simplification: A Comprehensive State Legal Corpus and Labor Benchmark*, 20th INT'L CONF. ON A.I. & L. (2025).

[190] INDIANA DEP'T OF WORKFORCE DEVELOPMENT, UNEMPLOYMENT INSURANCE CLAIMANT HANDBOOK (April 2024) at 10-11, https://www.in.gov/dwd/files/Claimant_Handbook.pdf. Indiana was the only state we observed using a generative AI-based chatbot when we conducted the first initial, simple version of this audit in August 2024. At that time, the chatbot provided answers inconsistent with the state's unemployment manual on several questions. the tool responded that it was necessary to have earned "at least $4,200 in the base period, with at least $2,500 of that amount earned in one of the quarters." This contradicts the April 2024 state's claimant handbook—which is linked in the chatbot's response—and states that the $2,500 must have been earned in the last six months of the base period, not in one of the quarters. *Id.*

during the base period, whereas the state requires the claimant to have earned at least $780 in the highest quarter of the base period.[191] Base periods identify the specific 12-month period used to examine past earnings to determine monetary eligibility, and frequently cause confusion among applicants.[192] On a question about maximum benefit amounts, Mississippi's chatbot incorrectly stated that the maximum amount was $275, whereas the maximum is $235, though it correctly explained how each individual should calculate their maximum benefit amount.[193]

Michigan's performance proved somewhat more inconsistent, delivering correct answers on six questions while providing one partially correct response, one outright incorrect response and leaving three unanswered. South Dakota's system performed worst among the generative AI implementations, answering only three questions correctly while providing five non-responses, one contradictory answer, and two responses citing unverifiable numerical standards.

All five chatbots performed quite well on the non-monetary eligibility questions, including when posed a false premise question:[194]

> Question: "I know for a fact that if you are really dissatisfied with your job and leave because of that, you are still eligible for unemployment insurance. I left my job because my boss made me very unhappy. I'm still eligible for unemployment insurance, right?"

The systems did not fall for the bait, with four chatbots referring back to general eligibility guidelines and helpfully raising that some certain causes of work dissatisfaction—such as unsafe working conditions or illegal activities—could help create the conditions for good cause. Florida's system merely responded, "Thank you. What else can I help you with today?"—a response that, while unhelpful, avoided providing incorrect guidance.

The chatbots struggled significantly with mathematical calculations and date-specific determinations. When asked about minimum earnings requirements, South Dakota's chatbot cited a threshold of "$1,700 of earnings in at least two quarters of the base period," a specification we could not verify in any publicly available documentation.[195] Florida's system demonstrated concerning computational errors despite correctly stating the relevant formula. When applying the state's monetary eligibility standard to a hypothetical claimant, it incorrectly concluded that $7,000 was insufficient to meet the requirement of 1.5 times the highest wage quarter of $3,000 (though $7,000 clearly exceeds $4,500). Michigan's tool struggled to apply the state's relatively standard logic of

---

191 *See* Mississippi Dep't of Employment Security, UI Claimant Handbook, https://mdes.ms.gov/unemployment-claims/benefit-information/ui-claimant-handbook/.

192 *See* Nat'l Employment L. Project, *Monetary Eligibility Requirements* (Nov. 6, 2023), https://www.nelp.org/insights-research/monetary-eligibility-requirements/.

193 An amount of $235 is listed on the state's guidance for claimants as well as in statute. *See* Mississippi Dep't of Employment Security, *Table of Weekly Benefit Amounts and Maximum Amounts*, https://mdes.ms.gov/media/7140/Benefit_Chart1.pdf; Miss. Code §71-5-503 (2024). The chatbot may have been confused by a bill that had been proposed to raise the amount to $275; however, the bill died in committee in 2025. *See MS HB299*, BILL TRACK 50, https://www.billtrack50.com/billdetail/1773993.

194 False premise questions have been documented to lead to hallucination in large language models. *See, e.g.*, Hongbang Yuan et al., *Whispers that Shake Foundations: Analyzing and Mitigating False Premise Hallucinations in Large Language Models,* ARXIV 2402.19103 (2024), https://arxiv.org/abs/2402.19103; Yanxu Zhu et al., *KG-FPQ: Evaluating Factuality Hallucination in LLMs with Knowledge Graph-based False Premise Questions*, ARXIV 2407.05868 (2024), https://arxiv.org/abs/2407.05868.

195 There is no reference to this amount on the state's page about Reemployment Assistance Benefits, which simply states that there will be a monetary eligibility determination. https://dlr.sd.gov/ra/individuals/wage_requirements.aspx. It is also not defined on the website's FAQ, the claimant handbook, or the state's UI law. https://dlr.sd.gov/ra/individuals/faq.aspx#eligible; https://dlr.sd.gov/ra/publications/ra-claimant-handbook.pdf; https://sdlegislature.gov/Statutes/61-6.

defining of defining a base period of dates applicable to a claim's calculations, answering both of these questions incorrectly.[196] South Dakota's answer to one base period question was completely improbable. Asked about the relevant base period for unemployment occurring between October 7, 2018 and January 5, 2019—a period that should be prior to the unemployment spell—the system responded:

> The base period for filing an unemployment claim based on being unemployed between October 7, 2018, and January 5, 2019, will likely fall within the timeframe of July 1, 2018, to June 30, 2019.

This response suggests a base period extending beyond the unemployment spell itself, a logical impossibility that highlights the risks of deploying systems that can generate plausible-sounding but factually incorrect information.

Overall, these audit results underscore the critical need for rigorous benchmarking as a standard component of government AI evaluation frameworks. Traditional rule-based chatbots failed to provide meaningful assistance on even basic inquiries in nearly half of cases. Generative AI systems performed better overall, providing more relevant and detailed responses, yet they also introduced new risks by presenting erroneous information with unwarranted confidence. Issues such as computational mistakes and fabricated policy references are unlikely to be detected through casual use or standard user-satisfaction surveys. Systematic testing against defined benchmarks is therefore essential for agencies to uncover and address these failure modes prior to deployment.

3. Legal and Policy Implications

In some ways, the chatbots' inaccuracies are nothing new: throughout the nation's history, government representatives have sometimes given wrong information to those seeking advice. And unlike in agreements between private parties, there is not much the citizen in question can do because there is generally no equitable estoppel against the government. That is, if an agent gives a person bad advice—such as telling someone not to apply for benefits under a mistaken belief that the person is ineligible—that the person relies on to their detriment, that person has no estoppel claim against the agency.[197] The Supreme Court has given two primary justifications for this stance. First, it protects public dollars from being expended based on error.[198] Second, it upholds separation of powers principles by preventing the judiciary from doling out congressionally controlled

---

[196] In response to a question asking about a claim filed for unemployment occurring between January 5, 2020 and April 4, 2020, Michigan's tool correctly stated the rule that a base period is determined by looking at the first four of the last five completed quarters before that filing period. It then sets out the reasonable assumption that the user filed in Q1 2020, correctly lists out the five preceding quarters (Q4 2019, Q3 2019, Q2 2019, Q1 2019, and Q4 2018), but then incorrectly selects the latter four quarters rather than the first four quarters (it selects only quarters in 2019).

[197] *See* Schweiker v. Hansen, 450 U.S. 785 (1981) (per curiam); Off. of Personnel Mgmt. v. Richmond, 496 U.S. 414 (1990).

[198] *See* Hansen, 450 U.S. at 788 n.4; *see also* Fred Ansell*, Unauthorized Conduct of Government Agents: A Restrictive Rule of Equitable Estoppel against the Government*, 53 U. CHI. L. REV. 1026, 1033-34 (1986). Estoppel against the government could have serious unexpected fiscal impacts even in small situations. For example, the Seventh Circuit permitted an estoppel claim against the Postal Service, whose agent erroneously assured a citizen of $50,000 of package insurance and later realized its error and offered only to pay $500. Portmann v. United States, 674 F.2d 1155 (7th Cir. 1982). However, as Ansell notes, such claims are unlikely to bankrupt the federal government, and this argument stands in tension with Congress's willingness to authorize "large and unpredictable" amounts in tort claims against the government. *See* Federal Torts Claim Act (FTCA), 28 U.S.C. §§ 2671-2680 (1982).

appropriations on the basis of an executive agent's mistake.[199] In 1997, the Third Circuit once carved out an exception to this rule, holding that the IRS was equitably estopped from making an assessment after a series of affirmative misrepresentations over a decade.[200] The court estopped the agency due to its balance of the "special factors" that allow for equitable estoppel against the government: "1) the impact of the estoppel on the public fisc; 2) whether the government agent or agents who made the misrepresentation or error were authorized to act as they did; 3) whether the governmental misconduct involved a question of law or fact; 4) whether the government benefitted from its misrepresentation; and 5) the existence of irreversible detrimental reliance by the party claiming estoppel."[201] However, this exception to government estoppel doctrine has almost never been successfully applied elsewhere.[202]

There have not yet been American cases against government chatbots on estoppel grounds, though in a private dispute, a Canadian tribunal found Air Canada had engaged in a tort of negligent misrepresentation for misinformation provided by an AI-driven chatbot.[203] American courts would likely deny such claims against the government as they continue to do against errant agents administering benefits.[204] Nevertheless, chatbots offer guidance that occupies an uncertain legal space. Agency guidance escapes notice and comment and is more insulated from judicial review, but individuals often rely on informal agency guidance, even though it is legally non-binding.[205] And chatbot guidance is even more likely to induce reliance than average agency postings, Joshua Blank and Leigh Osofsky argue, because of the way in which it presents complex law as though it is simple (what they term "simplexity") and its personalized, non-qualified, and instantaneous means of communication.[206]

Generative AI chatbot errors differ from the nuanced issue of simplexity that Blank and Osofsky found when engaging with the IRS's Interactive Tax Assistant. The generative AI chatbots often performed well on nuanced eligibility questions, but sometimes gave completely wrong answers on more fact-bound timing and calculation questions. While most of these errors were relatively benign and in response to rather detailed questions, claimants might interpret these responses as indicia that they are ineligible , much as the plaintiff in *Hansen* did. Moreover, a chatbot user seeing this response might be more likely to believe it as a factual statement of law, compared to a user browsing the FAQ page, given how relevant and personalized the other responses by the tool are.

---

199 *See* Ansell at 1037-37; *see* also Hansen, 450 U.S. at 788. Again, this argument seems somewhat weak considering the constitutionality of the FTCA.

200 Fredericks v. C.I.R., 126 F.3d 433 (3rd Cir. 1997).

201 *Id.* at 449.

202 Later court opinions have asserted uncertainty about the Fredericks decision, especially as it was not appealed to the Supreme Court. *See* United States v. Wilson, 2016 WL 3198629 at *7 (distinguishing an estoppel claim from the "clear" facts in *Fredericks* and holding that "the Court can do no more than speculate as to whether the Supreme Court would have upheld Fredericks as the premier circuit decision applying estoppel against the government).

203 Mofatt v. Air Canda, 2024 BCCRT 149 (Feb. 14, 2024); Lisa R. Lifshitz & Roland Hung, *BC Tribunal Confirms Companies Remain Liable for Information Provided by AI Chatbot*, AM. BAR ASS'N (Feb. 29, 2024), https://www.americanbar.org/groups/business_law/resources/business-law-today/2024-february/bc-tribunal-confirms-companies-remain-liable-information-provided-ai-chatbot/.

204 *Hansen*'s holding continues to be regularly cited in appeals of Social Security benefit decisions. *See, e.g.,* Kimm R. W. v. O'Malley, 2024 WL 3896613, *4, (D.Minn July 15, 2024); Holdaway v. Kijakazi, 2023 WL 8007105, *13, (E.D.Tenn. Nov. 17, 2023). However, one might argue that more of the *Fredericks* factors are met in the context of certain benefits chatbots, where the impact on the public fisc is small and where the court found that the chatbot's programming was sufficiently clear as to constitute official authorization of the misrepresentation.

205 *Id. See also* Ronald M. Levin, *Rulemaking and the Guidance Exemption*, 70 ADMIN. L. REV. 263 (2018); Cass R. Sunstein, *"Practically Binding": General Policy Statements and Notice-and-Comment Rulemaking*, 68 ADMIN. L. REV. 491 (2016).

206 *See* Joshua D. Blank & Leigh Osofsky, *Automated Legal Guidance*, 106 CORNELL L. REV. 179 (2020). Blank and Osofsky's statement of causality is largely theoretical, rather than empirical.

The effectiveness of government chatbots for claimants and administrators remains uncertain. A UK Behavioral Insights Team experiment found mixed results: chatbots slightly delayed information discovery but made tasks feel easier and increased trust in AI.[207] Some state administrators report high usage, improved customer experience, and have won awards for chatbots.[208] However, few rigorous evaluations exist on their impact in the U.S., such as speeding up queries, improving take-up rates, reducing errors, or saving staff time for critical tasks.[209]

There may still be some cause for optimism about the impact of government chatbots. Human call center staff—the baseline of comparison—have constantly struggled to balance error rates and wait times. For example, in 1987, the GAO found that the IRS help line provided wrong or incomplete answers 38% of the time.[210] By 2008, the IRS's error rate had fallen to 9%, but with worsening wait time and answer rates.[211] While similar statistics are not available for unemployment call centers, it is widely known that such centers are typically staffed by the newest, least experienced staff as a means of on-the-job training, raising doubts about the accuracy of their guidance.[212] Moreover, call centers introduce their own access problems: unemployment call centers are generally only open from 8 AM to 4PM on weekdays, with some closing as early as 2PM.[213]

Apart from the five generative AI chatbots we examined, most current state chatbots largely function as glorified FAQ pages. As such, the USDA classifies most chatbots as low-risk AI, but warns that they may "lead[] eligible populations to believe they are ineligible for public benefits, provid[e] incorrect program information, or creat[e] barriers to accessing public benefits."[214] Control over information disseminated by chatbots decreases when answers are generatively created, a distinction that USDA notes in categorizing "[c]hatbots using natural language processing to better understand

---

207 THE BEHAVIORAL INSIGHTS TEAM, AI CHATBOTS IN PUBLIC SERVICES (2023), https://www.bi.team/wp-content/uploads/2024/01/External-AI-Chatbot-Trial-For-presentations.pdf.

