Using Artificial Intelligence to Identify Administrative Errors in Unemployment Insurance

Abstract

Administrative errors in unemployment insurance (UI) decisions give rise to a public values conflict between efficiency and efficacy. We analyze whether artificial intelligence (AI) – in particular, methods in machine learning (ML) – can be used to detect administrative errors in UI claims decisions, both in terms of accuracy and normative tradeoffs. We use 16 years of US Department of Labor audit and policy data on UI claims to analyze the accuracy of 7 different random forest and deep learning models. We further test weighting schemas and synthetic data approaches to correcting imbalances in the training data. A random forest model using gradient descent boosting is more accurate, along several measures, and preferable in terms of public values, than every deep learning model tested. Adjusting model weights produces significant recall improvements for low-*n* outcomes, at the expense of precision. Synthetic data produces attenuated improvements and drawbacks relative to weights.

Keywords: artificial intelligence, machine learning, social policy, unemployment insurance, administrative errors

1. Introduction

Public organizations, especially welfare agencies, regularly make decisions affecting how scarce goods are distributed. Some of these decisions are *administrative errors*—decisions that will turn out to have been wrong or mistaken. Detecting and reducing such administrative errors is central to quality control in the public sector. As the scale and speed of administrative decisions outstrips the capacity of human agents to manage, many organizations are adopting artificial intelligence (AI) to quickly audit large volumes of administrative data. Adopting AI for this purpose enjoys bipartisan support in the United States; in reviewing the IRS' use of AI one Republican member of congress referred to AI as "the ultimate auditor" (Heckman, 2020). Yet little is known about the true performance of these systems, or the distributive consequences of their use in auditing public spending and services.

This article examines the use of AI to support the identification of administrative errors. We concentrate on the case of US State workforce agencies and their decisions of Unemployment Insurance (UI) claims. In deciding claims, mistakes consist in over- or under-paying claimants. For example, claims that should be approved but are wrongfully denied constitute underpayments of the full dollar amount claimed. We use data collected by the US Department of Labor (DOL) to train different Machine Learning (ML) based AI systems to predict administrative errors in this setting and analyze their performance. We then discuss whether such algorithms can and should be used to effectively detect mistaken decisions.

We are motivated by the question of whether AI *can* and *should* be used to detect administrative errors. The article makes first steps towards answering this question and improving AI-driven decision-making in government by examining prominent ML-based AI technologies[1] for a specific task currently performed by human agents (auditors). We use labeled audit data that are similar to those that are likely employed in developing such systems in practice. The article compares the performance of these different techniques and discusses their goodness of fit relative to these data. We thus contribute to the literature on technological innovation in the public sector, as well as administrative errors and improper payments.

AI holds out the promise to further a classic public value: *efficiency.* But greater efficiency, such as reducing costs, may come at the expense of at least one other important public value: *efficacy*, that is, making sure that those who are eligible for services receive them. Administrative errors hence have a public values conflict at their center. We argue that reducing administrative errors helps to overcome this conflict — *but only when the objective function is to identify and minimize all error types, not just politically salient ones*. With greater accuracy in decision making, multiple public values can be furthered at the same time.

[1] Machine learning is the current dominant paradigm of AI architectures. Its successes across a variety of task domains has leant it something like an aura of omnipotence in media and marketing, and as a result 'ML' is often interchangeably used with 'AI.' We explain our use of these terms in section 2.1 below.

This is particularly important today. UI fraud has long been a topic of attention and concern. In the US, 296,749 cases of fraud were identified in 2019, amounting to $366.8 million (Pallasch, 2020). During the COVID-19 pandemic the number of fraudulent claims has increased significantly (FBI National Press Office, 2020), resulting in "tens of billions of dollars in payments" to illegal claimants (Cowley 2022). At the same time, the pandemic has brought attention to non–fraudulent over- and underpayments of UI benefits and to insufficient timeliness with which UI claims are decided (De La Garza, 2020). The Bank of America, for example, which in 12 states delivers benefits payments as prepaid debit cards, underpaid around 100,000 people a total amount of several hundred million USD over 2020-21 (Cowley 2022).

The importance of recognizing systemic administrative errors in programs like UI has long been emphasized (Kingson & Levin, 1984). That AI could be used to identify administrative errors is also well-known (Bullock et al., 2020; Charette, 2018). Some classifiers have been used in other studies to detect fraud in Medicare payments (Bauder & Khoshgoftaar, 2017; Farbmacher et al., 2020; Juan Liu et al., 2016). But the question of which concrete AI-technologies *can* and *should* be used to address administrative errors, in a given application domain, has not been answered in full.

We answer this question in three ways. First, we analyze the normative considerations involved in using AI to audit administrative performance in social benefits programs. Second, we test the effect of adjusting the decision weights of our best-performing classifier on its ability to correctly identify under-payments in particular. Third, we test whether generating and training the model on synthetic data improves or otherwise affects system performance.

Briefly stated, we find that a random forest classifier using gradient descent boosting, CatBoost, is superior to several different deep learning-based classifiers both for accuracy and explainability. Furthermore, while all of the classifiers tested perform relatively poorly on classifying underpayment errors, CatBoost's performance was substantially improved by using an alternative decision weight scheme, as well as by generating and training the model on synthetic data. These performance improvements, however, came at a cost to performance on predicting error-free claims as well as overpayments. Both the improvements and the losses were significantly more pronounced for weighting adjustments than for synthetic data. Our findings suggest that the use of AI for auditing UI claims can satisfy multiple public values, but this requires particular attention when evaluating alternative technologies before adoption and implementation.

2. Background

The potential for Machine Learning-based Artificial Intelligence (AI) to transform economic and social activity has captured the world's attention over the past several years. Despite its reputation as a laggard with respect to innovation, the public sector has also embraced the promise of AI (Glaze et al., 2022). Governments are investing heavily in research and development of AI and deploying autonomous and intelligent tools and systems (Perrault et al.,

2019). These investments have led in turn to AI's use in various areas of public service, including the use of facial recognition and predictive analytics (i.e., predictive policing) in law enforcement; virtual assistants in human services and military recruiting; detecting improper payments and fraud in social services; optimizing food safety inspections and discovering new drugs in public health; and cybersecurity (Eggers et al., 2017; Engstrom et al., 2020; Wirtz et al., 2019). The use of AI in government decision-making, however, is contested: optimism about the technology's ability to improve efficiency and reduce administrative burdens is tempered by evidence of systematic harmful discrimination in system outputs, the inability to adequately explain how some of the most powerful AI architectures arrive at their decision outputs, and the risk of serious wrong to individuals and societies (Buolamwini & Gebru, 2018; Byrnes, 2016; Courtland, 2018; Levy et al., 2021; Young, Himmelreich, Bullock, et al., 2021).

2.1 Conceptualizing Artificial Intelligence

It is necessary to unpack "AI" as a term before proceeding further, because its capabilities are distinct from other forms of algorithmic governance in profound ways. This is not a straightforward task, however, because "AI" is an amorphous term. In the broadest sense, "AI" includes systems designed to think or act either like rational human beings (Russell and Norvig 2015). For the purposes of this article, we use "AI" to refer specifically to *systems that employ machine learning (ML) approaches to make inferences — decisions — from data without explicit programming* (Witten et al., 2016). There are numerous different ML architectures, each with their own advantages and disadvantages. Two of the most widely adopted and well-known ML architectures are support vector machines and artificial neural networks; complex variants of the latter are also known as "deep learning" systems (Le et al., 2013). Although our empirical analysis concentrates on the ML paradigm in AI in particular, we talk about "AI" more generally in this paper (a) because "AI" is the term typically used in public administration for this topic, even if ML is meant; (b) because much of our normative contribution applies not only to ML but to AI generally, regardless of the specific paradigm; and (c) our normative contributions also overlap with research on "explainable AI (XAI)" which deals with the tradeoff between complexity and performance in machine learning-based AI decision-making.[2]

Machine learning has three important dimensions. The first is its reliance on data that can be read, manipulated, and processed by standard computational methods, also known as "machine readable" data. Access to machine readable data is of such fundamental importance to AI that data digitization and operation efforts account for more than half of all AI-related investment (The Economist, 2020). The second dimension of ML is its learning process. This process takes three general forms: supervised, unsupervised, and reinforcement (Russell & Norvig, 2015). ML's final dimension is the use of stochastic learning and optimization techniques in many of its functional forms. This introduces inherent uncertainty in ML processes in exchange for significant performance increases.

[2] See e.g., https://www.darpa.mil/program/explainable-artificial-intelligence

These features differentiate AI as used here from other approaches to artificial intelligence such as "expert systems." Expert systems use predetermined criteria to analyze structured data. Except for "edge case" inputs that result in system failure, any output can be used to determine the initial input factors through reverse process engineering. Unlike expert systems, most modern AI applications — and all of the more powerful, complex forms, such as neural networks — are stochastic. This means that AI is not deterministic in the same way as expert systems; you cannot reverse engineer a decision reached by an AI with certainty despite having complete knowledge of all available input variables/features. This is reflected in how computer scientists evaluate AI decision performance: AI is evaluated probabilistically (e.g., a 93% success rate at making the correct classification decision), rather than in terms of absolute fidelity. Moreover, the final architecture of any AI system after it has been trained is fundamentally unique — even when the trainer uses identical initial architectures and training data. This critical distinction makes the use of AI in administrative decision-making more like the use of human labor in real time: just as human agents will systematically differ — even slightly — in their weighing of identical decision criteria, so too will AI.

2.2 Artificial Intelligence in Public Administration and Management

The theoretical and practical implications of this technology have not gone unnoticed in the research community. Public administration scholarship has seen a marked increase over the last several years in the attention paid to public sector AI implementation. Theoretically driven work has identified different conceptual approaches for understanding and governing public sector AI. These include networked, intersectoral approaches (Wirtz et al., 2020; Wirtz & Müller, 2019); those using a systems engineering approach to understand the relationship between social and technological characteristics and relations (Janssen et al., 2020; Janssen & Kuk, 2016); and approaches that focus on the type of task, its associated information requirements, and the level of discretion required (Bullock, 2019; van der Voort et al., 2019; Young et al., 2019).

Other theoretical, and often normative, work considers some of the particular challenges governments face when using AI. One of these is AI's inverse relationship between analytic power and explainability. As the dimensionality of analysis increases, it becomes correspondingly harder to explain the decision process developed by the AI in ways intelligible to humans (Danks, Forthcoming). A specific sub-domain of AI research — Explainable AI (XAI) — focuses on this problem (de Bruijn et al., 2021). The second challenge is AI's proclivity for optimizing its decision approach in ways contrary to the user's broader values. Harmless examples include AI systems trained to beat speed or score records in video games, succeeding by effectively hacking the game instead of playing it (Amodei et al., 2016). In public organizations, however, AI can optimize around systematic biases embedded in training and reinforcement data, harming vulnerable individuals and populations, or violating legal statute and ethical norms on equality under the law (Janssen & Kuk, 2016; Young, Himmelreich, Bullock, et al., 2021). As with explainability, this problem motivates active research on data governance and bias minimization in AI (Janssen et al., 2020).

Empirical work on public sector AI is similarly multifaceted. Some seek to understand the factors influencing the implementation decision (Alshahrani et al., 2021; Mikalef et al., 2021). Another vein of empirical research focuses on the individuals and their experience with public sector AI. These include studies of individuals' acceptance of being subject to AI-enabled processes, and citizens' level of trust in AI-mediated interactions with government (Doberstein et al., 2021; Huang et al., 2021; Ingrams et al., 2021). Others examine the effect of AI implementation on the exercise of discretion — the latitude individual agents have to shape administrative decisions and tasks –within public organizations (Bullock et al., 2020; Criado et al., 2020; Flügge et al., 2021). Another branch of organizational-level research includes studies of AI's potential to improve public sector organizational processes, such as optimizing public transportation in response to citizen requests (Kim & Hong, 2021).

This research contributes to this latter line of empirical inquiry by simulating the use of AI as an audit support tool. Because administrative decisions regarding program eligibility and appropriate benefit levels carry significant distributional consequences at both the individual and social level, we also contribute to the normative debates on government use of AI.

AI and algorithmic decision making play an increasingly prominent role in the public sector. It has been used to confirm the identity of taxpayers online, to enforce regulations at the SEC, and to determine whether a suspect is granted bail, and whether an immigrant is detained. AI is used in criminal justice, public health, child-welfare, education, policing, and regulatory enforcement (Bullock et al., 2020; Engstrom et al., 2020; Levy et al., 2021). One example: As the COVID-19 pandemic ravaged prisons, who had to stay in prison and who was released to shelter at home was determined by predictions of inmates' recidivism risk.[3]

In each decision, the stakes are high and the impacts are profound (Eubanks, 2018; O'Neil, 2017). Thus, general theoretical frameworks are being developed to guide the development and deployment of AI in the public sector (Young et al., 2019). In the next section, we lay out the normative issues embedded in our study before turning to the empirical policy context and data generative process.

2.3 Public Values of Unemployment Insurance

This article has not only a descriptive but also an evaluative aim, namely, to provide a normative lens on the use of AI for unemployment insurance (UI) and improper payments. In the following, we draw on the well-known approach of *public values* to present a way of evaluating the use of AI to detect improper payments in UI (Fukumoto & Bozeman, 2019; Jørgensen & Bozeman, 2007; Nabatchi, 2018). We first offer an analysis of what values an UI should further. This clarifies the dimensions of the normative assessment. We then, in a second step, defend the plausibility of three propositions on when and how AI should be used to detect improper payments in UI.

[3] The *CARES Act* of 2020 extended the maximum time of existing home confinement rules, which already require a risk assessment. This risk assessment is performed by PATTERN (Prisoner Assessment Tool Targeting Estimated Risk and Need) which was created in response to the *First Step Act* of 2018.

For this analysis, we need to step back from our focus on administrative errors and consider the larger process of UI claims. For simplicity, we model the UI claim application process as two-stage selection operating on an initial set of individuals, the population. Each individual in the population is either eligible or ineligible for UI. The first stage — which cannot be observed in our data — consists in individuals' decisions to file an UI claim. Presumably, most of those who decide to file a UI claim are in fact eligible, but some are not — some might think that they are eligible although they are not, and some of those who file might attempt to defraud UI. Likewise, most of those who decide *not* to file a UI claim are *in*eligible, but some *eligible* individuals do *not* file either because they might not know that they are eligible for UI or they decide against filing for other reasons. From the first stage, we concentrate on those individuals who filed claims for UI, some of which are ineligible.

The second stage, from which our data are sampled, consists in the eligibility determination and claims decision by state workforce agencies. Again, presumably most of those who are deemed eligible are in fact eligible, but some are not. And most of those who are deemed ineligible are truly ineligible, but some are not. Errors can be *overpayments*, that is, individuals who were deemed eligible although they are in fact not eligible, as well as *underpayments*, that is, individuals who were deemed ineligible although they are in fact eligible.[4]

This is, of course, a vastly simplified and highly schematic model of how UI works. For example, decisions about eligibility and fraud are not all made at the same time — but we collapsed these processes into the second stage. Moreover, determinations are not a binary classification, as we assume here. Yet, despite these simplifications, this model allows us to bring out some of the normatively relevant features of UI. The model allows us to approach the questions of when and how AI *should* be used to detect improper payments.

A prominent method for normative analysis in public administration is the public values approach. Questions around ethical and professional values are fundamental to the field of public administration. They have played a prominent role in the field's history — from Frank Goodnow's and Leonard White's views on managerialism vs. legalism, via the Simon–Waldo debate, to the Minnowbrook conferences — and such discussions play an increasingly prominent role today (Van der Wal et al., 2015; Van der Wal et al., 2011) [add reference]. In addition to debating the inventory of values, their content, and their importance in the abstract, value analyses have been used to develop sector- or service-specific normative frameworks, for example for public infrastructure and utilities in the face of privatization (De Bruijn and Dicke, 2006) [add reference]. Our contribution here aims to likewise offer an application-specific value-based evaluative framework. At the same time, however, it should be acknowledged that the approach of public values has serious potential limitations: The collection of values can seem arbitrary, values are seen as exogenous to public administration such as originating from

[4] Under- and over-payments can also occur when claimants were both duly eligible to receive benefits and received them; some receive more than they were supposed to, and some less.

branches of government, the law, politics, or democratic ideals, and "laundry lists" insufficiently arbitrate value conflicts (Heath 2020, 51-53) [add reference].

We place our normative contribution within the recent literature on public values (Fukumoto & Bozeman, 2019; Jørgensen & Bozeman, 2007; Nabatchi, 2018). Drawing on such existing inventories of public values, we argue that at least four values are relevant to UI. We argue that UI should simultaneously further efficiency, efficacy, integrity, and equity. Each of these values has different dimensions or aspects that we articulate below. Our focus is not on the values themselves nor on their foundations, but on conflicts between them. That is, we contend neither that the following values are exhaustive of the values relevant to UI, nor do we suggest that these four values are more important than other values not mentioned here. Instead, the following four values are noteworthy for UI because they tend to conflict with one another especially in conflicts that manifest in policies that govern agency decisions. AI embodies such policies and centralizes them at scale. The conflicts between the four values are thus central to evaluate the use of AI in supporting agency decision-making. In the interest of ecumenism, we articulate the values in a way that should be acceptable to many without relying on any particular moral or political theory. As such, UI should further at least the following four principal values:

1. **Efficacy**: provide insurance payments to those who are eligible in a convenient and timely manner.
    a. **Opportunity**: enable individuals who are likely eligible to apply, e.g., offer an application process that is convenient for eligible claimants.
    b. **Payment**: render goods/services for eligible claims to claimants quickly.
    c. **Avoid underpayment**: minimize under-payment, i.e., reduce false negative eligibility.
2. **Efficiency**: reduce unnecessary monetary cost.
    a. **Cost**: minimize cost of administering insurance claims.
    b. **Avoid overpayment**: minimize sum of overpayment amounts, i.e., reduce false positive eligibility.
3. **Integrity**: make determinations based on consistent policy; ensure quality and safety of both model and data.
    a. **Audit:** audit and assess quality of the training data.
    b. **Privacy**: protect sensitive claimant data; avoid data breaches.
    c. **Explainability**: Provide explanations of how eligibility decisions are made (i) in general, and/or (ii) in any particular case.
    d. **Agency**: at least at high-stakes decisions, retain the authority and agency to decide what data, methods, and modeling techniques are used.
4. **Equity**: avoid that protected groups are disadvantaged or discriminated against.
    a. **Error Parity**: the false negative rate, i.e., rate of underpayment, in determining eligibility should be the same across all protected groups.

b. **Predictive Parity**: the probability that an individual, who was determined to be ineligible, is actually ineligible should be the same across all protected groups.

Efficacy formulates one central aim of UI, namely, to provide insurance payments. This value is fundamental and intrinsic to the very idea of UI. Notable, given our two-stage model of the process, the value of efficacy is relevant already at the point at which individuals decide whether or not to claim UI — what we called the first stage (*opportunity*). In the second stage, when workforce agencies make claim decisions, efficacy demands that underpayments are avoided and that payments are made quickly (*payment* and *avoid underpayment*).