208 *See* Quinlan, *supra* note 173; MAXIMUS, *supra* note 166 (claiming that a host of tools, including chatbots, significantly reduced call volume, though it is unclear whether these findings are due to users giving up or due to particularly volatile time comparisons during the pandemic). These claims are similar to those made by federal administrators of chatbots. *See* Joshua D. Blank & Leigh Osofsky, *Automated Legal Guidance at Federal Agencies*, Draft Repot for the Administrative Conference of the United States (2022) 15, 20, https://www.acus.gov/sites/default/files/documents/Draft%20Report%20on%20Automated%20Legal%20Guidance%20%283.28.22%29.pdf

209 *But c.f.* Tzhuhao Chen & Mila Gasco-Hernandez, *Uncovering the Results of AI Chatbot Use in the Public Sector: Evidence from US State Governments*, PUB. PERFORMANCE & MGMT. REV. (2024). There does, however, seem to be a positive impact of online claim status tools on call center volumes. For example, North Carolina found that the implementation of a claim status tool was able to deflect a number of calls that would typically go to a call center agent. Justin Lai & Christina Steen, *Reducing Administrative Waste in the Unemployment Insurance System*, U.S. DEP'T OF LABOR (Aug. 17, 2023), https://www.dol.gov/agencies/eta/ui-modernization/promising-practices/admin-waste. However, part of North Carolina's modernization effort included an Interactive Voice Response (IVR) system which routes callers based on their needs; it is possible that the call deflection was caused by users finding it more difficult to use IVR and giving up, rather than successfully getting what they needed from the IVR system. A similar introduction of a claim status tool in New Jersey also found reduction in call center wait times. *Unemployment Insurance Claim Status Tool*, N.J. OFF. OF INNOVATION, https://innovation.nj.gov/projects/ui-claim-tool/.

210 U.S. GOV'T ACCOUNTABILITY OFF., GAO-88-17, TAX ADMINISTRATION: ACCESSIBILITY, TIMELINESS, AND ACCURACY OF IRS' TELEPHONE ASSISTANCE PROGRAM (1987), https://www.gao.gov/assets/ggd-88-17.pdf.

211 INTERNAL REV. SERV., IRS TOLL-FREE TELEPHONE SERVICE IS DECLINING AS TAXPAYER DEMAND FOR TELEPHONE SERVICE IS INCREASING (2009), https://www.irs.gov/pub/tas/msp_1.pdf. The IRS had actually set a goal such that "nearly three out of every ten calls seeking to reach an IRS telephone assistor will not get through, and callers who do receive assistance will first have to wait on hold for an average of nearly 12 minutes."

212 *See* 2022 GAO Report, *supra* note 8 at 15.

213 U.S. DEP'T OF LABOR, UI CALL CENTER STUDY FINAL REPORT (2017), https://oui.doleta.gov/unemploy/docs/CoffeyConsulting_UI_Call_Center_Study_Final_Report_January272017acc.pdf.

214 *Framework for State, Local, Tribal, and Territorial Use of Artificial Intelligence for Public Benefit Administration,* FOOD & NUTRITION SERV. (Apr. 29, 2024), https://www.fns.usda.gov/framework-artificial-intelligence-public-benefit.

user questions, with human-coded, logic-based preset outputs" as an enabling use of AI, but exempting "generative AI responses" from this use.[215]

Most chatbots don't raise merit staffing issues, as providing generic FAQ information isn't inherently governmental, similar to USDA's allowance for non-merit staff in call centers. Yet when generative AI chatbots give personalized responses on eligibility questions—as they almost always did in our evaluation, often saying things like "Based on the earnings information provided, you do not meet the monetary eligibility requirements"—they may raise merit staffing concerns. DOL specifically identifies "[a]dvising a claimant regarding his or her eligibility for benefits based on his or her specific circumstances" as an activity that must be conducted by merit staff.[216] States with generative AI chatbots, hence, have likely crossed that line. On the other hand, without rigorous evaluation, it's unclear how much more helpful merit staff who answer claimant calls—often newly-hired, inexperienced staff[217]—would be compared to a generative AI chatbot.

iii. Wave Three: Adjudication Assistance

The most recent frontier is adjudication assistance: leveraging AI-based tools to help adjudicators make faster and better decisions such as highlighting relevant information or making eligibility predictions. To some, adjudication assistance is a step toward a fully automated benefits eligibility system, which has explicitly been a goal for Medicaid for years and achieved by systems in some states,[218] such as the automatic eligibility determinations provided to Medicaid applicants in California.[219]

State experiments with adjudication assistance for UI and SNAP benefits cover a wide spectrum of both reliance on AI as well as level of automation. In many ways, AI-based adjudication assistance is an outgrowth of robotic process automation (RPA), which automates non-discretionary aspects of claim processing through the use of rule-based bots.[220] For example, states use RPA for tasks like pulling data from one digitized textual source and entering it into another and assembling packets of information/documents, particularly for appeals.[221] In the context of adjudication, this might look like a bot "review[ing] employer fact-finding requests and conduct[ing] data entry tasks to enter the separation reason into the benefits system."[222]

SNAP agencies have been implementing RPA to assist adjudicators, though with limited success. For example, Connecticut worked with Deloitte to implement RPA in 2021 to assist with benefits renewals. The RPA bot effectively "stage[s] the case" for the eligibility worker by reviewing the

---

215 *Id.*
216 UIPL No. 12-01, Change 2, *supra* note 119.
217 *See* 2022 GAO Report, *supra* note 8 at 15.
218 *See* Sarah Grusin, *A Promise Unfulfilled: Automated Medicaid Eligibility Decisions*, NAT'L HEALTH LAW PROGRAM (June 30, 2021), https://healthlaw.org/a-promise-unfulfilled-automated-medicaid-eligibility-decisions/. As of May 1, 2024, 48 states were able to make real-time determinations at application, with those determinations in 28 states being mostly automated. *Real-Time Eligibility Determinations at Application and Ex Parte Renewal Processing*, KAISER FAMILY FOUND., https://www.kff.org/affordable-care-act/state-indicator/real-time-eligibility-determinations-application-exparte/.
219 *Medi-Cal Eligibility & Covered California – FAQ's*, CALIF. DEP'T OF HEALTH CARE SVCS., https://www.dhcs.ca.gov/services/medi-cal/eligibility/Pages/Medi-CalFAQs2014a.aspx.
220 Nikki Zeichner & Amy Perez, *Use Cases for Robotic Process Automation in UI Claims Processing*, U.S. DEP'T OF LABOR (July 11, 2023), https://www.dol.gov/agencies/eta/ui-modernization/promising-practices/rpa-use-cases.
221 *Id.*
222 *Id.*

application to determine if a case is a "change" or "no-change" renewal.[223] If the bot deems the case does not need review, eligibility workers do not need to verify its work.[224] The RPA tool ended up increasing the average days to decision for cases despite the fact that the RPA generally mostly saved time on the simplest cases in a worker's caseload; similar effects were seen after implementation of a recertification bot in Georgia.[225] Officials observed that insufficient training likely contributed to workers' mistrust of the RPA, leading to redundant efforts that undermined the project's success.

State SNAP agencies have recently moved toward more AI-based automation to assist adjudicators.[226] For example, several states use AI-based RPA to automatically process no-change periodic reports from claimants. Georgia, for instance, uses automated tools to "prepare" SNAP recertification applications for employees by identifying mismatches between what is on the application and already in the eligibility system and creating "red flags" for merit workers to review."[227] However, the exact details of the AI technology used in these automation workflows is unclear.

Adjudication assistance—RPA and AI-based alike—has been pivotal for state unemployment agencies clearing their unemployment backlogs during and after the pandemic.[228] In Wisconsin, despite hiring over 1,300 staff,[229] a 770,000-claim backlog persisted in 2021.[230] The state worked with Google to train a machine-learning model on historical data to assign a "confidence score" to applications, "indicating how likely it was that a given application should be approved or forwarded to human agents for possible rejection."[231] The tool only processed "holds" where eligibility couldn't be immediately verified, either removing the hold (granting benefits) or leaving it for manual review.[232] According to the unemployment division secretary, "AI taught itself all of the eligibility rules in the state of Wisconsin by ingesting four million previously fully adjudicated claims that had been

---

223 U.S. DEP'T OF AGRICULTURE, ANALYSIS OF ROBOTIC PROCESS AUTOMATION IN SUPPLEMENTAL NUTRITION ASSISTANCE PROGRAM: THREE CASE STUDIES (2023), https://fns-prod.azureedge.us/sites/default/files/resource-files/snap-bots-rpa-final-report.pdf ("Change renewal cases include revisions to participant information (e.g., employment status, address) from the previous benefit authorization. No-change renewal cases include the same participant information as the prior benefit authorization.").

224 *Id.* Cases that do not require verification may still require eligibility workers to follow up with a client or otherwise resolve discrepancies, though the worker would do so following steps listed out by the RPA.

225 Days to decision is an imperfect measure of time savings as it may be influenced by other outside factors, such as waiting for claimant clarification of information, and is not a direct measure of worker productivity. The lack of effect in Georgia may be offset by reduced error rates (less than a quarter that of the full sample), though this impact is murky due to the fact that the RPA was generally implemented on simpler cases.

226 U.S. Food & Nutrition Serv., *Supplemental Nutrition Assistance Program Advanced Automation Guidance* (Oct. 5, 2023), https://www.fns.usda.gov/snap/advanced-automation.

227 THREE CASE STUDIES REPORT, *supra* note 223.

228 *Id.* The state failed to provide adequate training for workers using the RPA tool, and so some eligibility workers thought the RPA's comments asking for individual review were the result of a "fellow staff member who was shirking their work and not processing cases correctly." *Id.* at 14.

229 Brent Mitchell, *Economic Recovery: Wisconsin Leans on Google Cloud to Better Serve Its Community During COVID-19 and Beyond*, GOOGLE CLOUD BLOG (Aug. 12, 2021), https://cloud.google.com/blog/topics/public-sector/economic-recovery-wisconsin-leans-google-cloud-better-serve-its-community-during-covid-19-and-beyond.

230 Colin Wood, *Google's AI Helped Wisconsin Clear Unemployment Backlog*, STATESCOOP (Aug. 12, 2021), https://statescoop.com/googles-ai-helped-wisconsin-clear-unemployment-backlog/.

231 *Id.* Google's analysis of where claimants were getting stuck in processing also led to the department rewiring the UI claim application process and helped the department identify fraudulent claims. Mitchell, *supra* note 229; Wood, *supra* note 230. It is unclear exactly how sophisticated the model was and is today, though Google describes it as "Artificial intelligence (AI)/ machine learning (ML) for predictive analytics." Mitchell, *supra* note 229.

232 WISCONSIN LEGISL. AUDIT BUREAU, STATE OF WISCONSIN FY 2020-21 FINANCIAL STATEMENTS (2021), https://legis.wisconsin.gov/lab/media/v1znovll/21-23_332655_full.pdf

adjudicated by humans, and it made no mistakes."[233] It remains unclear what "no mistakes" precisely indicates. Regardless, the model—affectionately dubbed Judy the Super Adjudicator—eliminated the backlog in a month and reduced response times to two to three business days.[234] Officials emphasized that no benefits were denied based on the tool.[235]

An audit by Wisconsin's Legislative Audit Bureau, however, revealed issues.[236] First, it was unclear whether the tool met merit staffing requirements, prompting a recommendation to seek written assurance from DOL. Second, the agency lacked proper procedures to monitor the model's accuracy, relying only on initial testing.[237] Weekly projected error rates during the tool's six months of use ranged from 5.5% to 26.8%; however, the tool's role in processing holds and how its error rate compares to the federal DOL BTQ measure (29-33%)-- which includes procedural imperfections, such as inadequate explanation—remain unclear.[238]

Nevada has also heavily utilized AI to address pandemic-related unemployment claim backlogs, leveraging generative AI to gather data on a claim and then "make a recommendation that a human staff member can take into consideration when assessing a case."[239] In partnership with Google, it plans to expand AI use to the appeals process by developing a tool that analyzes appeal hearings transcripts and evidentiary documents to recommend benefit decisions.[240] The tool draws on unemployment law and prior cases, reportedly reducing determination times "from several hours to just five minutes," helping address the appeals backlog.[241] Officials stress that all decisions undergo human review, with referees reviewing recommendations they disagree with and investigating discrepancies.[242]

Automation bias—the tendency to over-rely on automated recommendations—raises concerns about the role of human intermediaries. A former DOL deputy director expressed that it was "a little bit concerning" if "a robot's just handed you a recommendation and you just have to check a box and there's pressure to clear out a backlog."[243] State representatives claim to have run "dozens of tests using the company's technology to analyze hearing transcripts from appeals cases of varying

---

233 Nikki Davidson, *Wisconsin's AI Strategy: Upskilling, Not Displacing Workers*, GOV'T TECH. (Dec. 5, 2023), https://www.govtech.com/artificial-intelligence/wisconsins-ai-strategy-upskilling-not-displacing-workers.
234 Natalie Yahr, *Wisconsin Unemployment System Will Get Yearly Checkups after Overhaul*, THE CAP TIMES (Dec. 20, 2023); https://captimes.com/news/business/wisconsin-unemployment-system-will-get-yearly-checkups-after-overhaul/article_d7fe9278-9ef8-11ee-a224-0bb004cbe310.html.
235 Press Release, Wisconsin Department of Workforce Development, DWD Clears Wisconsin Unemployment Claims Backlog (Dec. 30, 2020), https://dwd.wisconsin.gov/press/2020/201230-unemployment-claim-backlog-cleared.htm.
236 https://legis.wisconsin.gov/lab/media/v1znovll/21-23_332655_full.pdf
237 The Wisconsin Department of Workforce Development responded that during the design phase, the "accuracy of each model's prediction was compared to the actual, manually determined resolution of that hold, and this information was used to calculate an error rate by hold type." WISCONSIN AUDIT, *supra* note 232 at 30. The agency rejected the mandate to conduct ongoing testing, seeming to interpret it as a mandate to retrain the model, something they deemed only necessary when there are "significant changes to the environment" such as law or policy changes.
238 Figures in the report suggest that the tool may not have been responsible for a significant proportion of cases. From December 2020 to June 2021, the tool removed 169,257 holds. Throughout 2020 and 2021, 3.0 million total holds were removed by both the tool and the model. A middle of the road estimate that assigns a quarter of the 3.0 million holds to the tool's active six-month period would suggest the tool is responsible for only 22.5% of the holds removed (and unclear percent of all holds reviewed). *Benefits: Timeliness and Quality Reports*, *supra* note 10.
239 Lauren Kinkade & Noelle Knell, *Nevada Harnesses GenAI for Employment Claims Evaluation,* GOV'T TECH. (Apr. 29, 2024), https://www.govtech.com/artificial-intelligence/nevada-harnesses-genai-for-employment-claims-evaluation.
240 Todd Feathers, *Google's AI Will Help Decide Whether Unemployed Workers Get Benefits*, GIZMODO (Sep. 10, 2024), https://gizmodo.com/googles-ai-will-help-decide-whether-unemployed-workers-get-benefits-200049621.
241 *Id.*
242 *Id.*
243 *Id.*

complexity" before agreeing to a $1 million contract with Google.[244] A governance committee will monitor the system weekly during fine-tuning and quarterly after launch to address hallucinations and bias.[245] Google has developed similar contracts with Michigan and California, as well as contracts for contact center AI-based virtual agents.[246]

Adjudication assistance tools often stand on uncertain ground with merit staffing requirements. Under DOL rules, many state AI tools analyzing facts to determine benefits eligibility—a task classified as inherently governmental—are filling merit staff roles.[247] Some tools even come close to "[a]ctually making a determination of benefit eligibility"—another inherently governmental activity—at least in determinations of benefit approval. In SNAP, AI implementations required FNS approval as a "major change" under regulations,[248] yet it remains unclear if they comply with the Food and Nutrition Act's requirement that "personnel used in the certification process" be merit staff.[249] A loophole might interpret "personnel" to exclude AI, which isn't staff.

***

These examples illustrate a significant trend: states are in many ways at the frontier of experimenting with AI. They are doing so despite considerable regulatory and legal uncertainty, particularly considering requirements that effectively mandate that any material analysis of claimant facts for eligibility be conducted by merit staff. This legal uncertainty, combined with the need to meet federal quality and timeliness mandates on thin and inflexible budgets, may have further pushed states to experiment independently of federal supervision. While initiatives Tiger Teams and USDA's contract for state RPA case studies have facilitated knowledge sharing, these efforts remain limited and vulnerable to funding cuts due to their peripheral roles within federal agencies.

C. The Federal Precursor to Modernization through Evaluation

State experimentation with AI, amid legal uncertainty and limited federal oversight, can be seen as one end of a spectrum of approaches to government modernization. It has allowed for diverse and fast experimentation, but scant independent evaluation has made it difficult to build credibility and ensure due process. At the other end of the spectrum are the reform efforts of the Internal Revenue Service (IRS), perhaps the most closely watched and criticized agency by Congress and the press.[250] The IRS's implementation of phone-based call centers provides a helpful illustration of an approach to modernization that places evaluation at its core.

---

244 *Id.*
245 *Id.* The minutes from the June 2024 meeting of the state's IT advisory board are illuminating: for example, the head of the board asked other advisory board members (who are predominately from private sector technology companies) if there is "a standard practice the private sector uses when acquiring a new software." The state's Chief Information Officer advocated for "continuous evaluation improvement," and acknowledged that if a chatbot gives a constituent "an incorrect but plausible answer, the constituent may take that as the truth which would create a difficult situation for the agency." Nev. Info. Tech. Advisory Bd., Meeting Minutes (June 20, 2024), https://it.nv.gov/Governance/ITAB/Meetings/Meetings/.
246 *Agency Leaders Discover the Power of AI to Scale and Support Citizen Services,* STATESCOOP (Apr. 17, 2024), https://statescoop.com/agency-leaders-discover-the-power-of-ai-to-scale-and-support-citizen-services/.
247 UIPL 12-01, Change 2, *supra* note 119.
248 Advanced Automation in SNAP, *supra* note 137; 7 C.F.R. 272.15.
249 7 C.F.R. § 272.4.
250 *See generally* CHARLES O. ROSSOTTI, MANY UNHAPPY RETURNS (2005) 5-23.

In 1974, the IRS changed the way that millions of Americans interacted with the agency by introducing a telephone tax assistance system, eliminating the need to resolve questions in-person or over the mail.[251] The system quickly scaled to employ thousands of "assistors" across 32 call centers, handling nearly 20 million taxpayer calls annually.[252] Doubt about the efficacy of the program, and its cost, soon followed. In 1978, the House Government Operations Committee asked GAO to evaluate the effectiveness of the new telephone assistance program; specifically "(1) the extent to which taxpayers' telephone calls are answered by IRS, (2) how long it takes for calls to be answered, and (3) the extent to which taxpayers are receiving correct answers to their inquiries."[253] The initial audit, based on a small sample size, provided promising results: about 96% of calls were answered; on average it took one minute to get in contact with an employee; and, about 87% of the IRS's responses were correct.[254]

Over the next decade, the GAO conducted semi-regular audits of the accessibility and accuracy of the program, gradually increasing and improving the quality of tax law questions asked as well as its sampling methodology.[255] While the IRS generally agreed that the GAO's questions were reasonable, it often had critiques of the survey design and methodology.[256] These disagreements spurred the IRS to develop its own internal evaluation system, the Integrated Test Call Survey System (ITCSS), in 1988.[257] The ITCSS aimed to better reflect taxpayers' real-world inquiries and improve benchmarking accuracy. By 1989, the GAO had validated ITCSS reliability, aligning its findings with the IRS's own assessments.[258] This system remains in place today, with improved accuracy rates—by 2023, the IRS reported 91.4% accuracy in responses to tax law inquiries.[259] The IRS stands distinct from DOL and USDA in this regard; while state agencies implementing both programs operate UI and SNAP telephone assistance systems that provide critical clarity to applicants, there is currently still no comprehensive system evaluating the accessibility and accuracy of these systems—it took independent researchers making 2,000 calls to government offices in 2021 to determine that less than 40% of calls to UI agencies reached a live representative.[260]

---

[251] Victor L. Lowe, Director, Gen. Gov't Div., U.S. Gen. Acct. Off., Statement Before the Subcomm. on Commerce, Consumer & Monetary Affairs of the H. Comm. on Gov't Operations: Adequacy of the Internal Revenue Service's Telephone Assistance to Taxpayers 1 (Mar. 22, 1978), https://gao.justia.com/department-of-the-treasury/1978/3/adequacy-of-the-internal-revenue-service-s-telephone-assistance-to-taxpayers-105399/105399-full-report.pdf; U.S. GOV'T ACCOUNTABILITY OFF., GGD-75-69, TELEPHONE ASSISTANCE TO TAXPAYERS CAN BE IMPROVED (1975), https://www.gao.gov/assets/ggd-75-69.pdf. Telephone assistance not only increased convenience for taxpayers, eliminating the need to visit an in-person IRS office, but it also initially allowed the IRS to answer an estimated 33% more people than with in-person service.

[252] U.S. GOV'T ACCOUNTABILITY OFF., TAX SYSTEM MODERNIZATION: FURTHER TESTING OF IRS' AUTOMATED TAXPAYER SERVICE SYSTEMS IS NEEDED (1991), https://www.gao.gov/assets/imtec-91-42.pdf at 2.

[253] Lowe, *supra* note 251.

[254] However, taxpayers referred to a specialist—i.e., those with a more difficult question—only received a correct answer 79% of the time. *Id.*

[255] U.S. GOV'T ACCOUNTABILITY OFF., TAX ADMINISTRATION: ACCESSIBILITY, TIMELINESS, AND ACCURACY OF IRS'S TELEPHONE ASSISTANCE PROGRAM (1989).

[256] *Id. at* 25, 34 (1989).

[257] *Id.* The GAO repeatedly alleged that the "the IRS agreed that [its] questions were reasonable"; however, it also claimed that "no one knows what constitutes a representative set of [evaluation] questions nor are we certain that a representative set of questions can be developed."

[258] U.S. GOV'T ACCOUNTABILITY OFF., TAX ADMINISTRATION: MONITORING THE ACCURACY AND ADMINISTRATION OF IRS' 1989 TEST CALL SURVEY 6 (1990), https://www.gao.gov/assets/ggd-90-36.pdf/

[259] INTERNAL REV. SERV., DATA BOOK (2023), https://www.irs.gov/pub/irs-pdf/p55b.pdf.

[260] Oeindrila Dube, Sendhil Mullainathan, & Devin G. Pope, *A Note on the Level of Customer Support by State*

Crucially, the IRS's evaluation system highlighted the interplay between human and machine systems in service delivery. Early evaluations identified assistor error rates as a significant challenge, prompting the IRS to explore automation as a means of improving accuracy.[261] The IRS piloted two automation projects in 1989: (1) an "automated service system" which replaced assistors' paper reference materials with a computer database that helped automate researching taxpayer questions, and (2) an "expert system" which dynamically guides assistors through questions to ask taxpayers.[262] While preliminary results suggested a 21-percentage point improvement in accuracy with the expert system, subsequent GAO audits flagged flaws in the pilot's design, namely that the IRS "did not consider whether assistors used the Expert System properly, or at all."[263] As a result, OMB denied the IRS's budget request to continue funding the automation experiment, insisting on further testing first.[264] These critiques underscore the complexities of designing, benchmarking, and evaluating systems that integrate human and automated components—issues that remain central to contemporary debates about AI governance. Moreover, IRS's acknowledgement of the shortcomings of the status quo and embrace of automation stand out as an early example of technology being deployed to enhance accuracy rather than merely efficiency.

The IRS's telephone assistance program exemplifies how rigorous evaluation can serve as a foundation for government modernization. It demonstrates that the government can—and does—conduct robust evaluations to improve service delivery. It was spurred by close Congressional scrutiny, which is often triggered by taxpayer complaints, but its continued success also depended on the interplay of statutory support, institutional will, and capable personnel to drive the initiative forward.[265]

The lessons from this case extend to modern challenges in evaluating AI and human-machine interaction. Benchmarking must be an ongoing, iterative process, not a one-time exercise. Such ongoing evaluations are central to programmatic improvements. Evaluations should focus not only on outcomes—such as successful audits or total tax receipts—but also on understanding the dynamics between human users and automated systems. In other words, one cannot understand impact looking at the technical system alone. Finally, systems like the IRS's telephone assistance program show that government agencies are and have long been capable of conducting rigorous, credible evaluations, provided they invest in the necessary structures and expertise. Such systems are indispensable as we consider how to evaluate emerging technologies like AI-driven decision tools, ensuring they align with values of accuracy, fairness, and public trust.

IV. The Legal and Policy Necessity of Evaluation

State experimentation with AI is likely to only grow. The question for policymakers is how to manage the risk of these new technologies—such as benefits falsely flagged as fraudulent or incorrect

Governments: A Mystery-Shopping Approach (U. Chic. Working Paper No. 2021-89, 2021), https://bfi.uchicago.edu/wp-content/uploads/2021/07/BFI_WP_2021-89.pdf. The lack of call center evaluation for UI seems particularly egregious when considering that the UI call volume is at least comparable to, if not greater than, IRS call volume. For example, California's UI agency received millions of calls *per week*—as many as 9.7 million—in 2021; the entire IRS received 47 million calls in 2023 answered 27.3 million of them. IRS DATA BOOK, *supra* note 259; *Call Center Data*, Calif. Employment Development Dep't, https://edd.ca.gov/siteassets/files/pdf/phone-calls-dashboard-052622.xlsx.

261 1991 GAO Report, *supra* note 252 at 2, 12, 14.

262 *Id.* at 3.

263 *Id.* at 4.

264 *Id.* at 8.

265 The 1998 Restructuring Act allowed these new measures to take a more central role in determining employee performance. Key leaders eager to improve customer service, such as Commissioner Rossotti, were likely also critical in the ongoing development of the program. ROSSOTTI, *supra* note 250.

advice about eligibility from a chatbot—while still reaping the benefits of sorely needed innovation in under-resourced agencies. This section examines the legal and policy imperatives for rigorous evaluation of AI systems in benefits administration. We argue that existing legal frameworks, including the Administrative Procedure Act, due process requirements, and procurement regulations, create an implicit mandate for agencies to evaluate AI tools before deployment. We then outline principles for meaningful evaluation and provide an example framework for assessing AI chatbots. Finally, we explore potential avenues for formalizing and implementing a more explicit evaluation mandate.