Efficacy is grounded in widely recognized public values. Insofar as the workforce agencies carry out existing law, "efficacy" refers to the satisfaction of individual legal rights and procedural due process. Efficacy is hence rooted in a legal value frame (Nabatchi, 2018). In the case of UI specifically, efficacy moreover rests on public values in the category "public sector's contribution to society" (Jørgensen & Bozeman, 2007), such as social cohesion, altruism, and human dignity. This is because the administration of UI — supporting those eligible for unemployment benefits — is arguably a form of altruistic cooperation that furthers social cohesion and ought to respect human dignity.

However, although the public values to which "efficacy" refers are widely recognized, the name "efficacy" is not often used in the public values literature.[5] This might be because efficacy is implied by efficiency, at least on one definition of "efficiency".[6] Moreover, this might be because of a difference in topical focus: Efficacy is a property of a concrete government program, such as our topic here. By contrast, it seems less natural to speak of the "efficacy of public administration" as a whole, which is the topic of the public values literature.

Efficiency, the second principal value, reflects a fiduciary obligation to avoid unnecessary costs. It means cost-savings and cost-efficiency (*cost*) and is hence rooted in a market-based value frame (Nabatchi, 2018; Stone, 2011), and in long-standing discussions in public administration around the role of managerial and business values (Bozeman, 2007). Beyond administrative expenses, efficiency is increased with the accuracy of claim decisions. Specifically, unnecessary costs are reduced when overpayments are reduced (*avoid overpayment*).

Third, the value of *integrity* reflects a different kind of due process consideration. If a workforce agency had an AI model that is highly accurate in determining eligibility determinations, this model could still be lacking in public values, despite its good performance. Integrity refers to concerns about auditability, privacy, explainability, as well as agency or sovereignty over central modeling choices.

---

[5] Pugh (1991, p. 10), however, lists "efficacy" as one of the main content values of the bureaucratic ethos.

[6] When "efficiency" is defined as a ratio of output to input that is to be maximized, then this implies both efficacy (i.e. increase "output") as well as cost-savings (i.e. decrease "input").

One important dimension of integrity is explainability. When AI is used for decision-making, explainability requires that this use of AI is both scrutable (in some sense) as well as intuitive. ML models often fail on both counts (Selbst & Barocas, 2018). This distinction is important: an explanation might be correct but hard to understand for anyone but experts. The question of what makes a "good" explanation thus raises substantive, difficult, and relevant issues of ethics, philosophy, and policy (Danks, Forthcoming).

Explainability is an important value for intrinsic as well as for instrumental reasons. Intrinsically, the idea that a government agency can explain its decisions, even when they are made by AI, is rooted in democratic theory (Binns, 2019). Instrumentally, explainability, firstly, could help prevent future mistakes — assuming that it allows that causes of administrative errors are recognized, understood, and rectified faster (Young, Himmelreich, Bullock, et al., 2021). Secondly, explainability is instrumentally valuable insofar as it allows to document reasons for which decisions were reached. In this way, explainability instrumentally promotes the public value of procedural due process or rule of law (Jørgensen & Bozeman, 2007; Nabatchi, 2018). Thirdly, explainability is necessary to protect citizens' means for self-advocacy (Vredenburgh, 2022). Finally, explainability might be a central tool for AI governance (Danks, Forthcoming).

Fourth, eligibility determinations for UI should also be non-discriminatory. This is reflected in the value of *equity*. We operationalize "equity" using two well-known definitions of statistical fairness. On one definition, *error parity*, fairness is achieved just in case the rate of underpayments is the same across all protected groups. On another definition, *predictive parity*, fairness is achieved just in case the probability that an individual, who was determined to be ineligible for UI, is in fact ineligible should be the same across all protected groups.

This, in sum, gives us a public values framework for a normative evaluation of UI. Efficiency and efficacy arguably conceptualize what motivated the creation of the UI system, what informs its legal background, what shapes how UI is administered, and what the public expects of UI. Integrity and equity are crucial public values that have their origin variably in constitutional law, administrative law, deliberative democracy, theories of citizenship, and, again, legitimate expectations towards UI.

2.4 Conflicts Between Public Values of Unemployment Insurance

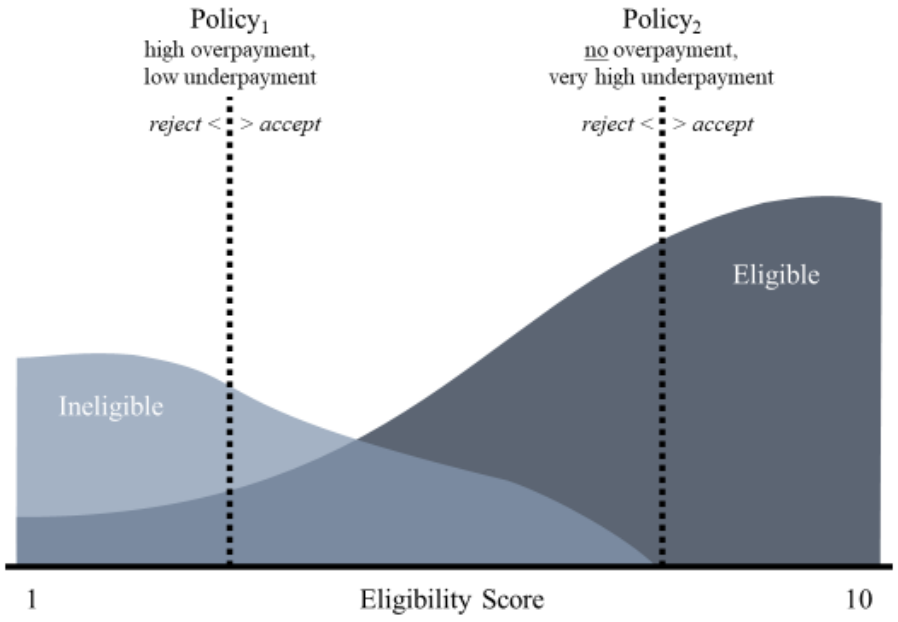
Conflicts between public values are a foundational topic in public administration (Van der Wal et al., 2011). The public values of UI can conflict insofar as they require policy makers and administrators to make choices that further one value at the expense of another and *vice versa*. For example, the application process could be automated further to reduce costs — in line with the value that we call *cost* above — but this might diminish the quality of service that claimants experience and hence be a reduction of what we call above the value of *opportunity*. This is a clear conflict of public values in the first stage of the claims process.

2.4.1 The Efficiency–Efficacy conflict in claims decisions

Efficiency and efficacy likewise conflict in the second stage, when claims are decided. Efficiency demands to avoid *over*payment, whereas efficacy demands to avoid *under*payment. An increase in one value leads to a decrease in the other. This conflict arises because claim eligibility is hard to measure. Eligibility is estimated and mistakes happen. Workforce agencies estimate whether a submitted claim is eligible, but these estimates are not perfectly accurate. We call their estimates the "eligibility score."

The public values conflict consists in the fact that, in terms of this eligibility score, the distributions of claims that are in fact eligible and those that are not, overlap. This is a form of uncertainty. In the worst case, for any given eligibility score, a claim may be eligible, or it may not be. To deal with this uncertainty, workforce agencies hence need to decide whether to accept a claim, given a certain eligibility score. Where should they draw the line so that all claims with an eligibility score that is below this point are rejected and all claims with an eligibility score that is above this line are accepted? Figure 1, adapted from Young et al. (2021) illustrates this conflict.

Figure 1. Illustrated conflict between avoiding over- and under-payments by decision policy.



This conflict between efficiency and efficacy arises even when no eligibility scores are used explicitly. Classical statistical hypothesis testing teaches that efforts to reduce the odds of doing the wrong thing (a Type I error) generally increase the odds of not doing the right thing (a

Type II error). It is therefore reasonable to assume that efforts to prevent improper overpayments correspondingly make it more likely that improper underpayments will occur.

This conflict can play out along multiple causal pathways. For example, fraud might be reduced by requiring claimants to prove having qualified dependents in triplicate instead of a single source. This requirement may cause some recipients to become ineligible even though they are, in fact, eligible. Underpayment errors would result on the margin for the appropriate dependent allowance, or in full if the recipient's claim is placed on administrative hold pending determination or if a claimant elects to not to complete their application or appeal in light of the additional documentation required.

It should be noted that different errors are associated with vastly different practical results. Failing to detect an underpayment is different from failing to detect an overpayment, both from ethical and economic perspectives. From the claimants perspective, failing to detect an underpayment is usually worse than failing to detect an overpayment. The practical consequences of not receiving claims because an agency mistakenly determined them to be ineligible are severe (De La Garza, 2020; Eubanks, 2018).

This conflict between efficiency and efficacy is relevant for two reasons. First, the conflict can inform an analysis of the legal and policy history. We argue in the next section that social insurance programs in the US tend to focus on efficiency. Second, the conflict motivates our investigation into the use of AI to identify administrative errors, which we undertake in the subsequent sections. Avoiding under- and over-payment is immediately relevant to automatic classification (because these two values are related to avoiding false negatives and positives respectively). AI could be a way of partially overcoming the conflict between efficiency and efficacy.

2.4.2 Conflicts of Integrity and Equity

Two further conflicts within these public values should be mentioned — very briefly since in the literature on the ethics of AI, both are considered rather general and well-known tradeoffs (Andrus et al., 2021; Kleinberg et al., 2016, 2018). First, equity and privacy (a dimension of integrity) might conflict. To check whether a system upholds equity, one might need to reduce privacy. Specifically, to check that a system does not discriminate based on protected characteristics—such as gender, age, sexual orientation, or race—one needs data on these protected characteristics, which, depending on the nature of the data, can be a significant privacy risk.

Second, there is a public values conflict internal to *equity*. The two dimensions of equity, *error parity* and *predictive parity*, are themselves in tension with one another under realistic conditions (Kleinberg et al., 2016). This is a so-far under-investigated and un-addressed challenge for equity in welfare systems. Some contend that this conflict can be seen as a conflict between procedural fairness (such as equal treatment, here: predictive parity) and reasons of justice (here: error parity), similarly to how affirmative action furthers justice at the expense of procedural fairness (Vrendenburgh, Forthcoming).

2.5 Empirical Context

Unemployment Insurance in the United States was created by the 1935 Social Security Act. Employees who meet eligibility criteria can apply to receive these benefits if their current employment tenure ends. It is a devolved program; the federal government covers overhead costs and provides broad-based oversight, while State governments finance and implement the program individually. This devolution also applies to eligibility and benefit amount criteria, which can vary substantially across States and within any given State over time. Such changes are frequent, because UI is both a primary tool for countercyclical economic stimulus in recessionary periods when unemployment increases, and a highly visible and politicized program at both the state and national level. This combination of institutional variation and high political salience makes UI a highly complex administrative program. Its complexity, in turn, increases the risk of administrative errors — mistakes in determining the eligibility status of a claim and/or the amount of benefit the claimant is entitled to receive. When such errors occur frequently enough in programs that are as large as welfare, Medicare, or UI, the result is a significant misallocation of public money.

2.5.1 Institutional History: Focus on Overpayments

The US federal government began to take a hardline approach on combating administrative errors in UI during the 1970s in a "war on fraud and error" (Brodkin & Lipsky, 1983; Matt & Cook, 1993). As the name of this 'war' would suggest, the political goal of reducing or eliminating UI payments made to individuals who were not only receiving benefits they were statutorily ineligible for, but were actively defrauding the government in the process (Mendeloff, 1977).

The current process for auditing unemployment insurance claims is known as the Benefits Accuracy Measurement (BAM) program. BAM auditors take a statistically representative, randomized sample of all claims submitted to each State unemployment agency and audit this sample to find errors in both claimant data and administrative decisions on eligibility and benefit level determination. Over the past 20 years federal oversight has increasingly focused on the identification and recovery of improper payments — particularly overpayments — made through social insurance policies, including unemployment insurance. As part of this focus Congress has passed several laws producing performance management requirements for State unemployment administrators to demonstrate that they are proactively seeking, identifying, and correcting overpayments in their unemployment insurance programs. However, no such statutory focus exists with respect to the *under*payment of unemployment insurance benefits. This discrepancy recently came to the fore during the COVID-19 induced unemployment surge in the United States during 2020. Individual and collective harms from systematic wrongful denial and underpayment of unemployment insurance benefits proved to be substantial. At the same time, little has been done at the policy level to identify solutions for this problem going forward.

This research tests whether AI systems can help administrators identify all types of administrative errors in unemployment claim decisions — both overpayments and underpayments, as well as any other systematized errors. This is particularly relevant for practitioners because unemployment insurance is such a highly complex and complicated benefit program. The high level of variation between States and within States over time suggests that the underlying error rate of UI administration is high. To the extent that AI-based automation can both increase the scope and scale while decreasing the time and labor costs required to perform such audits, there are strong normative claims for using this technology on efficiency as well as efficacy grounds. But we argue that a more complete normative consideration also requires an empirical assessment of whether such AI systems may also improve our ability to identify and correct underpayments to eligible beneficiaries.

3. Methods

3.1 Data

The dataset used in this analysis was sampled from the Benefits Accuracy Measurement system of the Department of Labor. This dataset contains information about randomly sampled investigations into UI claims: specifically, federally-collected information of improper UI payments. There are 785,159 observations in the final, analysis-ready dataset. Each observation contains information about one unemployment benefits claim made during the years 2002–2018 in the form of 228 features (variables). These features contain personal information of the claimant (date of birth, gender, race, etc), information about the last employment of the claimant (occupation code, salary, etc), information about the interaction between the claimant and agency (how the claim was filed, whether it was submitted on time or not, etc.), as well as other information. Additionally, we include data on state-level policy differences in UI eligibility requirements and benefit determination criteria, merged by State and corresponding year.

3.2 Analytic Strategy

In this work we treat the problem of detecting administrative errors as a classification problem. Classification (sometimes called categorization) is one of the most fundamental problems in machine learning. It assumes that all observations or datapoints (sometimes called data samples) belong to a limited set of categories or classes, and the goal is to predict the category of the datapoint from its features. A simple example of a classification problem is spam detection, in which new incoming emails should be categorized into two categories: “spam” and “non-spam.” Each email has a set of features, such as its content, its length and the words used, as well as from where the email was sent. To solve the classification problem, an algorithm uses these features of past emails together with each email’s known category — spam or non-spam — to train a new algorithm, a model, that describes those features that distinguish spam emails from non-spam emails. This new algorithm is then applied to incoming emails to determine whether they are spam or not.

In this work, we similarly use a supervised learning approach. Such an approach assumes that the categories, denoted by labels (such as “spam” and “non-spam”), are given for the sample, so that patterns between features and labels can be learned. In this work, we use machine learning algorithms to predict administrative errors. Thus, the features in our data describe various pieces of information about UI claims and class labels describe the types of possible errors (e.g., “No error,” “Underpayment,” etc). Different methods are available to train a model with the labeled samples in order to predict labels on the unlabeled instances. Next, we describe the various machine learning approaches that we used and their obtained results.

3.2.1 Overview of Machine Learning Classification

As discussed above, classification in machine learning can be defined as a problem of predicting labels of the data points (data samples). Each data point has a corresponding vector of the features, which describe that sample. For example, for the spam detection problem, possible features include the number of words in the text, specific words (or their combinations) used in the email, url addresses in the email (which can be used to detect suspicious websites), time of the email, etc. Features can appear in different formats: categorical features, which have values drawn from a limited set of possible values (for example, dog vs. cat); numerical features, which take on a numerical value; text features; image-based features; and others. Many machine learning algorithms expect input features to be in numerical form, and various approaches exist to convert non-numerical features into numerical features.

In addition to these features, a machine learning algorithm also needs a subset of data for which samples are labeled with their category or class. Ultimately, the machine learning model will be trained on this subset, and patterns connecting the features to the class label will be extracted. In supervised machine learning, these categories are assumed to be known — for example, they could be annotated manually by humans, or the process of data collection may be such that categories are obtained alongside data, and so on.

These labeled samples are used to train the model. During the training process, the machine learning algorithm tries to infer patterns in the features to distinguish different categories — for example, “spam” emails may have some specific words that are rarely used in the usual emails, or they may be sent from some specific set of domains. In contrast, “non-spam” emails may be more likely to come from the same domain as the recipient of the email, or sent from addresses that have already been replied to by the recipient. The trained model is used to predict categories of non-labeled samples in an automated way. In addition to making predictions about unlabeled data points, ML algorithms can be used to better understand the data. For instance, algorithms can identify those features that are most important for classification, giving the analyst insight into properties distinguishing between classes of data points.

For certain classes of algorithms (e.g., those based on decision trees), feature importance can be extracted directly from the model. For other algorithms, it is possible to infer the most important features through local permutation of the samples (i.e., make slight modifications to feature values and observe how the predicted label changes) (Baehrens et al., 2010). In the case

of longitudinal data, such algorithms may help to identify changes over time. Additionally, as different ML models operate under specific assumptions about structure of the patterns that may exist in the data, even the performance of various algorithms may reveal additional information, leading to better understanding of the dataset.

The choice of a classification algorithm gives rise to public values conflicts. Insofar as different machine learning algorithms are better or worse at enabling an increased understanding of the data or of the classification, as described here, the algorithms fulfill the public values of integrity, especially its dimensions of audit and explanation, to different degrees. Some algorithms might be more accurate at the expense of fewer insights into feature importance. Or some algorithms might be more accurate at the expense of needing more, and perhaps more sensitive data — a potential privacy risk. The choice of a classification algorithm thus may give rise to conflicts between efficacy and efficiency on the one hand and integrity on the other.

3.2.2 CatBoost: Decision Tree-Based Classification

For the sake of performing a thorough analysis, we use several classification algorithms, which are listed in the Experimental Setup section. Here, for purposes of illustration, we describe one of them: the CatBoost algorithm. CatBoost belongs to the family of gradient boosting classifiers. This type of model combines multiple decision trees into a single model. A decision tree is a flowchart-like object in which feature values are used to determine which branch in the flowchart to take, until a prediction is arrived at (Jain & Srivastava, 2013). During construction of the decision tree, the algorithm iteratively selects the feature that is the most useful for predicting the label, and builds a corresponding branch of the tree, thus splitting the dataset according to the selected feature.

CatBoost (and gradient boosting models in general) are based on an important observation: combining predictions of multiple simple models is better than using a single model. An entire class of models builds a set of decision trees such that every new tree is trained to correct errors of previously trained trees. Several models operate under this general framework, including XGBoost (eXtreme Gradient Boosting (Chen et al., 2015)), LightGBM (Ke et al., 2017), AdaBoost (Friedman et al., 2000), and CatBoost.

Although these algorithms often perform similarly, the main advantage of CatBoost over other algorithms from the family of gradient boosting models is that it was created to address a problem known as *prediction shift*. Prediction shift can be defined as a variation of target leakage, which happens when classifiers during training implicitly get access to the information about the label of the sample in the way that is impossible for unlabeled data during prediction. This happens because when trees are trained iteratively, the gradients of the error reveal information about the target variables.

CatBoost addresses prediction shift with ordered boosting (a weighted sampling method). In this work, we use the “MultiClass” optimization from the original CatBoost implementation, which uses a Multiclass Cross-Entropy Loss (log-loss) function given by:

$$MCE = \frac{\sum_{i=1}^{N} \square w_i \log\left(\frac{e^{a_{it_i}}}{\sum_{j=0}^{M-1} \square e^{a_{ij}}}\right)}{\sum_{i=1}^{N} \square w_i},$$

where a_{ij} represents the predicted probability that element i belongs to class j and the w_i values represent weights associated with each element.