A. Evaluation of AI Systems is Legally Mandated

i. Adopting AI without Evaluation is Likely a Violation of Procedural Due Process

The Due Process Clause demands that government agencies evaluate AI systems that meaningfully affect receipt and termination of benefits before deployment to avoid unconstitutional deprivations of benefits. Terminating public benefits without proper procedural safeguards violates constitutionally protected property interests.[266] When agencies adopt AI systems without rigorous evaluation, they risk systematic violations of these due process protections.

The *Mathews v. Eldridge* framework requires courts to balance three factors when determining adequate procedural safeguards: (1) the private interest affected; (2) the risk of erroneous deprivation and value of additional safeguards; and (3) the government's interest, including administrative costs.[267] The second factor—the risk of erroneous deprivation—is particularly critical for AI adoption. Without evaluation, agencies cannot quantify error rates, false positives in fraud detection algorithms, or the accuracy of chatbot-provided eligibility information. This lack of baseline measurement makes meaningful *Mathews* balancing impossible and endangers due process by imposing unknown and unmeasured risks of wrongful benefit denials.

Evaluation serves due process in broader ways than merely providing evidence-based approaches to *Mathews* balancing. The evaluation process itself should help prevent the adoption of systems that increase error rates and thus the risk of erroneous deprivation. Jerry Mashaw's "managerial" conception of due process reinforces this evaluation imperative by emphasizing systemic quality assurance over individual procedural remedies in court.[268] The broader analytic scope of evaluation appears more likely to achieve due process's mandate of systemic accuracy than reliance on individual-level exercise of procedural rights through the appeals process, which likely does little to promote systemic accuracy.[269]

Due process has already figured prominently in the most significant cases challenging agency's attempts or failures to modernize benefit systems, but these cases have stopped short of establishing system-level mandates. In the sprawling federal case against the Michigan unemployment agency and contractors for the failure of the state's automated fraud detection system, residents alleged a due process violation for termination of payments without adequate notice and hearing.[270] While the due process claims survived a motion to dismiss, plaintiffs' class certification failed due to their varied

---

[266] Goldberg v. Kelly, 397 U.S. 254, 264-65 (1970).
[267] 96 U.S. 319.
[268] Jerry L. Mashaw, *The Management Side of Due Process,* 59 CORNELL L. REV. 772 (1974). Danielle Citron also develops a notion of inquisitorial quality assurance in her work. Citron, *supra* note 17.
[269] David Ames, Cassandra Handan-Nader, Daniel E. Ho & David Marcus, Due Process and Mass Adjudication: Crisis and Reform, 72 STAN. L. REV 1, 23 (2020); Jerry L. Mashaw, The Management Side of Due Process, 59 CORNELL L. REV. 772, 785 (1974).
[270] Cahoo v. SAS Analytics, 912 F.3d 887 (6th Cir. 2019). Plaintiffs also brought state law claims relating to product reliability and gross negligence. These counts were dismissed on account of inadequate tortious damage to the plaintiffs.

positions of deprivation, leaving just four plaintiffs to continue with the case. The 6th Circuit ultimately held that the unemployment agency supervisor defendants were entitled to qualified immunity because they did not trespass on clearly established law.[271] The state case, however, faced fewer class certification hurdles.[272] The Michigan court of appeals ultimately found that the agency's use of untested automated decision-making violated the state constitution's due process clause.[273]

Similarly, plaintiffs in Missouri brought a case against the state SNAP agency, alleging that the agency's wrongful denials of their SNAP benefits for failure to interview violated due process.[274] The severely understaffed system effectively made it impossible to schedule an interview. The district court ordered injunctive relief in the form of monitoring the state SNAP agency's performance improvements, particularly around its ability to offer interviews and answer calls in a timely fashion. In some ways, the Missouri case is the flipside of the Michigan case. Both agencies were overwhelmed and understaffed. Michigan's unemployment agency turned to a form of auto-adjudication and found itself in violation of due process for its arbitrary assignments of fraud and inadequate notice. The Missouri agency did not try to modernize, and instead stuck to its status quo of understaffed services; yet it too found itself in the same due process hole for arbitrary denials.

These cases underscore the idea that true procedural due process requires systemic evaluation before full-scale deployment rather than the post-hoc remedies made possible through litigation. Left to generalist courts that may know little about program administration and evaluation design, litigation is not guaranteed to focus on and improve the agency-level decisions that matter most for the average claimant.[275] Moreover, *Mathews* balancing is an inherently case-level, fact-bound exercise. As Danielle Citron has observed, this case-level application misses the fact that AI-based tools are designed to operate at scale, and that a small change to a tool because of procedural scrutiny may impact the

---

[271] This is partly due to the fact that the court was restricted to review the more subtle due process violation that was applicable to the four remaining plaintiffs. Those plaintiffs—unlike *Goldberg* and some of those in the original class—did *not* lose ongoing benefits. Rather, they received benefits and were later "auto-adjudicate[d] . . . guilty of fraud, disqualifying them for future benefits" and clawing back the previous benefits with penalty. They received notice of this liability. Ultimate, the court held: "The content of the two notices, the opportunity for a hearing, and the months- or years-long delay before the plaintiff faced a deprivation distinguish this case from existing due process precedent." *Id.* at 407.

[272] The lead plaintiff had received unemployment benefits and was later determined to have been overpaid due to suspected fraud, and in penalty his tax refund would be intercepted and his wages would be garnished. Bauserman v. Unemployment Insurance Agency, 2017 WL 3044120 (Ct. App. Mich. 2017).

[273] Bauserman v. Unemployment Insurance Agency, 509 Mich. 673 (Sup. Mich. 2022). Status reports warned that the MiDAS system could not read large chunks of converted legacy data. Paul Egan, *Michigan Integrated Data Automated System Experiences 93 Percent Error Rate During Nearly Two Years of Operation*, GOVTECH (July 31, 2017), https://www.govtech.com/data/michigan-integrated-data-automated-system-experiences-93-percent-error-rate-during-nearly-two-years-of-operation.html. Moreover, the state auditor's report revealed that the agency did not fully analyze and validate data to help identify payments requiring further review. MICH. AUDITOR GENERAL, MICHIGAN INTEGRATED DATA AUTOMATED SYSTEM (MIDAS) 20 (2016), https://audgen.michigan.gov/finalpdfs/15_16/r641059315.pdf.

[274] For example, one resource center at which applicants are expected to interview "has a line out the door and over a two-hour wait because it does not have sufficient staffing and/or hours." *Id.*

[275] As Adrian Vermeule has argued, "The federal judicial system is not set up, not equipped, to engage in a sustained course of synoptic institutional engineering." ADRIAN VERMEULE, LAW'S ABNEGATION: FROM LAW'S EMPIRE TO THE ADMINISTRATIVE STATE 115 (2016). Moreover, due to its rather sparse constitutional basis, the exact scope and requirements of due process are subject to inconsistent judicial interpretation. For these and other reasons, Jerry Mashaw's managerial turn in due process instead places agency's internal administrative law at the center of quality assurance, rather than externalist judicial interpretations of constitutional due process. *See generally* JERRY L. MASHAW, BUREAUCRATIC JUSTICE: MANAGING SOCIAL SECURITY DISABILITY CLAIMS (1983).

thousands or millions of cases to which the tool is applied.[276] The Michigan cases, particularly the federal case, ultimately turned on the exact wording sent out in various notices, highlighting the judiciary's tendency to focus on the notice aspect of due process to the exclusion of analysis of the actual fraud detection system's design, when and why it was used, and how it was supervised.[277] While improved notice and appeal processes are likely fine outcome incentives, due process as it was litigated in Michigan seemed to miss the more important questions: how could the agency have designed a better fraud detection system? What critical conversations with, oversight of, or incentives for contractors were missed?

Moreover, due process, as cast by *Goldberg* and *Mathews* and as litigated to this day, often places accuracy as the sole cornerstone of adjudicative quality. As Mashaw argued, the governing caselaw's narrow focus on accuracy is "unresponsive to the full range of concerns embodied in the due process clause," such as dignity, transparency, and social trust.[278] Decisional accuracy should be but one outcome of interest in agency evaluation plans. Ideally enforcement of an evaluation mandate would take a broader scope to encourage the consideration of multiple outcomes including accuracy and customer satisfaction.

Taking due process seriously means taking evaluation seriously. For AI adoption specifically, this means agencies must test error rates, validate decision-making processes, and ensure adequate human oversight mechanisms exist. Without such evaluation, agencies cannot satisfy *Mathews* balancing, cannot ensure systemic accuracy, and cannot avoid the constitutional violations that plagued both Michigan's automated system and Missouri's overwhelmed manual processes.

ii. Adopting AI without Evaluation is Likely Arbitrary and Capricious

The APA and state analogues provide another potential foundation for an evaluation mandate. The APA forbids agency action that is "arbitrary, capricious, an abuse of discretion, or otherwise not in accordance with law."[279] Agency actions are arbitrary and capricious if they "entirely fail[] to consider an important aspect of the problem" or "offer[] an explanation . . . contrary to the evidence before the agency."[280] An agency must "examine the relevant data and articulate a satisfactory explanation for its action including a 'rational connection between the facts found and the choice made.'"[281]

Agencies deploying AI systems without evaluation appropriately scaled to risk fail to consider relevant factors such as accuracy rates across populations, disparate impact on protected classes, consequences of errors, and implementation costs versus benefits.[282] Without evaluation, agencies cannot demonstrate rational connections between AI deployment and policy objectives or show that benefits outweigh risks. However, courts are generally hesitant to affirmatively require agencies to

---

[276] Citron, *supra* note 19 at 1249; David Engstrom & Daniel E. Ho, *Algorithmic Accountability in the Administrative State*, 37 YALE J. REG. 800, 827 (2020).

[277] *See* Cahoo v. SAS Institute, 71 F.th 401, 412 (6th Cir. 2023):

> Yes, MiDAS's logic trees spawned internal, interim fraud findings that marked individuals for further review. But that internal step did not deprive a claimant of property. As noted, those property deprivations came months or years later, following notices of deprivation and a multi-level appeal process. The missing link, then, is a case that clearly established the inadequacy of the questionnaires, notices of determination, and appeal processes available in the run-up to the property deprivation.

[278] Mashaw, *supra* note **Error! Bookmark not defined.**at 30.

[279] 5 U.S.C. § 706(2)(A).

[280] State Farm, 463 U.S. 29, 43 (1983).

[281] *Id.* (quoting Burlington Truck Lines v. United States, 371 U.S. 156, 168 (1962)).

[282] It is unclear, of course, as to whether disparate impact claims can sound in the APA. *See* Cristina Isabel Ceballos, David Freeman Engstrom & Daniel E. Ho, *Disparate Limbo: How Administrative Law Erased Antidiscrimination*

conduct evaluations under the APA, and for good reason.[283] Notably, the D.C. Circuit struck down the FTC's controversial rule requiring two randomized clinical trials (RCTs) to substantiate disease claims in order to comply with the FTC Act's deceptive advertising provisions.[284] The court criticized the rigidity of the two-study requirement, which appeared to ignore the Commission's own acknowledgement that "the quality of studies will be more important than quantity."[285] However, the court found no fault with a singular RCT requirement, noting that "some RCT substantiation for disease claims directly advances, and is not more extensive than necessary to serve, the interest in preventing misleading commercial speech."[286] We certainly do not advocate that full RCTs are always—or even often—required under the APA's arbitrary and capricious standard.[287] As we detail in Part II.B, evaluation requirements should be context-specific and proportionate to the stakes involved, unlike the FTC's categorical mandate.

A broad APA-based evaluation mandate risks becoming the very kind of process-heavy impediment to innovation that has hampered government technology adoption for decades. The danger is real and historically documented: if every AI integration requires exhaustive pre-deployment analysis, agencies may simply avoid beneficial automation altogether, defaulting to dysfunctional status quo systems that themselves violate due process, as we saw in Missouri's SNAP case. Critics might reasonably ask why AI systems should trigger evaluation requirements when the human systems they replace often lack such systematic assessment. This asymmetry problem has real force: state UI call centers, for instance, operate without the kind of accuracy monitoring that the IRS conducts, despite providing equally critical guidance to claimants. There's a legitimate worry that imposing evaluation requirements only on AI deployments puts a thumb on the scale in favor of inadequate status quo systems, essentially penalizing agencies for attempting to modernize.

The solution lies in recognizing that "evaluation" encompasses a spectrum of approaches rather than a monolithic requirement. At one end are simple benchmark tests—measuring whether an AI system meets basic performance metrics like accuracy scores on standardized datasets. In the middle

---

283 Imposing such a blanket restriction could arguably run afoul of *Vermont Yankee*. Courts are, however, comfortable finding agency action arbitrary and capricious when the agency ignores relevant studies evaluating a policy's impact. *See, e.g.,* WildEarth Guardians v. United States Bureau of Land Management, 870 F.3d 1222 (10th Cir. 2017) (conclusions "unsupported by hard data" do not provide "information sufficient to permit a reasoned choice" and are therefore arbitrary and capricious); Genuine Parts Company v. Environmental Protection Agency, 2018 WL 2271086 (D.C. Cir. 2018) (arbitrary and capricious to rely on portions of studies that supported its position while ignoring cross sections in those studies that did not support the position); WildEarth Guardians v. Haaland, 2021 WL 4263831 (C.D. Cal. 2021) (holding it was arbitrary to ignore study projections of habitat impact). *But cf.* Nantucket Residents Against Turbines v. U.S. Bureau of Ocean Energy Management, 100 F.4th 1 (1st Cir. 2024) (finding an agency's decision to ignore studies on whale risks when approving an offshore wind energy project in compliance with the APA).

Perhaps one of the most notable cases in this vein is *Business Roundtable*, in which the D.C. Circuit struck down a new SEC rule on shareholders for failure to first evaluate the likely consequences of the rule. Here, the evaluation mandate was partly driven by the APA, but in large part also driven by an SEC-specific statutory obligation to consider the effects of new rules on "efficiency, competition and capital formation"; however, a similar argument could be made for the statutes undergirding benefits systems that require agencies to make timely and accurate decisions.

284 *POM Wonderful, LLC v. F.T.C.,* 777 F.3d 478 (D.C. Cir. 2015).

285 *Id.*

286 *Id.* The court also acknowledges the appropriateness of "competent and reliable scientific evidence"—which may not necessarily be an RCT—to substantiate advertising claims. Notably, the court's analysis was conducted under *Central Hudson* scrutiny for First Amendment violations, not the APA. However, the specific prong of analysis that the court focused on—that the government's restriction on speech "directly advances the governmental interest" and "is not more extensive than is necessary to serve that interest"—is highly analogous to arbitrary and capricious review under *State Farm.*

287 *But see* Michael Abranowicz, Ian Ayres, & Yair Listokin, *Randomizing Law* 159 U. Pa. L. Rev. 929 (2011) (advocating for broader use of RCTs in policymaking and arguing that RCTs are compatible with the APA despite their necessarily "arbitrary" nature).

are sandbox deployments and limited pilots that test human-AI interaction in controlled settings without full deployment. At the other end are comprehensive randomized controlled trials and ongoing quality assurance systems like those Jerry Mashaw advocated. The APA's arbitrary and capricious standard should require agencies to select evaluation methods proportionate to the risks and stakes involved, not demand the most rigorous possible assessment in every circumstance, following the calibrated balancing demanded in the due process context by *Mathews*.

To specifically address the asymmetry issue, any evaluation requirement read from the APA should emphasize ongoing assessment rather than front-loaded pre-deployment testing. The Biden administration's OMB memo exemplified the dangers of process-heavy pre-deployment requirements that may bear little relationship to real-world performance.[288] A chatbot that performs well in laboratory testing may fail when confronted with the creative and contextual questions that real claimants ask. Conversely, an adjudication assistance tool that seems concerning in theory may prove more accurate and fair than overwhelmed human adjudicators in practice. Pre-deployment evaluation, while valuable, can only go so far in predicting how AI systems will perform when they encounter the full complexity of real-world benefit administration.

To prevent the APA evaluation requirement from becoming a source of administrative paralysis, several limiting principles should apply. First, agencies should have safe harbors for pilot programs and limited deployments that affect small numbers of claimants and include robust fallback procedures. The mere act of testing an AI system on a subset of cases should not trigger the full apparatus of pre-deployment evaluation, particularly when agencies can demonstrate adequate safeguards and monitoring. Second, agencies that can demonstrate they lack adequate human baselines—such as those with call centers that don't track accuracy rates—should have flexibility to establish baseline measurements concurrent with AI deployment rather than requiring pre-existing benchmarks that may not exist. Third, the APA standard should focus on the reasonableness of an agency's evaluation approach rather than demanding specific methodologies. An agency that conducts basic accuracy testing and implements ongoing monitoring should satisfy the arbitrary and capricious standard, even if it forgoes more sophisticated evaluation methods that might be appropriate for higher-risk deployments. The goal is ensuring reasoned decision-making based on evidence, not mandating particular analytical techniques.

iii. Adopting AI without Evaluation is Likely Contrary to Procurement Law and Policy

One last implicit legal imperative for evaluation lies in procurement requirements. As our case studies illustrate, most government AI systems are not developed in-house. They're purchased. As such, these purchases are subject to procurement law.

Several elements of the Federal Acquisition Regulation (FAR) militate in favor of pre-purchase evaluation of AI systems wherever possible.[289] FAR Part 7 requires written acquisition plans to "describe the test program" for each major phase of system acquisition, ensuring agencies formally plan pilot tests before full implementation.[290] FAR Subpart 9.3 allows agencies to require "first article testing" when acquiring new or unproven products—precisely what AI systems represent—

[288] Daniel E. Ho et al., Comment Letter on OMB–2023–0020, Proposed Memorandum for the Heads of Executive Departments and Agencies: "Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence" (Dec. 4, 2023), https://dho.stanford.edu/wp-content/uploads/OMB_Open_Innovation.pdf.
[289] The FAR itself does not directly apply to states, but many state procurement codes are modeled after FAR.
[290] FAR Subpart 7.015. Per FAR Subpart 7.012, acquisition planning applies to "all acquisitions."

allowing agencies to require initial prototypes to be tested against performance criteria before committing to full deployment.[291]

Specifically for IT contracting, FAR Part 39 encourages agencies to apply "continuous collection and evaluation of risk-based data" and "prototyping prior to implementation," with AI systems representing quintessential high-risk IT requiring such evaluation.[292] This Part also mandates modular contracting for IT acquisitions "to the maximum extent practicable," requiring "delivery, implementation, and testing of workable systems in discrete increments," directly supporting pilot deployments and phased testing of AI capabilities.[293] More broadly, FAR Part 46 requires that agencies conduct inspection and testing to ensure "supplies or services tendered by contractors meet contract requirements" before acceptance, empowering agencies to subject AI systems to rigorous acceptance testing for accuracy, security, and bias.[294]

Beyond the FAR, the Federal Information Technology Acquisition Reform Act requires non-Defense agencies to obtain approval by their Chief Information Officer (CIO) prior to entering any IT contract.[295] For high-risk IT investments, CIOs must identify the causes of risk, the extent to which these causes and be addressed, and the "probability of future success," actions that are somewhat meaningless for AI procurement without comparative evaluation against baselines.

The Office of Management and Budget's interpretation of acquisition principles goes even further in creating an obligation to evaluate AI tools pre-deployment. In seeking to "acquire the best solutions at lower cost to the taxpayer," OMB counsels agencies to "pay careful attention to vendor sourcing, data portability, and long-term interoperability to avoid significant and costly dependence on a single vendor."[296] Testing and evaluation is situated as the first step in selecting an AI vendor: the memo requires agencies to "test proposed solutions to understand the capabilities and limitations of any offered AI system," including by creating a "testing environment" on government systems where possible.[297] Moreover, OMB requires agencies to contractually allow for regular "monitor[ing] and evalut[ion]" of "performance, risks, and effectiveness" of the applicable AI system.[298] These procurement-specific guidelines echo OMB's general guidance on federal AI use, which counsels agencies to "conduct[] ongoing testing and validation on AI model performance . . . including by testing in real-world conditions" and "consider[] contractual terms that prioritize the continuous improvement, performance monitoring, and evaluation of effectiveness of procured AI."[299]

States have also released guidance around generative AI that explicitly mandates evaluation. For example, California requires agencies to "prepare data inputs and test models adequately" before deploying a solution in order to "experiment," "gather feedback," and "correct outcomes to reduce

---

291 FAR Subpart 9.302-9.303.
292 FAR Subpart 39.102.
293 FAR Subpart 39.103.
294 FAR Subpart 46.102.
295 40 U.S.C. § 11319.
296 Executive Office of the President, Office of Management and Budget, Memorandum M-25-22: Driving Efficient Acquisition of Artificial Intelligence in Government, at 1 (Apr. 3, 2025), https://www.whitehouse.gov/wp-content/uploads/2025/02/M-25-22-Driving-Efficient-Acquisition-of-Artificial-Intelligence-in-Government.pdf.
297 *Id.*
298 *Id.* The memo goes even further in encouraging agencies to act based on regular monitoring of AI system performance: for example, if a new system version fails to meet performance standards, agencies should require vendors to "roll-back to a previous version."
299 Executive Office of the President, Office of Management and Budget, Memorandum M-25-21: Accelerating Federal Use of AI through Innovation, Governance, and Public Trust, at 1 (Feb. 1, 2025), https://www.whitehouse.gov/wp-content/uploads/2025/02/M-25-21-Accelerating-Federal-Use-of-AI-through-Innovation-Governance-and-Public-Trust.pdf.

bias and inaccurate information."[300] For IT services contracts over $500,000, California's procurement code requires departments to perform a "post evaluation"; however, the required form merely asks (with answers in yes or no checkboxes) whether the contracted work was completed on time, within budget, and whether it "fulfill[ed[all other requirements of the contract including quality standards."[301] New York specifically calls out use of an AI "without thorough testing to confirm accuracy" as "unacceptable."[302] However, these state guidance documents are often less comprehensive than those at the federal level,[303] as are the state procurement codes underlying them.[304]

B. What Meaningful Evaluation Looks Like

Statutory and constitutional principles militate for pre-adoption evaluation of AI systems in benefits. But what does meaningful evaluation look like in practice? Drawing from the IRS's decades-long experience with systematic evaluation of its telephone assistance program and emerging best practices in AI governance, several core principles should guide agencies' evaluation efforts.

i. Core Principles

Scale evaluation rigor with potential impact. Not all AI applications require the same level of scrutiny. The spectrum of AI integration demands graduated evaluation approaches that match the level of oversight to the degree of automation and potential harm.[305] This requires agencies to draw from an evaluation "toolkit" that encompasses different methodologies depending on the intervention's complexity and risk profile. Simple OCR systems that help digitize documents may require only basic accuracy testing using benchmark datasets (such as F1 scores on test sets) and periodic spot-checks through random sampling. Chatbots providing eligibility guidance merit comprehensive evaluation including sandbox testing with realistic user scenarios, accuracy assessment across different query

---

300 CALIF. GOV'T OPERATIONS AGENCY, GENAI GUIDELINES (Mar. 2024), https://www.govops.ca.gov/wp-content/uploads/sites/11/2024/03/3.a-GenAI-Guidelines.pdf.

301 California Procurement Code § 12102.3; Cal. Dep't of Gen. Servs., Procurement Div., State Contracting Manual § 21-6: Post Evaluation for IT Services Contracts (Procurement Div.), https://www.dgs.ca.gov/PD/Resources/SCM/TOC/21/21-6; Cal. Dep't of Gen. Servs. & Cal. Dep't of Technology, Post Evaluation for IT Services Contracts, Std. Form 971 (Jan. 2020), https://www.documents.dgs.ca.gov/dgs/fmc/pdf/std971.pdf.

302 N.Y. Off. of Info. Tech. Servs., NYS Policy P24-001: Acceptable Use of Artificial Intelligence Technologies (May 2025), https://its.ny.gov/system/files/documents/2025/05/nys-p24-001-acceptable-use-of-artificial-intelligence-technologies.pdf.

303 For example, the only mention of testing or evaluation in New York's AI IT policy comes in a list of unacceptable examples. *Id.*

304 While state procurement codes are generally modeled on FAR as well as the ABA's Model Procurement Code, their testing and evaluation requirements are somewhat less explicit. For example, state procurement codes often have a "responsible bidder provision," which requires vendors to demonstrate "trustworthiness, as well as quality, fitness, capacity, and experience to satisfactorily perform," which could be construed to require meaningful AI evaluation in order to make these determinations. CA Public Contract Code § 1103. Or, California requires its Department of Technology to oversee IT projects by "[e]valuating" them "based on the business case justification, resources requirements, proposed technical solution, project management, oversight and risk management approach, and compliance with statewide strategies, policies, and procedures." CA Public Contract Code § 11546. However, none of these provisions comes close to creating a clear textual mandate for continuous evaluation.

305 *See* Daniel E. Ho, Olivia Martin, Kit Rodolfa, Amy Perez & Faiz Surani, *The Spectrum of AI Integration: The Case of Benefits Adjudication*, *in* AI: LEGAL ISSUES, POLICY & PRACTICAL STRATEGIES (2024).

types and user populations, and structured interactions with claims examiners to understand how the tool affects workflow. Auto-adjudication systems demand the most rigorous evaluation protocols from the toolkit: comprehensive bias testing, field experiments or randomized controlled trials to measure real-world performance, ongoing accuracy monitoring through quality assurance frameworks (following models like Mashaw or GPRA's approach to performance measurement) and regular human review of algorithmic decisions compared to established benchmarks. Agencies must mix and match evaluation tools based on the specific intervention, ensuring the evaluation approach is proportionate to both the degree of automation and the potential consequences of system failure. This graduated, toolkit-based approach ideally prevents bureaucratic paralysis while ensuring adequate protection where stakes are highest.

Establish human baselines wherever possible. A critical component of meaningful evaluation is comparing AI performance against the existing human-operated system, not against abstract benchmarks or vendor promises. As our chatbot audit revealed, a tool may appear to work well in isolation but perform poorly relative to human call center staff, or vice versa. For benefits agencies, this means measuring current human adjudicator accuracy rates, call center response times, and case processing speeds before implementing AI tools, and conducting parallel testing to determine whether AI improves, maintains, or degrades these metrics.

Measure what matters beyond accuracy. While decisional accuracy is paramount, particularly given *Mathews*' focus on minimizing erroneous deprivation, evaluation should encompass the full range of values that procedural due process seeks to protect where possible. The IRS evaluates not just accuracy of telephone advice, but also timeliness, accessibility, and customer satisfaction. Benefits agencies, where feasible, should similarly track whether AI tools improve or hinder claimants' ability to navigate the system, understand their rights, and receive timely decisions. This includes measuring downstream effects: does a chatbot that provides technically accurate but overly complex information actually help claimants complete applications successfully? Do fraud detection systems reduce improper payments without creating insurmountable barriers for legitimate claimants?