Additionally, CatBoost generates new features during the training process. These features are defined as a combination of existing features. As the number of possible combinations of features on the big dataset is huge, CatBoost constructs new features according to a greedy heuristic to approximate the true global optima.

3.3 Experimental Setup

3.3.1. Preprocessing of the dataset

To prepare the dataset for analysis, we first removed all features that could act as a direct proxy to the target variables. Next, for all models except CatBoost, all non-numeric features were converted to a categorical binary representation (one-hot encoding). As this representation is very sparse, some of the features were removed to reduce dimensionality of the data. CatBoost is able to work with some non-numeric (e.g., text features) features by design, assuming their categorical type.

3.3.2. Algorithms tested

We evaluated several different ML classifier algorithms to test for variation in performance against the audit data. These include:

1. **Logistic regression** (LR). This algorithm is simple, but is commonly used for classification problems. While originally designed for the binary classification, it can be used for multi-class classification with one-vs-all scheme (where a different LR model is trained for each class).
2. **Random Forests** (RF). This algorithm is an ensemble-based algorithm which uses multiple decision trees trained on the different subsets of the data.
3. **CatBoost**. This algorithm, described in more detail above, is based on the idea of Gradient Boosting.
4. Several deep learning-based models:
    a. **TabNet** (Arik & Pfister, 2019). This DL model was created by Google for neural network learning on tabular data. It uses sequential attention and, similarly to the Gradient Boosting models, allows insight into structural patterns in the data.
    b. **DeepFM** (Guo et al., 2018). This model was created for use in recommender systems and uses factorization machines together with neural networks. While it

was not created for tabular data specifically, it may be used with this type of data, and we include it for completeness.

c. **WideDeep** (Cheng et al., 2016). This neural network was created by Google for recommender systems. It combines "wide" linear models and "deep" neural networks. Similarly to DeepFM, it can be used for the tabular data.

d. **Deep & Cross Network** (DCN) (Wang et al., 2017). This model is based on the idea of feature crossing (generation of new features based on existing ones). This model also works well for the tabular data.

To evaluate the performance of a model, we need to compare predictions on a subset of data with known labels. This subset of data is usually referred to as a "test" set or "holdout" set. This data should be excluded from the training process, as it is important that the model is not familiar with the testing data.

Two scenarios were considered. In the first scenario the whole dataset was split into two parts — the "training" set and the "test" set, consisting of 80% and 20% of the data, respectively. The "training" set was used for the training of the model and the "test" set was used for the evaluation. In the second scenario, dataset models were evaluated only on one specific year of claims and trained on the data from one (a) or three (b) previous years. This setting is more realistic, as in this case the model is not able to utilize patterns observable in the current year and has to rely only on previous observations.

3.4 Evaluation Metrics

Each sample of the dataset is a UI claim that was randomly selected to be investigated for incorrect payments. There are several different categories that datapoint can belong to: "No error," "Overpayment," "Underpayment," or "Wrong issue." A "No error" label means that the claim was processed without any errors, "Overpayment" denotes a claim with excessively paid benefits, and "Underpayment" labels claims with incorrectly low benefits. "Wrong issue" means that an error unrelated to payment in the claim was made.

Table 1. Audit determination outcomes in the data.

Outcome	Count	Proportion
No Error	629,445	0.807
Overpayment	74,983	0.096
Underpayment	59,080	0.076
Wrong Issue	16,090	0.021

Total	779,598	1

The dataset is highly imbalanced, as shown in Table 1. Most of the claims belong to the "No Error" class. This fact needs to be accounted for during model evaluation. To understand why, suppose that we evaluate the model just by percentage of claims with correctly inferred class. In this case a simple classifier that makes the "No Error" prediction every time would have a relatively high accuracy of 80%, while being absolutely useless at the same time. To address this, one can use the *F1-score* to evaluate accuracy. To define F1-score we need to define some other metrics that are commonly used in evaluation of ML models:

1. Precision: the fraction of the correctly predicted instances from a class over all predicted instances of the class. In other words, precision with respect to a particular class is defined as the number of true positives divided by sum of true positives and false positives:

$$P = TP/(TP+TN).$$

2. Recall: the fraction of the relevant instances which were correctly predicted. This can also be defined as the number of true positives divided by the sum of true positives and false negatives:

$$R = TP/(TP+FN).$$

The F1-score is the harmonic mean of these two measures:

$$F_1 = \frac{2P*R}{P+R}$$

For the multi-class setting there are two ways to compute the F1-score:

1. Precision, recall, and F1-score may be computed separately for each class. The *macro-F1 score* is then computed as the unweighted mean of these class F1-scores.

2. Precision and recall may be computed globally for all classes at the same time. The *micro-F1 score* is computed from this single precision and recall value.

3.5 Addressing Imbalances in the Dataset

As the dataset is very imbalanced, each model tends to predict most of the claims as belonging to the "No error" class. This leads to low recall performance of less frequent classes (i.e., many members of these classes are not identified).

Such an imbalanced dataset is typical for any empirical context that involves the identification of administrative errors, insofar as errors are the exception. How such imbalances

can be addressed is thus a relevant question for any context where AI is used to detect administrative errors.

The issue of low recall performance can be addressed in different ways; for example, by using different weighting schemes or employing techniques aimed to rebalance dataset (undersampling, oversampling, generation of balanced synthetic data). Most of these approaches exhibit a tradeoff between precision and recall. Precision may be sacrificed in order to obtain higher recall.

This tradeoff affects public values. It constitutes another instance of the conflict between efficacy and efficiency. On the one hand, reducing precision for detecting over- and underpayment reduces efficiency. More claims are flagged as administrative errors by mistake, and investigated without finding an error (*cost*). On the other hand, increasing recall for underpayment errors is a matter or efficacy. It means that more of the underpayment errors, which might otherwise go unnoticed, are detected by the algorithm (*avoid underpayment*). At the same time, increasing recall for overpayment, however, increases efficiency again, since more of the true overpayments errors are detected as such (*avoid overpayment*). Given the normative and empirical relevance of imbalanced datasets, we explore different methods, we perform two experiments, described next.

3.5.1 Weighting schemes

The simplest way to increase recall of the minority classes is to change class weights in the model (for those algorithms that support class weights). Intuitively, class weights affect the penalty for misclassification of different classes. The higher class weight is comparatively to the weights of other classes, the higher is the expected penalty for misclassification for this category.

There are multiple different ways to set class weights. In our experiments, we set weights according to the "balanced" heuristics suggested by (King & Zeng, 2001):

$$w_c = N / (\ k \ * |c|) \ ,$$

where w_c denotes the weight of class c, N stands for the size of the dataset, k denotes the number of classes in the dataset, and $|c|$ denotes the number of samples that belong to class c.

3.5.2 Synthetic data

Another common technique for addressing the problem of an imbalanced dataset is to generate synthetic data which follows a similar distribution as samples from the minority classes. This synthetic data is merged into the training set to make proportions of the classes balanced. Ultimately, the model is evaluated on the test set (which does not have any synthetic samples). For this experiment, we used the SMOTENC (Chawla et al., 2011) algorithm for generating synthetic data, which operates under the following general principle:

1. Select a random sample from the least frequent class;

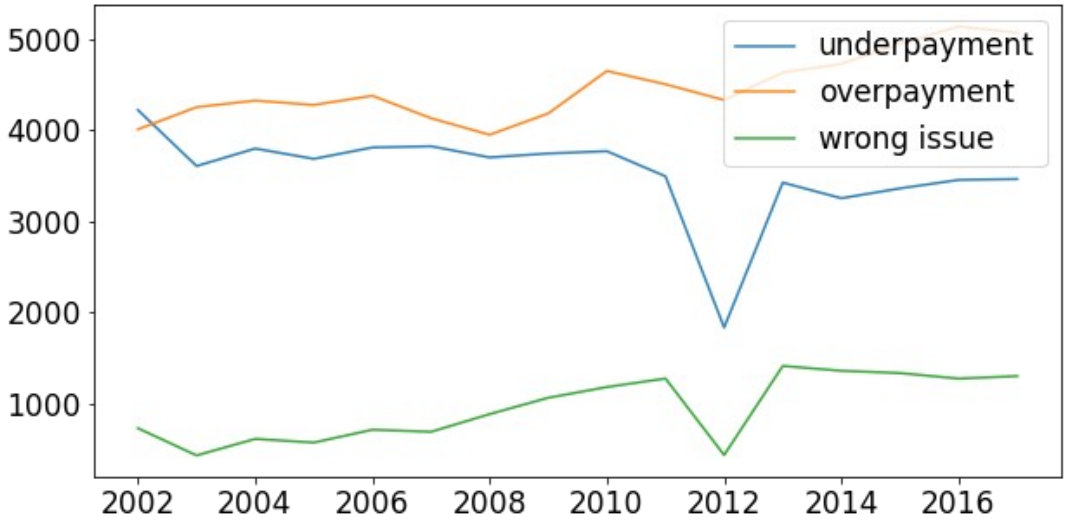
2. Find the *k* most similar samples in the dataset (in our experiments *k* was set to 5; similarity between samples is defined as Value Difference Function (Cost & Salzberg, 1993); and
3. Generate a new sample in the next fashion: sample numeric features between values of random neighbor and the features original example. Set categorical features to be equal to the most common category from the neighbors of the same class.

4. Results

4.1 Descriptive Statistics

Administrative politics and economic situations change over time, which affects patterns of administrative errors as well. Figure 2 shows the trend of increasing overpayment error rates with an approximately 25% increase from 2002 to 2018. The rate of underpayment was slowly decreasing over the period of observation.

Figure 2. Improper payment error trends over the sample period.

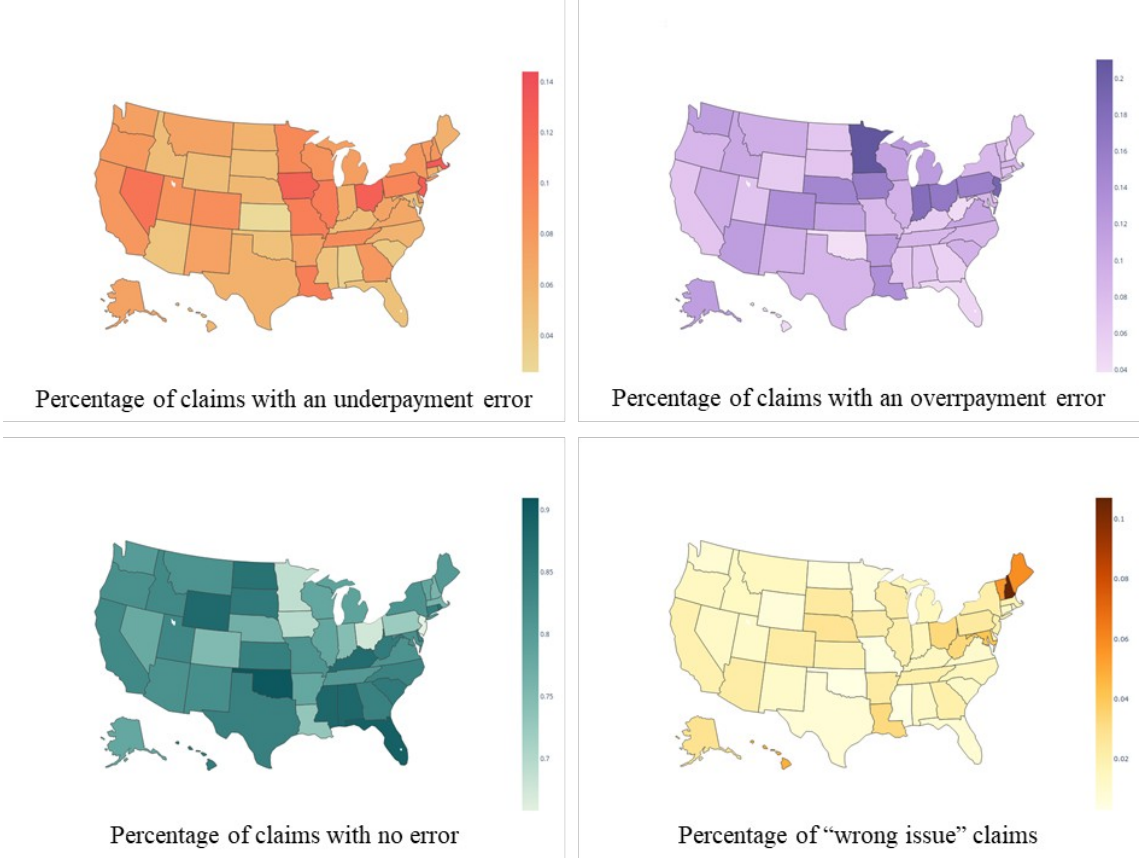


This descriptive statistic can be evaluated with the public values framework outlined above. Insofar as underpayment errors are trending downward and overpayment errors upward over time, the UI system would have improved over time in efficacy (*avoid underpayment*) but worsened in efficiency (*avoid overpayment*).

The error types also vary by State, as seen in Figure 3, which shows aggregated results for all years. These differences may be explainable by differences in State policies. For example, Ohio has a high rate of underpayment errors (one of the highest in the country), but a comparatively low rate of fraud. Non payment-related errors appear to be particularly clustered in the New England region of the Northeast, particularly in New Hampshire. Across all states and outlying territories, the total cost of overpayments across all years of observation is $69

million. Total underpayments are approximately $5 million. This variance might also be indicative of different decisions in resolving public values conflicts.

Figure 3. Relative rate of improper payment errors by type of error across states.



4.2 Classification Analysis

In this section we report our estimates of whether machine learning can be used to predict administrative errors made in processing claims using features of the claim and prior information on administrative error, and utilize trained models to obtain better understanding of the data. Our goal is to predict the class (category) of the claim. Recall from Section 3 that we considered two experimental settings. In the first setting, we combined all years of data and performed a randomized split of the data into training and test sets. Results of this setup can be found in Tables 2 and 3. The best-performing classifier's score for both precision and recall by each class of outcome are bolded. As can be seen, the performance of the CatBoost is the highest, with Random Forest in second place. Logistic Regression failed to generalize and performed poorly.

Table 2. F-Scores (micro/macro) by classifier type by class using data from all time periods.

Classifier	F1-score (macro)	F1-score (micro)

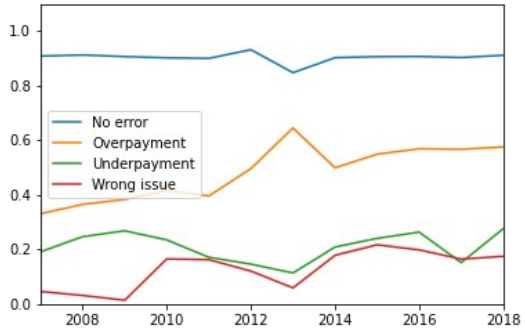
LR	0.230	0.721
RF	0.452	0.805
CatBoost	**0.491**	**0.823**
TabNet	0.366	0.781
DeepFM	0.274	0.730
Wide & Deep	0.271	0.733
DCN	0.384	0.777

Table 3. Recall and Precision by Classifier Type by Class using data from all time periods.

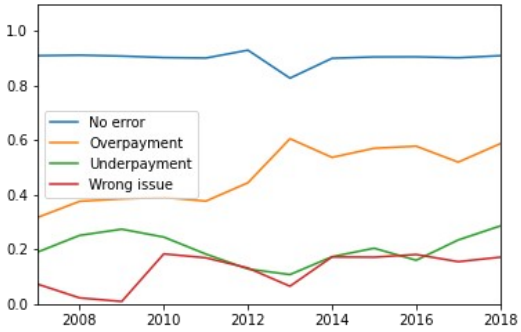
	No errors		**Overpayment**		**Underpayment**		**Wrong Issue**	
Classifier	**Pr.**	**Rec.**	**Pr.**	**Rec.**	**Pr.**	**Rec.**	**Pr.**	**Rec.**
Logistic Reg.	0.807	**0.995**	0.250	0.015	0.061	0.001	0.000	0.000
Random Forest	0.851	0.986	**0.756**	0.374	0.651	0.113	0.702	0.119
CatBoost	**0.865**	0.980	0.739	**0.486**	**0.686**	**0.171**	**0.745**	**0.106**
TabNet	0.842	0.975	0.627	0.404	0.605	0.036	0.500	0.001
DeepFM	0.818	0.952	0.277	0.168	0.071	0.003	0.000	0.000
Wide & Deep	0.815	0.971	0.394	0.122	0.044	0.006	0.000	0.000
DCN	0.835	0.983	0.710	0.290	0.438	0.039	0.475	0.090

In the second experimental setting, the dataset was split in a temporal fashion. The model was trained only on the data from the previous one or three years and evaluated only on the next year. Results for CatBoost (which performed the best) can be found in Figure 5.

Figure 5. F1 scores for CatBoost trained on data from prior year (a) or prior 3 years (b).



(a) F1 score for CatBoost model trained on data from 1 previous year and evaluated on the next year.



(b) F1 score for CatBoost model trained on the data from 3 previous years and evaluated on the next year.

As part of its output, CatBoost is able to provide a ranking of the features based on how important they were to the classification. Features which were identified as especially important can be found in Figure 6. A description of these features can be found in Table 4. Most of these important features can be grouped into one of these three sets:

1. Features which describe the individual's previous occupation, including salary;
2. Features related to time (year of the claim, etc); and
3. Features with information about administrative decisions made prior to the benefit audit.

Figure 6. CatBoost Feature Importance.