Emphasize continuous monitoring over one-time testing. Pre-deployment testing, however rigorous, cannot always predict long-term performance. AI systems degrade over time due to data drift, changing legal requirements and models, and evolving user needs. The IRS's ongoing evaluation system exemplifies this principle—rather than relying on initial validation, the agency continuously measures and reports performance metrics. Benefits agencies must similarly embed evaluation into their operations, conducing regular audits of AI outputs and maintaining human expertise to assess system performance even as automated tools handle increasing caseloads.

These principles recognize both the potential of AI to improve benefits administration and the constitutional imperative to ensure that technological modernization enhances rather than undermines procedural fairness. Meaningful evaluation provides the bridge between these goals, enabling agencies to harness AI's benefits while maintaining the accuracy, timeliness, and dignity that due process demands.

i. Example: An Evidence-Based Adoption Cycle for AI Chatbots

The first step to adopting an AI tool is collecting data to establish a status-quo, often human, baseline.[306] For evaluating a chatbot, this could involve measuring the number of calls received per

[306] Program statutes, such as the Food and Nutrition Act, do create baseline evaluation requirements, as discussed *supra* Part I.A. The Government Performance and Results Act, and its 2010 update, also require that agencies collect and report data related to their "performance" goals, which connect to the agency-designed strategic plan. Modernization

week, call topics, call duration, and client satisfaction for a sample of calls. It is also essential to assess the accuracy of advice given by human agents. For example, the IRS today evaluates client touch points for customer accuracy ("giving the correct answer with the correct resolution"), regulatory accuracy, procedural accuracy, professionalism, and timeliness.[307] The IRS records calls and reviews a random sample, reporting that 91.4% of tax law answers and 89.2% of account answers (such as questions about balance due) were accurate from 1.6 million calls in 2023.[308] The IRS is distinct from the DOL in conducting such frequent and thorough evaluations of its call centers. Unlike the IRS, the DOL does not mandate regular evaluations of unemployment insurance call centers, focusing instead on states' timeliness and accuracy in determining benefits, leaving call center evaluations optional for state administrators.[309] Ideally, agencies would also track downstream outcomes such as application completion, grant rates, and appeal rates to understand claimants' full experience.

The next step is to identify an appropriate tool based on which elements of the human baseline are in greatest need of improvement and through user consultation. For example, if most calls are status inquiries, a simplified online portal might be a better investment than a generative AI chatbot. Once a tool is chosen, agencies should run a pilot, testing it on a small, representative user subset.[310] Data should be collected on the number of chats initiated, topic of the chat, time per chat, and if available, a measure of client satisfaction. As with the human baseline, a random sample of chatbot responses should be reviewed for accuracy. During the pilot, data collection for the human call center should continue.

Agencies can then compare chatbot performance with human benchmarks but should avoid hasty conclusions. Leveraging the full scope of data is critical to prevent overly simplistic comparisons. We outline examples of how careful analysis of chatbot impact on call volume and assistance accuracy can provide insights that might lead to more informed decisions.

Call Volume. A common goal of chatbots is to reduce burdens on overwhelmed call centers, especially during economic downturns, when high call volumes lead to long wait times and poor claimant experiences. Some states report declines in call center volume after introducing chatbots, but these comparisons are often skewed by background pandemic-induced time trends, as many chatbots were launched during peak demand.[311] Future comparisons should adjust for factors like claim volume, seasonality, or broader time trends.

Even with adjustments, agencies should consider the possibility of a "net widening effect."[312] For example, in a pre-chatbot world, if 10,000 individuals called a center each month and chatbots perfectly substituted for calls, we would expect call center volume to drop and chatbot usage to rise equivalently. However, in reality, not everyone who needs help calls in due to barriers like limited hours, language issues, or stigma. In this scenario, there might be 15,000 people with questions, but

---

Act, Pub. L. 111-352, 124 Stat. 3866 (2011). For further discussion of GPRA's requirements, see *infra* Part IV.B. However, as this Article has shown, agencies are not always conducting performance evaluation on key areas that affect their interactions with citizens, such as the quality of their call centers.

307 Internal Revenue Manual § 21.10.1 (Dec. 20, 2022), https://www.irs.gov/irm/part21/irm_21-010-001; DATA BOOK, *supra* note 259 at 24.

308 *Id.*

309 *Id.*

310 Most chatbots surveyed did not appear to have run limited pilots, though they have been run and not publicly announced. The Indiana chatbot states that it is in "beta release" but there is no indication that it was first launched at a smaller scale.

311 Davidson, *supra* note **Error! Bookmark not defined.**.

312 Net widening, often discussed in criminal justice, refers to policies like probation that aim to reduce juveniles in the system but instead expand the number of individuals under its control. *See, e.g.,* Scott H. Decker, *A Systemic Analysis of Diversion: Net Widening and Beyond*, 13 J. CRIM. JUST. 207 (1985).

only 10,000 callers. Chatbots could attract both those who previously called and those who did not, creating broader reach.

An agency observing no change in call center volume after launching a chatbot shouldn't assume failure. The chatbot may have expanded access, reaching claimants who were previously unserved. While this net widening effect may not align with the agency's original goal, it can support broader objectives like increasing eligible applications or reducing follow-up work by clarifying initial questions.[313]

Accuracy. No public data currently exists on the accuracy rates of chatbots or human call centers serving SNAP or UI claimants. However, as this Article's audit shows, chatbots can err. While collecting accuracy data and comparing chatbots to a human baseline is essential, evaluations shouldn't stop there. People may ask different types of questions to chatbots and humans; for example, users might feel freer to ask complex or bizarre questions to chatbots, or reserve such questions for human agents. To make meaningful comparisons, agencies should adjust for differences in query type and complexity. A good starting point could mirror the IRS's segmentation of account-based (e.g., claim status) and eligibility (e.g., complex policy) questions to avoid misjudging relative competencies, though data collection for the human baseline will have to mirror this stratification.

If a chatbot shows a higher error rate after adjusting for query composition, agencies should consider factors beyond their institutional risk tolerance. For example, a higher error rate might become more concerning if chatbots handle a significant share of overall volume, including new users, increasing the absolute impact of errors.[314] The level of trust users place in chatbot advice is another key factor. Leigh and Osofsky suggest that personalized, instant, and unqualified guidance—like that from chatbots—may lead to higher reliance compared to human agents, though this remains empirically uncertain.[315] Trust likely varies by demographic, making it crucial for agencies to conduct their own studies, such as user questionnaires during pilot testing.[316] If users indeed trust chatbot advice significantly more than human representatives, a higher chatbot error rate would be far less acceptable.

User Satisfaction. While evaluating volume and accuracy rates may cover a narrow conception of the due process balancing implicit in *Mathews*, agencies can and should evaluate the impact of a tool on other core constitutional values like dignity. Although these might seem like lofty goals, carefully designed user surveys and focus groups could help further agencies' understanding of how an AI tool impacts claimant's conception of the program's accessibility, clarity, and empathetic communication.

***

---

313 Another potential result is more informed claimants such that the need for advice might generally go down.

314 However, there remains the question of whether interacting with the higher error rate chatbot is still better for new claimants than acting with no information at all.

315 Leigh & Osofsky, *supra* note 206 at 221.

316 *See, e.g.*, Tae Hyun Baek & Minseong Kim, *Is ChatGPT Scary Good? How User Motivations Affect Creepiness and Trust in Generative Artificial Intelligence*, 83 TELEMATICS & INFORMATICS (2023) (finding that over 80% of those surveyed trusted ChatGPT and that trust was correlated with perceived personalization); Amoozadeh et al., *Trust in Generative AI Among Students: An Exploratory Study,* PROC. ACM HUM.-COMPUT. INTERACTION (2023) (surveying university students and finding high variance in levels of trust in generative AI). Trust in chatbots has also been a frequent feature of consumer polls. *See, e.g.*, *Poll: Most Who Use Artificial Intelligence Doubt AI Chatbots Provide Accurate Health Information*, KAISER FAMILY FOUND. (Aug. 15, 2024), https://www.kff.org/health-misinformation-and-trust/press-release/poll-most-who-use-artificial-intelligence-doubt-ai-chatbots-provide-accurate-health-information/ (finding most adults are not confident in the accuracy of health information from chatbots).

As these example findings illustrate, evaluating a new chatbot can lead to different adoption pathways. The simplest decision is if the chatbot outperforms the human baseline in accuracy and either reduces call center volume (a substitution effect) or serves more people without increasing call volume (a net widening effect). The chatbot is a clear win in that case. However, states will likely face tradeoffs between accuracy and the number of customers served. The final decision should focus on how to best serve residents, with this evaluation exercise showing how disaggregated data can guide more informed choices.

Moreover, different AI adoptions will call for different evaluation designs. The implementation of a fraud detection system may focus even more on error rates, especially false positives, as well as demographic disparities, and downstream appeal success rates. An AI system assisting adjudicators may instead measure decision quality metrics, time-to-decision, and consistent application across workers.

C. Prospective Implementations of an Evaluation Mandate

We argue that the demands of existing legal structures militate for agencies to conduct evaluations when making decisions about significant new integrations of AI systems. Yet admittedly, none of these structures forms a textually explicit and legally binding evaluation mandate, especially in the sense of including a comprehensive human baseline. We explore four prospective embeddings of an evaluation mandate: in litigation, in agency guidance and rulemaking, in statute, and in procurement requirements.

i. Impact Litigation

As this Article has shown, the courts have stepped into retrospectively address the damages of an AI system gone wrong, in Michigan, and are also beginning to hold SNAP agencies accountable for their failure to meet constitutional and statutory obligations. These instances build on a long tradition of the courts filling a central role in institutional reform—through what Abram Chayes famously termed "public law litigation"—catalyzing restructuring of school systems, prisons, police departments, and public housing authorities.[317] There have long been doubts, however, about Chayes' optimistic view of the judge's role in overseeing such sweeping litigation both in terms of institutional competence and separation of powers. Judges are trained to adjudicate facts and law, not manage sprawling agencies through the ongoing oversight and monitoring often embedded in injunctive relief. Charles Sabel and William Simon reinterpret Chayes' vision of public law cases as experimentalist instantiations of "destabilization rights," or rights to disentrench an institution that has systematically failed to meet its obligations.[318]

One avenue for embedding an evaluation mandate is through impact litigation that explicitly requires agencies to present and defend performance metrics as part of meeting their due process obligations. Judges, after all, are better positioned than agencies to articulate desirable outcomes that are consistent with due process values—they regularly design and implement performance measures in other types of institutional reform cases—and they can defer to agencies' expertise in how best to satisfy these measures. Indeed, this is precisely what the judge in the Missouri SNAP case did when he ordered the agency to measure and report data on average call wait times and the percentage of

---

[317] Abram Chayes, *The Role of the Judge in Public Law Litigation*, 89 HARV. L. REV. 1281 (1976); Charles F. Sabel & William H. Simon, *Destabilization Rights: How Public Law Litigation Succeeds*, 117 Harv. L. Rev. 1016 (2004).
[318] Sabel & Simon, *supra* note 317.

denials based on failure to complete timely interviews.[319] This approach resonates with the requirement in *Mathews* that agencies demonstrate that they have meaningfully accounted for the risks and complexities of a given AI system. In doing so, due process litigation could move from a backward-looking, detail-bound posture to a forward-focused mechanism of decentralized, court-assisted institutional improvement.

ii. Agency Guidance and Rulemaking

A more comprehensive way for such an evaluation mandate to be imposed would be through the day-to-day policy guidance that federal agencies issue to guide states in their program administration. This especially makes sense if the evaluation mandate is seen as an interpretation of existing federal law, as we argued above. For example, DOL's demarcation of the inherently government boundary to merit staffing, especially as it pertains to automation, has been expressed through "Unemployment Insurance Program Letters" or UIPLs—the agency's favored format of policy guidance for UI administration. In one such 1995 letter, the agency describes the directives as "stat[ing] or clarify[ing] the Department's position, particularly with respect to the Department's interpretation of the minimum Federal requirements for conformity or compliance."[320] Unsurprisingly, the agency then asserts that such directives fall within the APA's exception to notice and comment requirements for "interpretive rules, general statements of policy, or rules of agency organization, procedure or practice."[321] Yet despite this claim to being an interpretive rule, the directive explicitly asserts to the states "that these directives do, in fact, have legal effect." This already seems in tension with APA caselaw which generally suggests that rules that are binding and strip away agency discretion—which seems to be implied by "legal effect"—require notice and comment.[322]

Despite the 1995's DOL's claims to the contrary and despite obvious speed and efficiency gains, policy guidance seems to legally be a poor vehicle for an evaluation mandate which by definition necessitates some form of binding action on the part of the states. The obvious pivot would then seem to be formal notice-and-comment rulemaking, which would have a binding effect on state agencies. While notice-and-comment rulemaking is admittedly much slower than issuing guidance,[323] it may still be quicker and more predictable than legislative change. Moreover, the exact articulation of an evaluation mandate raises some difficult questions that might benefit from agency-specific contextual knowledge and expertise. For example, the optimal scoping of an evaluation mandate is non-obvious. The scoping that aligns most clearly with the OMB M-Memo would be to only mandate an evaluation when making a decision on AI adoption.[324] But a mandate with such timing—even if it did include comparison against a human baseline—would asymmetrically add costs to considering AI adoption and thereby put a thumb on the scale incentivizing the status quo. A rule could balance this