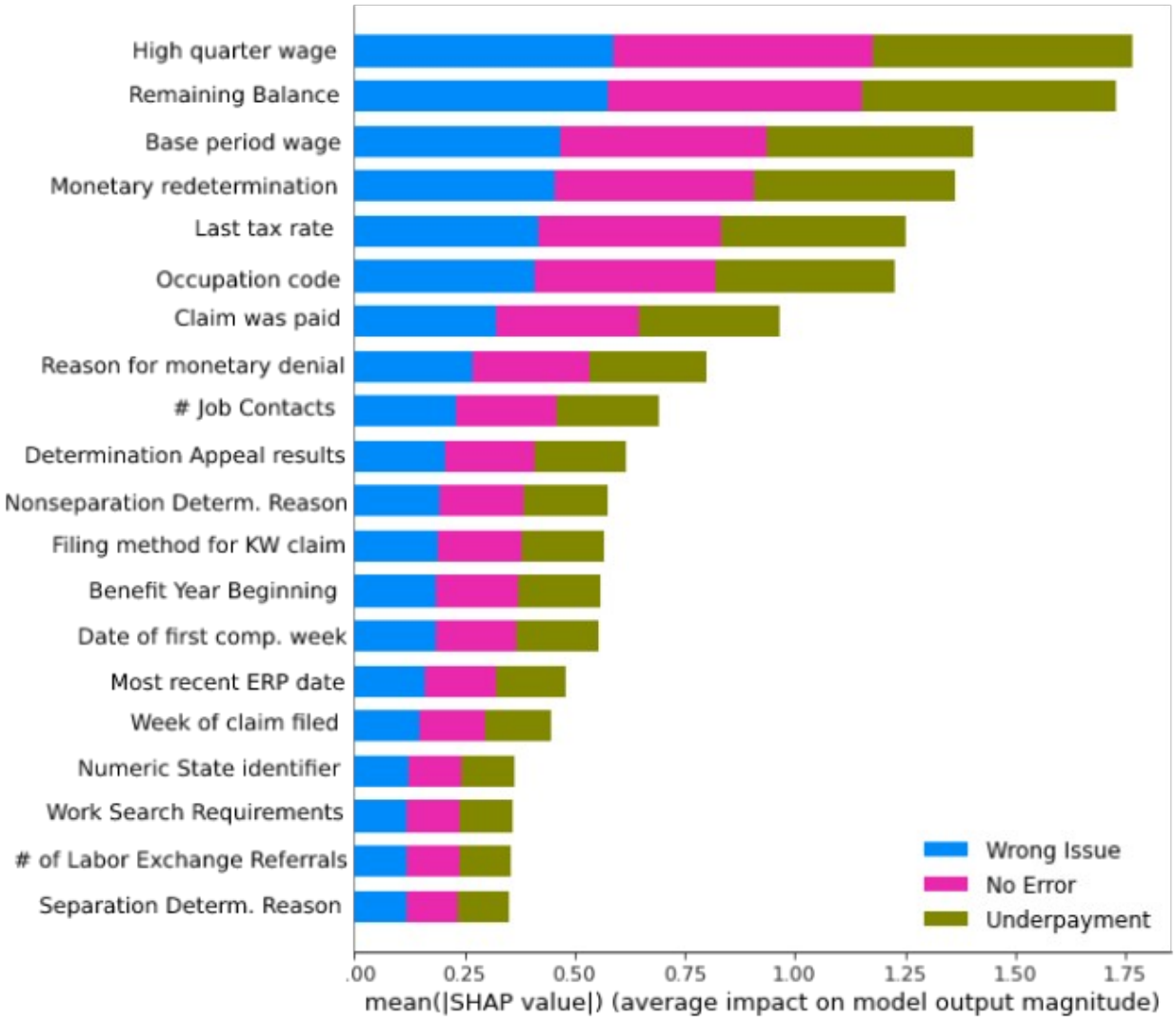


Table 4. Description of important features.

Feature Name	Feature Description
High quarter wages	Highest wages reported in the fiscal quarter before investigation.
Remaining Balance	Remaining Balance (RB) of claim as of key week ending date.
Base period wage	Base period wage before the investigation.
Monetary redetermination	Whether the State redetermined claimant's monetary eligibility.

Last tax rate	Last tax rate for the claimant.
Occupation code	Occupation code of the claimant’s last employment.
Claim was paid	Money was paid for the claim (claim was not denied).
Reason for monetary denial	Reason for monetary denial before investigation.
# Job Contacts	Number of Job Contacts Listed from any source.
Determination Appeal results	Results of Appeal of Initial Determination that denied eligibility.
Nonseparation Determ. Reason	Reason for Nonseparation Determination Before Investigation.
Filling method for KW claim	Filing method for claim
Benefit Year Beginning	Effective date of most recent new or transitional (not reopened or additional) claim.
Date of first comp. wek	Date of first compensable week.
Most recent ERP date	Date of claimant's most recent eligibility review up to Key Week.
Week of claim filed	Week in which claim was filed (beginning date).
Numeric State identifier	State
Work Search Requirements	Subject to Work Search Requirements
# of Labor Exchange Referrals	Number of times Employment Services referred claimant for employment during the current benefit year.
Separation Determ. Reason	Reason for Separation Determination Before Investigation.

4.2 Weighting schemes and synthetic data

Results for CatBoost, trained with the “balanced” heuristic weight scheme can be found in Table 5. There is significant improvement of the recall of minority classes at the expense of precision. At the same time recall of the “No errors” class dropped.

Table 5. Results for CatBoost using ‘balanced’ weighting scheme

	Audit Outcome Class			
Metric	**No Errors**	**Overpayment**	**Underpayment**	**Wrong Issue**
Precision	0.944	0.342	0.236	0.110
Recall	0.532	0.815	0.502	0.773
F-1 score (macro)	0.419			
F-1 score (micro)	0.624			

Results for the model, trained on the mix of real data and synthetic samples can be found in Table 6. Compared to the results for CatBoost reported in Table 3, one can notice a small average boost in recall for the "underpayment" and "wrong issue" classes. From the experimental setup, we can see that the greatest benefit from the synthetic data generation model is for recall on the smallest classes.

Table 6. Results for CatBoost model trained on synthetic data.

	Audit Outcome Class			
Metric	**No Errors**	**Overpayment**	**Underpayment**	**Wrong Issue**
Precision	0.862	0.678	0.554	0.491
Recall	0.967	0.446	0.180	0.150
F-1 score (macro)	0.488			
F-1 score (micro)	0.813			

5. Discussion

This article is motivated by the question of whether public managers could and should use AI to support auditing decision making in social service provision. We believe that the evidence from our experiments suggests a conditional "yes" to both questions. It is clear that even using limited publicly accessible data one can train classifiers that are capable of performing fairly well in identifying potential administrative errors. It would, however, be

unwise and potentially wrongful to use such a system as we have simulated, as a "human-out-of-the-loop" decision automation system. Instead, the focus of our investigation is auditing and quality control: the detection of administrative errors, not the use of AI to make UI claims decisions. It would perhaps be best to consider implementing AI as a basic decision support system that assists human auditors. Instead of relying on random sampling to audit claims decisions, AI could help auditors in identifying the subset of data where they are most likely to find problematic claims. This recommendation of a limited role of AI as a decision support system can also be rooted in the public value of integrity and, more specifically, *agency*, which cautions against a transfer of authority and decision-making power to automated systems (Young et al., 2019).

Our results suggest several implications for the use of artificial intelligence in the public sector. First, the overall classifier performance suggests that CatBoost is, among the set of classifiers we evaluate, the best for the BAM audit data. For now, let us concentrate on the public values of avoiding overpayments (a dimension of efficiency) and underpayments (a dimension of efficacy). CatBoost performs better than other algorithms we tested in either of these dimensions, as operationalized using the accuracy measures.

In some ways the result that CatBoost performs so well is counterintuitive; CatBoost uses decision trees and not the artificial neural networks that have captured most of the media and prior research interest with respect to modern AI. Of particular note is the fact that CatBoost dominated every other classifier with respect to precision and recall for *underpayment errors* as well as recall for overpayment errors, and is relatively closely matched with random forest for overpayment precision. But random forest also performs relatively poorly in most other error types; CatBoost is clearly a better algorithm across all decision outcomes.

The higher performance of the CatBoost in comparison to the DL models is, however, in line with other findings in computer science research. Gradient Boosting-based models (e.g. CatBoost, XGBoost or LightGBM) are known to often outperform neural networks on classifying tabular data (Borisov et al., 2021; Shwartz-Ziv & Armon, 2022). Existing Deep Learning models for the tabular data tend to surpass GBM-based methods mainly on big datasets with predominantly continuous features (Borisov et al, 2021), which is not the case for the BAM dataset.

It is also worth noting that CatBoost's performance tends to improve over time. This could indicate gradual improvement of the administrative procedures, and investigating the causes of this behavior is an avenue for future research. Taken as a whole, this is further evidence that public managers looking to implement AI in their organizational decision-making need to pay particularly close attention to the full spectrum of decision outcomes and their possible implications.

However, there is a risk that, if the public value of efficacy is not kept in view, an AI classification algorithm might be chosen that focuses only on efficiency, and might even be

better than CatBoost in this regard. Specifically, in our empirical context, if decision makers were only focusing on overpayments when evaluating model performance, they might be inclined to select an algorithm — or proprietary, "black box" system developed by a private vendor — that would lead to substantive underperformance for identifying other types of administrative errors.

Next, CatBoost also has an added normative benefit from the perspective of the public value of integrity, especially in the dimensions of audit and explainability. Because CatBoost is able to provide some clarity on the relative weighting or importance of different variables or features within the data, public managers, politicians, and the public all have a chance to make more informed decisions about whether the use of such a automated classifier is both of sufficient value to justify its use and does not make tradeoffs that violate either legal or normative obligations. CatBoost thus performs well in terms of the public values of auditability and explainability.

Furthermore, our experiments also provide novel evidence for researchers and practitioners with respect to different ways of addressing data set imbalance and the related public values conflicts. Table 7 reports the relative changes in performance metrics across different outcome classes as well as for micro- and macro-F1 scores for the two alternative training methods employed: changes to the weighting schema and the use of synthetic data. The difference between each of these alternatives is striking. Adjusting the weighting schema produced strong substantive changes particularly with respect to the recall capabilities for overpayments but also underpayments and wrong issue errors. However, this came at significant cost to precision for overpayments underpayments and wrong issue errors and this penalty is further evident in the harmonic mean scores for F1 macro and F1 micro. This happens because the algorithm is attempting to avoid misclassification of less frequent classes, and thus acts with more "suspicion." These results hence quantify the public values conflict between efficiency and efficacy described earlier.

Table 7. Percent Change in CatBoost Performance Relative to Naive Training on All Data

Alternative Training Method	**Metric**	**Outcome Class**			
		No Errors	**Overpayment**	**Underpayment**	**Wrong Issue**
Weighting Schema	Precision	9%	-54%	-66%	-85%
	Recall	-46%	68%	194%	629%
	F-1 score (macro)	-15%			
	F-1 score (micro)	-24%			
Synthetic Data	Precision	0%	-8%	-19%	-34%

	Recall	-1%	-8%	5%	42%
	F-1 score (macro)	-1%			
	F-1 score (micro)	-1%			

For the synthetic data approach, however, the differences are still present but significantly attenuated. In particular, the improvements to recall for underpayments as well as for wrong issues are significantly reduced but still positive. On the other hand, the synthetic data approach led to an overall reduction in both precision and recall for overpayment errors — although this change is significantly smaller in magnitude than for changes to the weighting schema. Similarly, both the F1 macro and F1 microscores are reduced as they are with the weighting schema changes but significantly less so in terms of magnitude. Another important point is that the simulated or synthetic data, in addition to producing some modest performance gains in recall for underpayments and wrong issue errors, also have an important role to play from the point of enacting privacy preserving methods for training public sector AI classifiers.