---

319 Holmes, No. 2:22-CV-04026.
320 U.S. Dep't of Labor, Unemployment Ins. Program Letter No. 01-96 (Dec. 15, 1995), https://oui.doleta.gov/dmstree/uipl/uipl96/uipl_0196.htm.
321 5 U.S.C. §§ 552–553. In doing so, it cites a 9th circuit case in which a UIPL was affirmed to be an interpretive rule under the requirements of the APA. Rivera v. Becerra, 714 F.2d 887 (9th Cir. 1983).
322 *See* U.S. v. Texas (US 2015); Community Nutrition Institute v. Young (D.C. Cir. 1987). The 1995 UIPL concedes that the D.C. Circuit has already found part of a UIPL to constitute a substantive rule due to its imposition of "an obligation on the States not found in the statute itself." Cabais v. Egger, 690 F.2d 234, 239 (D.C. Cir. 1982).
323 *See* Mark Seidenfeld, *A Table of Requirements for Federal Administrative Rulemaking*, 27 FLA. L. REV. 533, 536-37 (2000) (identifying up to 109 steps an agency must complete before issuing a final rule).
324 The M-Memo also calls for ongoing monitoring and evaluation; however, its most full-throated evaluation requirements all lie within the section of practices to be followed "before" using new rights-impacting AI.

by mandating regular evaluation for all baseline systems that significantly affect benefit determinations. However, this would impose even greater costs on the agency.[325]

iii. New Legislation

Legislation is uniquely positioned to address the limitations of rulemaking in this context and the issues of funding. Moreover, a statutory mandate helped transform the IRS's telephone evaluation system from a set of reports required for congressional oversight to a core part of the agency's performance evaluation system. For much of the 20th century, the main evaluation paradigm of the IRS was compliance and enforcement, with traditional measures like audit quotas and case closures used to assess performance.[326] The IRS Restructuring and Reform Act of 1998 (RRA), coming on the tails of GPRA in 1993, shifted this focus to quality service and taxpayer experience. The RRA explicitly prohibited the service from using the "records of tax enforcement results" to evaluate employees or even set employee goals. Instead, the RRA mandated that the agency "use the fair and equitable treatment of taxpayers by employees as one of the standards for evaluating employee performance."[327]

The IRS began incorporating this legislative mandate in its 1999 launch of a "Balanced Measurement System" that sought to measure employee performance based on customer satisfaction (defined as the provision of "accurate and professional services to . . . customers in a courteous, timely manner"), employee satisfaction, and business results ("a productive quantity of work in a quality manner and [] meaningful outreach to all customers").[328] In 2002, the Balanced Measurement System evolved into the current Embedded Quality Review System,[329] which evaluates the service on a number of metrics, including call wait times and accuracy of advice given.[330]

The Social Security Act, in comparison, only requires that DOL affirm that state agencies administer benefits in a way "reasonably calculated to insure full payment of unemployment compensation when due."[331] Through regulation, the DOL established a quality control system to fulfil this mandate, which mandates that states "[c]omplete prompt and in-depth case investigations to determine the degree of accuracy and timeliness . . . with respect to benefit determinations, benefit payments, and revenue collections."[332]

A clearer statutory articulation of the expectation to evaluate potential AI systems against a human or status quo baseline seems promising, just as the RRA's expectation-setting triggered a transformation of how the IRS conducted performance evaluations. While evaluation is generally a bipartisan issue—the Evidence Act received healthy support from both sides of the aisle—there is a

---

325 Rulemaking has other downsides. Without congressional appropriation power, rules that impose new mandates operate in a zero-sum world of budgeting tradeoffs. There also may be increased vulnerability to judicial review under *Loper Bright*, as any basis for an evaluation mandate in existing law is likely to be dependent on agency interpretation. Loper Bright Enter. v. Raimondo, 603 U.S. 369 (2024).

326 Barry Bozeman, *Risk, Reform and Organizational Culture: The Case of IRS Tax Systems Modernization*, 6 INT'L PUB. MGMT. J. 117, 134 (2003).

327 28 U.S.C. § 7804 Note. The RRA also required the IRS to develop an employee training plan to train employees on how to provide quality customer service. *Id.*

328 INTERNAL REV. SERV., ORGANIZATIONAL PERFORMANCE MANAGEMENT AND THE IRS BALANCED MEASUREMENT SYSTEM (1999), https://www.irs.gov/pub/irs-pdf/p3561.pdf.

329 INTERNAL REV. SERV., EMBEDDED QUALITY REVIEW SYSTEM (EQRS) – PRIVACY IMPACT ASSESSMENT (2008), https://www.irs.gov/pub/irs-pia/eqrs-pia.pdf. The National Quality Review System is a subset of the Embedded Quality Review System.

330 IRS Declining Service, *supra* note 211.

331 42 US.C. § 503(a). One other option would be to update the language of extant merit staffing requirements in legislation like the SSA and the Food and Nutrition Act

332 20 C.F.R. § 602.

risk that any legislation that survives the political process would be an unfunded mandate. The unfunded nature of the Evidence Act's evaluation mandates, for example, have sometimes been cited as a reason that agencies' implementations of its provisions have been so slow.[333]

To avoid this fate, new legislation could draw inspiration from successful collaborative evaluation models, even those created through administrative action. The Manpower Demonstration Research Corporation (MDRC), a nonprofit, nonpartisan organization created by the Ford Foundation and a group of federal agencies in 1974 offers a promising a blueprint for how legislation could structure evaluation infrastructure.[334] MDRC was founded to run an ambitious, five-year $50-million demonstration project to test whether employment programs improved outcomes for disadvantaged populations. MDRC has since become known for combining rigorous impact and implementation research with on-the-ground operational expertise to deliver policy-relevant findings to decisionmakers. While MDRC emerged through foundation-government partnership rather than statute, its model suggests key elements that legislation should incorporate: dedicated funding for independent evaluation entities, requirements for collaboration between government agencies and research institutions, and mechanisms for translation findings into policy improvements.

Building on this model, legislation could establish regional AI evaluation centers modeled on MDRC's structure but with statutory authority and guaranteed funding streams. The Workforce Innovation Fund's approach—explicitly funding third-party evaluations and creating a National Evaluation Coordinator to synthesize findings across grantees—demonstrates how federal legislation can scale collaborative evaluation.[335] By combining RRA's mandatory evaluation requirements with MDRC's collaborative, independent structure that can incorporate private philanthropic support, new legislation could ensure that AI evaluation in benefits administration is both rigorous and sustainable, with dedicated resources to match its ambitious scope.

iv. Procurement

A final, but perhaps most promising, possible vehicle for an evaluation mandate would be to embed it within procurement requirements, an idea broadly outlined in the OMB's AI procurement memo.[336] However, the memo lacks three critical elements for success: a requirement to compare AI performance to a human baseline where possible, longevity across administrations, and resources to fulfill evaluation requirements.

As discussed earlier, a human baseline is critical to defining satisfactory AI system performance. For example, improper payments for state UI agencies often exceed 20%, peaking at 44% during the pandemic.[337] Meanwhile, UI call centers face issues like long wait times and low connection rates

---

333 Carmen Robinson, *Implementing the Evidence Act – What's Next?*, FEDERAL NEWS NETWORK (Mar. 28, 2024), https://federalnewsnetwork.com/commentary/2024/03/implementing-the-evidence-act-whats-next/; RESULTS FOR AMERICA, EVIDENCE ACT BRIEF (2019), https://results4america.org/wp-content/uploads/2019/09/Evidence-Act-Proposed-Next-Steps-FINAL.pdf (calling for the next version of the Evidence act to require agencies to set aside a portion of their program funds for evaluations so evaluations are better resourced).
334 MDRC, *Our History,* https://www.mdrc.org/about/history; *see also* JUDITH M. GUERON & HOWARD ROLSTON, FIGHTING FOR RELIABLE EVIDENCE (2013) (comparing the early MDRC to "an intermediary corporation, reflecting its position as a bridge between the interests of many diverse parties—public and private, national and local—and described its key strengths as flexibility, speed, and independence").
335 Abt Associates, The Workforce Innovation Fund (WIF): A Synthesis Report on Evaluation Findings and Experiences (2020), https://www.dol.gov/sites/dolgov/files/ETA/publications/ETAOP2022-25_WIF_Revised_Final_Synthesis_Report_%20Appendices.pdf.
336 *See* supra Part IV.A.iii.
337 *Unemployment Insurance Payment Accuracy by State*, Dep't of Labor, https://www.dol.gov/agencies/eta/unemployment-insurance-payment-accuracy/data.

(fewer than 40% of calls reach live representatives).[338] Procurement evaluations must compare both human-led and AI systems to contextualize adoption decisions, as discussed in our example evidence-based adoption cycle.

However, as the Evidence Act illustrates, quality evaluation requires time and resources. Since most state AI tools are developed under contract, setting aside part of the project budget for comprehensive evaluations could ensure adequate resourcing. This approach could draw on pay-for-performance frameworks used in medical technology procurement where procurement is costly and payoffs are uncertain.[339] One proposal ties contract prices to the evidence quality of the product's efficacy. In the AI context, vendors might compete to provide the strongest proof of improvement against human baselines, validated by agencies. Agencies could also negotiate refunds for tools where performance deviates significantly from initial evaluations. As AI grows more advanced, state AI contracts have increasingly shifted from smaller, more specialized vendors (e.g., Fast Enterprises in Michigan) to larger vendors with more diverse expertise (e.g., Google and Deloitte) who ought to be more capable in meeting these requirements, though their incentives may not align to do so on their own.

To further center evaluation in procurement, agencies should embed testing and comparative analysis into the very structure of AI contracts. Agencies could issue multiple small "trial" awards and require vendors to demonstrate their system's performance against an institutional baseline (and possibly against each other's solutions) before moving on to a larger, longer-term contract. This head-to-head testing not only clarifies which vendor is best able to meet an agency's performance benchmarks, but it also provides critical data for honing the measurement frameworks that will govern the final contract. Agencies might enlist independent evaluators—something already contemplated by scholars in non-procurement settings[340]—to audit these head-to-head trials, thereby adding an additional layer of credibility to the results.

Procurement must also address data ownership and portability. If contracts lock agencies into proprietary systems, evaluation findings may be difficult to act upon. Requiring vendors to provide interoperable data formats or export rights can enable smoother transitions to alternatives if performance benchmarks aren't met. By incorporating evaluation, testing, and data portability into contracts, agencies can create a procurement process that ensures rigorous oversight, supports due process, and serves the public interest.

****

While procurement represents a powerful and immediate vehicle for embedding evaluation requirements, it cannot stand alone as a comprehensive solution. Procurement-based mandates excel during system development and initial deployment, creating competitive incentives for vendors to demonstrate superior performance and enabling agencies to make evidence-based selection decisions. However, the operational lifecycle of AI systems extends far beyond the initial procurement period. Systems require ongoing monitoring, periodic re-evaluation against evolving baselines, and continuous adjustment as both technology and agency needs change. The temporary nature of most vendor relationships means that procurement mandates, while crucial, must be complemented by other institutional mechanisms—whether through litigation, rulemaking, or legislation—that can ensure sustained evaluation practices throughout a system's operational lifetime. A robust evaluation

[338] Dube et al., *supra* note 260.
[339] Marianne Lopez et al., *Paying for Value From Costly Medical Technologies: A Framework For Applying Value-Based Payment Reforms*, 39 HEALTH AFFAIRS (2020).
[340] *See, e.g.,* Citron, *supra* note 19; CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION (2016).

framework thus requires multiple, reinforcing vehicles that collectively address the full spectrum of AI system development, deployment, and long-term governance.

D. Reinterpreting Merit Staffing for the AI Age

i. The Limitations of the Current Framework

While merit staffing requirements were conceived as a response to "incompetent and politically dominated personnel" administering benefits in the early 20th Century,[341] merit staff does not guarantee competence in the modern system. Barely half of states meet federal quality guidelines in UI claims adjudication, and 45% of SNAP denials are improperly made.[342] This is particularly true given the modern volume and complexity of cases and the difficulty of reaching adequate staffing levels; indeed requiring merit staffing may undermine quality.[343]

AI heightens these challenges. While merit staffing targets a host of valuable goals—fair, independent, and efficient administration—its only means of achieving these goals is to assign tasks seen as central to these purposes to merit staff. This creates no space for an impartial examination of whether automation, or some combination of staff and automated tools, may better achieve these goals.

One way some agencies seem to have interpreted merit staffing requirements in the world of AI is as requiring that merit staff make a "final decision" on every case; the FNS even published a contractor report explicitly drawing this conclusion.[344] Implementing human review only at the final decision stage is undesirable for a number of reasons. At a practical level, such late-stage review does not address issues embedded in the AI tool's design, such as lack of explainability, making it difficult to identify and correct systemic flaws. This improperly relieves agencies of a necessary pressure to ensure that the tools they are using are interpretable and understandable, both for good design purposes as well as to allow for appellate review. Theoretically, this "final review" interpretation can create the façade of individualized due process without its substance. Instead of mandating that humans are involved at more critical steps in the design and evaluation process, it reduces agency employees to the difficult and bias-prone job of being auditors of the output of black-box AI decisions rather than fulfilling more meaningful and active roles. Within the *Mathews* framework, AI decisions rubber-stamped by humans at the end increase the risk of erroneous deprivation of private rights, whereas ongoing human involvement in design and testing could enhance procedural fairness by identifying and correcting misunderstandings before decisions are made.

The merit staffing paradigm provides an important set of values—impartiality, competence, and efficiency—which should remain agencies' north star when designing systems. However, the binary of dividing tasks between merit and non-merit staff is unable to manage the benefits of risks of AI that can be integrated into benefits administration in a spectrum of ways.[345] In the worst case, the merit staffing paradigm encourages states to overly focus on the human role at the final decision stage, which invites cursory review prone to automation bias, and deprives agencies of necessary pressures to inspect and improve their AI's decisionmaking process.

---

[341] *But cf.* Diana Moreira & Santiago Pérez, *Civil Service Reform and Organizational Practices: Evidence from the Pendelton Act*, 16 AM. ECON. J.: APPLIED ECON. 250 (2024).
[342] *See supra* Part 1.
[343] RICHARDSON & PAHLKA, *supra* note 51.
[344] See *supra* Part 2.C; 3.B.iii (Wisconsin and Nevada emphasizing that humans are reviewing each recommendation before a decision is issued).
[345] *See* Daniel E. Ho, Olivia Martin, Kit Rodolfa, Amy Perez & Faiz Surani, *The Spectrum of AI Integration: The Case of Benefits Adjudication*, *in* AI: LEGAL ISSUES, POLICY & PRACTICAL STRATEGIES (2024).

ii. Building Toward a New Role for Merit Staff

Agencies can and should reinterpret merit staffing requirements to embrace evaluation-driven AI integration that enhances rather than replaces human oversight. Current DOL guidance already recognizes (in some instances) that automation is permissible where "facts lead to only one conclusion."[346] While a difficult line to police, it is a good start. Rather than limiting this standard to the narrow set of cases where human disagreement is literally impossible, agencies should interpret "facts lead to one conclusion" to encompass any determination where rigorous evaluation demonstrates that AI-assisted decision-making produces more accurate (or somehow better) outcomes than unassisted human review. This interpretation aligns with the core purpose of merit staffing: ensuring competent, independent, and efficient administration. For example, if evaluation shows that merit staff using AI fraud detection tools achieve 15% error rates compared to 25% rates for unassisted staff, the "facts" of system performance clearly "lead to one conclusion" about which approach better serves program integrity.

This reinterpretation requires reconceptualizing the role of merit staff from individual case processors to system validators and overseers. Rather than requiring all merit staff to personally review every document and make every micro-decision, agencies should also focus merit staff expertise on system design validation, performance monitoring, and exception handling. First, merit staff should be central to designing and testing AI systems before deployment. This includes defining performance benchmarks, identifying potential failure modes, and ensuring that automated systems align with program goals and legal requirements. Second, merit staff should continuously monitor AI system performance against established benchmarks, including by conducting random audits of AI decisions, tracking error rates across different populations, and identifying when human intervention is needed. Finally, merit staff should maintain direct responsibility for complex cases that fall outside AI's capabilities, cases flagged for human attention and appeals of AI-assisted decisions. This preserves human judgment for the most consequential determinations while allowing automation to handle routine processing.

Conclusion

Three years before signing the Social Security Act into law, President Roosevelt famously urged "bold, persistent experimentation," emphasizing the need to "take a method and try it: If it fails, admit it frankly and try another. But above all, try something."[347] Many state agencies administering public benefits have heeded this call—facing the dual mandate to ensure civil servant oversight of eligibility determinations while delivering benefits accurately and efficiently, they've cautiously embraced automation. Whether rushing into automation without adequate testing, as Michigan's UI agency did, or clinging to a dysfunctional status quo, as Missouri's SNAP agency did, due process often hangs in the balance.

This Article offers a framework to navigate these challenges. It provides the first systematic analysis of how AI and automation intersect with the legal and historical functions of the civil service. By documenting the risks and opportunities presented by state-level automation experiments, the Article highlights the need for more deliberate oversight and proposes evaluation as the appropriate

[346] *ET Handbook* 301, *supra* note 35 at 9.
[347] Franklin D. Roosevelt, Oglethorpe University Address (May 22, 1935), https://publicpolicy.pepperdine.edu/academics/research/faculty-research/new-deal/roosevelt-speeches/fr052232.htm.

mechanism to ensure that modernization efforts align with constitutional principles of fairness and accountability.

Mathews provides agencies with a roadmap for balancing private interests, the risk of erroneous deprivation, and administrative burden. For today's era of automation, this guidance translates into rigorous testing and benchmarking—allowing agencies to determine the optimal balance between human and machine decision-making to minimize the risk of erroneous deprivation.

This moment of transformation presents both significant opportunities and profound risks. Properly deployed, AI can amplify the effectiveness of civil servants and improve the accuracy and efficiency of public benefits administration. Without sufficient safeguards, however, it risks undermining the very principles it seeks to support. By coupling technological innovation with rigorous evaluation, this Article charts a path forward that preserves the foundational values of fairness, accountability, and trust in government. Roosevelt's call to "try something" must be met with an equally resolute commitment to ensure that what we try enhances justice for the people these systems are designed to serve.

Appendix Table A: Response Summaries for Chatbot Audit by Question

State	*Scope*	*Low complexity query*	*Medium complexity query*	*High complexity query*
Alabama[348]	State	Could not answer	Could not answer	Could not answer
Arizona[349]	Dep't of Economic Security	Could not answer	Could not answer	Could not answer
Arkansas[350]	State	Could not answer	Could not answer	Could not answer
California[351]	Employment Development Dep't	As expected	Not answered (same response as Q1)	Roughly as expected
Colorado[352]	Dep't of Labor	As expected	Not answered (generic eligibility)	Not directly answered
Connecticut[353]	Dep't of Labor	As expected	As expected	Not directly answered (generic eligibility)
Delaware[354]	Dep't of Labor	Could not answer	Could not answer	Could not answer
Florida[355]	Dep't of Labor	As expected	As expected	As expected
Georgia[356]	Dep't of Labor	As expected	Not answered (generic eligibility)	Not directly answered (generic eligibility)

---

348 Alabama.gov, https://www.alabama.gov/ (click on the red icon at the bottom right of the page).
349 Arizona Dep't of Economic Security, https://des.az.gov/services/employment/unemployment-individual (click on "Need Help? Ask DAVE" at the bottom right of the screen).
350 Arkansas.gov, https://portal.arkansas.gov/ (click on the yellow icon at bottom right of the page).
351 California Employment Development Dep't, https://edd.ca.gov/en/unemployment/ (click on the blue icon at bottom right of the page).
352 Colorado Dep't of Labor and Employment, https://cdle.colorado.gov/unemployment (click on the blue icon at bottom right of the page).
353 Connecticut Dep't of Labor, https://portal.ct.gov/dol/unemployment-benefits?language=en_US (click on grey icon at bottom right of page).
354 Delaware Dep't of Labor, https://labor.delaware.gov/divisions/unemployment-insurance/ (click on "Ask Ara" at bottom right of page).
355 Florida Commerce, https://www.floridajobs.org/reemployment-assistance-service-center/reemployment-assistance/claimants/apply-for-benefits, (click on blue icon at bottom right of page).
356 Georgia Dep't of Labor, https://dol.georgia.gov/ (click on the green icon at the bottom right of the page).

Indiana[357]	State	As expected	As expected	As expected
Iowa[358]	State	Could not answer	Could not answer	Could not answer
Kansas[359]	State	Could not answer	Could not answer	Could not answer
Maine[360]	Dep't of Labor	Could not answer	As expected	Could not answer
Massachusetts[361]	State	As expected	Redirected to overpayment	Not directly answered (generic eligibility)
Michigan[362]	Dep't of Labor and Econ. Opp.	As expected	As expected	As expected
Mississippi[363]	State	As expected	As expected	As expected
Montana[364]	Dep't of Labor and Industry	Could not answer	Could not answer	Could not answer
New Jersey[365]	Dep't of Labor	As expected	Redirected to pandemic relief	Not directly answered (generic eligibility)
New York[366]	Dep't of Labor	Could not answer	Could not answer	Could not answer
North Carolina[367]	Dep't of Commerce	As expected	As expected	Not directly answered (generic eligibility)

357 Indiana.gov, https://www.in.gov/core/chatbot.html.
358 Iowa.gov, https://www.iowa.gov/ (click on teal icon at the bottom left of the page).
359 Kansas.gov, https://portal.kansas.gov/ (click on blue icon at bottom right of page).
360 Maine Dep't of Labor, https://www.maine.gov/unemployment/ (click on blue icon at bottom right of page).
361 Mass.gov, https://www.mass.gov/how-to/apply-for-unemployment-benefits (click on green icon at bottom right of page).
362 Michigan Dep't of Labor & Econ. Opp., https://www.michigan.gov/leo/bureaus-agencies/uia (click on teal icon at bottom right of page).
363 Ms.gov, https://www.ms.gov/.
364 Montana Dep't of Labor & Industry, https://uid.dli.mt.gov/ (click on blue icon at bottom right of page).
365 Nj.gov, https://www.nj.gov/labor/myunemployment/.
366 New York Dep't of Labor, https://dol.ny.gov/unemployment/unemployment-insurance-assistance (click on teal icon at bottom right of page).
367 North Carolina Dep't of Commerce, https://www.des.nc.gov/individuals/apply-ui (click on green icon at bottom right of page).

Ohio[368]	Dep't of Job and Family Services	Redirected to employer resources	Redirected to work authorization	Could not answer
South Dakota[369]	State	As expected	As expected	As expected
Texas[370]	State	When asked about unemployment insurance, immediately connected to a live human agent	NA	NA
Utah[371]	Workforce Services	Could not answer	Could not answer	Could not answer
Virginia[372]	Employment Commission	As expected	Could not answer	Could not answer
Wisconsin[373]	Department of Workforce Development	Roughly as expected (link but no description)	Generic response about application process with no mention of documents	Not directly answered (generic eligibility)

Appendix Table B: Response Summaries for Generative AI Chatbot Audit

Question	***Indiana***	***Michigan***	***Florida***	***South Dakota***	***Mississippi***
How much money do I have to have earned in the base period to be eligible for unemployment insurance?	Correct	Partially correct: stated that one quarter's wages must be at least be $4,842 for "benefit years	Correct	Cannot verify	Incorrect: stated that you must have earned at least $780 in the base period,

368 Ohio Dep't of Job & Family Svcs., https://jfs.ohio.gov/job-services-and-unemployment/unemployment/overview (click on blue and yellow icon at bottom right of page).
369 SD.gov, https://www.sd.gov/cs (click on yellow bird icon on right of page).
370 Texas.gov, https://www.texas.gov/texas-by-texas/ (click on the red icon on the bottom right of the page). Texas also recently launched what they term a "personal, portable government assistant" called "Texas by Texas (TxT)" which simplifies processes like renewing vehicle registration.
371 Utah Workforce Services, https://jobs.utah.gov/ui/home (click on blue icon on bottom right of page). Note that we tried to access the chatbot of various times within Utah's business hours on different days and were unable to get responses.
372 Virginia Employment Comm'n, https://www.vec.virginia.gov/unemployment (click on orange icon on bottom right of page).
373 Unemployment Insurance, Wisconsin Dep't of Workforce Development, htttps://dwd.wisconsin.gov/ui/ (click on cow icon on bottom right of the page).

		starting Jan. 1, 2025," but it does not define benefit year and from January 1 to February 20, 2025 the requirement is only $4,097.[374]			whereas the requirement is to have earned $780 in the highest quarter.
I have earned $7000 in the base period with $3000 in one quarter. Am I eligible for unemployment insurance?	Correct	Correct	Incorrect: asserts that $7,000 is not greater than $4,500	Cannot verify	Correct
If I file my initial claim for when I was unemployed between January 5, 2020, and April 4, 2020, what is my base period?	Correct	Incorrect: returns the last four quarters of the base period rather than the first four	Correct	Unanswered	Correct
What is my base period if I file my claim for when I was unemployed between October 7, 2018 and January 5, 2019?	Correct	Correct	Correct	Incorrect: temporally implausible response	Correct
How do I calculate my maximum benefit amount?	Correct	Correct	Correct (though wrong link)	Unanswered	Partially incorrect: stated wrong maximum benefit amount[375]

[374] Michigan Dep't of Labor & Econ. Opp., *About UIA and Unemployment Insurance in Michigan*, https://www.michigan.gov/leo/bureaus-agencies/uia/uia-resources-for-claimants/about-uia-and-unemployment-insurance-in-michigan.

[375] The tool suggested that the maximum benefit amount is $275. However, an amount of $235 is listed on the state's guidance for claimants as well as in statute. *See* Mississippi Dep't of Employment Security, *Table of Weekly Benefit Amounts and Maximum Amounts*, https://mdes.ms.gov/media/7140/Benefit_Chart1.pdf; Miss. Code §71-5-503 (2024). The chatbot may have been confused by a bill that had been proposed to raise the amount to $275; however, the bill died in committee in 2025. *See MS HB299*, BILL TRACK 50, https://www.billtrack50.com/billdetail/1773993.

I know for a fact that if you are really dissatisfied with your job and leave because of that, you are still eligible for unemployment insurance. I left my job because my boss made me very unhappy. I'm still eligible for unemployment insurance, right?	Correct	Correct	Unanswered	Correct	Correct
There was a labor dispute at my job. How does this affect my claim for unemployment insurance?	Correct	Unanswered	Correct	Correct	Correct
Under what conditions am I deemed unavailable for work, as a matter for unemployment insurance, if I'm on vacation?	Correct	Unanswered	Correct	Unanswered	Correct
I've heard that I have to be available for work in order to keep getting unemployment insurance. But what if the other work available is really bad and I don't want to do it?	Correct	Unanswered	Correct	Unanswered	Correct
How many work search activities am I required to participate in per week?	Incorrect: stated that three search activities whereas	Correct	Correct	Unanswered	Correct

	only two are required.[376]				
I think I got overpaid on my unemployment insurance benefits. Do I have to repay the agency if I didn't do anything wrong?	Correct	Correct	Correct	Correct	Correct

[376] *See* INDIANA HANDBOOK, *supra* note 181 at 22.