As a whole our results suggest that there are substantive impacts to be had from making different design choices with respect to model training parameters and correcting for highly imbalanced training data. This is particularly salient for public sector use of AI for decision making. The choices at hand are technical — e.g. how to adjust the weights in the training algorithm to address the issue of imbalanced data? — but the choices also concern public values. Since this kind of conflict will arise for any highly imbalanced dataset, and since all observational data about administrative errors is usually highly imbalanced, and since AI systems will likely be trained on these data, our experiments highlight the importance of critically attending to technical decisions: *Seemingly technical decisions in model selection and training carry substantive implications for public values*.

This also calls to attention the nature of the data used in this analysis. These data are very noisy across different classes. One potential explanation for this noise is a limitation to this article: we might not have enough information in the publicly accessible unemployment insurance audit data to have a true understanding of the underlying data generative process (DGP). Underpayment errors in particular are very difficult to distinguish from no error classes in our sampling data. These errors may have no systemic cause; they may just be a product of a stochastic underlying error rate in the UI claims process. An alternative explanation is that the data generative process that stems from the BAM audit rules and how they systematically differ between searches for overpayment and as particularly fraudulent errors and underpayments may in fact systematically bias the audit data towards finding overpayment errors at the expense of underpayments. The effect of historical embedding of different normative, political priorities in public sector DGPs on present-day efforts to implement AI in these contexts is an under-examined problem in the public administration and policy literature, making it a strong candidate for future research.

It is important to note that this dimensionality limitation, owing to the public facing nature of our training data, may also be artificially limiting the performance of neural net-based classifiers such as TabNet. Future research that can leverage secure micro level data employed by state workforce agencies in their more thorough audits and reviews of unemployment insurance claims would be a necessary and useful next step to continue to improve our understanding of the value and potential of AI-based decision support in public sector environments.

6. Conclusion

In this paper, we have investigated the potential of Artificial Intelligence (AI) to identify administrative errors in the empirical context of Unemployment Insurance (UI) claims decisions. This investigation is motivated by the joint questions: Can AI be used to detect administrative errors and, if so, should it? Our answer is a qualified "yes."

We compared the performance of seven different random forest and deep learning models, we addressed the highly imbalanced data using weighting schemas and synthetic data, and we evaluated the models using different accuracy measures, as well as a public values framework. We found that CatBoost, a random forest model using gradient descent boosting, is more accurate along several measures, and preferable in terms of public values — since its classifications are not only more accurate across different error types but are also in a limited sense explainable — than every deep learning model we tested.

Future research should seek opportunities to overcome the data limitations we faced in this work. This would require partnerships with government agencies to secure access to restricted data not available to the public (or researchers). Additionally, future extensions could experiment with additional and more sophisticated optimization techniques, e.g. identifying pareto frontiers in multi-objective optimization. Finally, there is a need for additional empirical research on how AI is already used or being prepared for implementation by state workforce agencies.

Detecting administrative errors in UI, and prominently in public organizations and welfare systems generally, gives rise to a conflict between efficiency and efficacy. When it comes to AI, this conflict bottoms out in tradeoffs of recall and precision in AI training methods. When the objective function is to minimize *all* error types — not just politically salient ones such as avoiding overpayment — AI can help to overcome this conflict. Our recommendation is that AI can be used to support audits of administrative decisions. AI can be used to prioritize quality control to detect and reduce administrative errors.

This article provides insights on what problems need to be addressed along the way and which AI methods can and should be used. Given the limits of our analysis, we provided a public values framework that can support deliberations about the use of AI to detect administrative errors beyond the confines of the analyses of this article. There has been a historical and institutional emphasis on minimizing overpayments to the frequent exclusion of investments in

making UI more accessible and timelier for those who qualify. Using AI to audit claims data both highlights the need to explicitly balance the desire for efficiency with a need for efficacy and can serve as a tool towards that end.

References

Alshahrani, A., Dennehy, D., & Mäntymäki, M. (2021). An attention-based view of AI assimilation in public sector organizations: The case of Saudi Arabia. *Government Information Quarterly*, 101617. https://doi.org/10.1016/j.giq.2021.101617

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete Problems in AI Safety* (arXiv:1606.06565).

Andrus, M., Spitzer, E., Brown, J., & Xiang, A. (2021). What We Can't Measure, We Can't Understand: Challenges to Demographic Data Procurement in the Pursuit of Fairness. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 249–260. https://doi.org/10.1145/3442188.3445888

Arik, S. O., & Pfister, T. (2019). *TabNet: Attentive Interpretable Tabular Learning*.

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., & Mueller, K.-R. (2010). How to Explain Individual Classification Decisions. *Journal of Machine Learning Research*, *11*, 1803–1831.

Bauder, R. A., & Khoshgoftaar, T. M. (2017). Medicare fraud detection using machine learning methods. *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 858–865.

Binns, R. (2019). Human Judgement in Algorithmic Loops; Individual Justice and Automated Decision-Making. *Individual Justice and Automated Decision-Making (September 11, 2019)*.

Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., & Kasneci, G. (2021). Deep neural networks and tabular data: A survey. arXiv preprint arXiv:2110.01889.

Brodkin, E., & Lipsky, M. (1983). Quality control in AFDC as an administrative strategy. *Social Service Review*, *57*(1), 1–34.

Bullock, J. B. (2019). Artificial Intelligence, Discretion, and Bureaucracy. *American Review of Public Administration*, *49*(7), 751–761.

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

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research*, *81*, 15.

Byrnes, N. (2016, June 24). Why We Should Expect Algorithms to Be Biased. *MIT Technology Review*. https://www.technologyreview.com/2016/06/24/159118/why-we-should-expect-algorithms-to-be-biased/

Charette, R. (2018, January 24). *Michigan's MiDAS Unemployment System: Algorithm Alchemy Created Lead, Not Gold - IEEE Spectrum*. IEEE Spectrum: Technology, Engineering, and Science News. https://spectrum.ieee.org/riskfactor/computing/software/michigans-midas-unemployment-system-algorithm-alchemy-that-created-lead-not-gold

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2011). *SMOTE: Synthetic Minority Over-sampling Technique*.

Chen, T., He, T., Benesty, M., Khotilovich, V., & Tang, Y. (2015). Xgboost: Extreme gradient boosting. *R Package Version 0.4-2*, 1–4.

Cheng, H.-T., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., Anderson, G., Corrado, G., Chai, W., Ispir, M., Anil, R., Haque, Z., Hong, L., Jain, V., Liu, X., & Shah, H. (2016). Wide & Deep Learning for Recommender Systems. *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, 7–10. https://doi.org/10.1145/2988450.2988454

Cost, S., & Salzberg, S. (1993). A WEIGHTED NEAREST NEIGHBOR ALGORITHM FOR LEARNING WITH SYMBOLIC FEATURES. *Machine Learning*, *10*(1), 57–78.

Courtland, R. (2018). Bias detectives: The researchers striving to make algorithms fair. *Nature*, *558*, 357–360. https://doi.org/10.1038/d41586-018-05469-3

Criado, J. I., Valero, J., & Villodre, J. (2020). Algorithmic transparency and bureaucratic discretion: The case of SALER early warning system. *Information Polity*, *25*(4), 449–470.

Danks, D. (Forthcoming). Governance via Explainability. In J. B. Bullock, Y.-C. Chen, J. Himmelreich, V. M. Hudson, A. Korinek, M. M. Young, & B. Zhang (Eds.), *Oxford Handbook of the Governance of AI*. Oxford University Press.

de Bruijn, H., Warnier, M., & Janssen, M. (2021). The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. *Government Information Quarterly*, 101666. https://doi.org/10.1016/j.giq.2021.101666

De La Garza, A. (2020, May 28). *States' Automated Systems Are Trapping Citizens in Bureaucratic Nightmares With Their Lives on the Line*. Time. https://time.com/5840609/algorithm-unemployment/

Doberstein, C., Charbonneau, É., Morin, G., & Despatie, S. (2021). Measuring the Acceptability of Facial Recognition-Enabled Work Surveillance Cameras in the Public and Private Sector. *Public Performance & Management Review*, 1–30. https://doi.org/10.1080/15309576.2021.1931374

Eggers, W. D., Schatsky, D., & Viechnicki, P. (2017). *AI-augmented Government: Using cognitive technologies to redesign public sector work*. Deloitte Center for Government Insights.

Engstrom, D. F., Ho, D. E., Sharkey, C. M., & Cuéllar, M.-F. (2020). *Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies*. Administrative Conference of the United States.

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

Farbmacher, H., Löw, L., & Spindler, M. (2020). An explainable attention network for fraud detection in claims management. *Journal of Econometrics*. https://doi.org/10.1016/j.jeconom.2020.05.021

FBI National Press Office. (2020, July 6). *FBI Sees Spike in Fraudulent Unemployment Insurance Claims Filed Using Stolen Identities* [Press Release]. Fbi.Gov. https://www.fbi.gov/news/pressrel/press-releases/fbi-sees-spike-in-fraudulent-unemployment-insurance-claims-filed-using-stolen-identities

Flügge, A. A., Hildebrandt, T., & Møller, N. H. (2021). Street-Level Algorithms and AI in Bureaucratic Decision-Making: A Caseworker Perspective. In *Proc. ACM Hum.-Comput. Interact.* (Vol. 5, Issue CSCW1, p. Article 40). Association for Computing Machinery.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *The Annals of Statistics*, *28*(2), 337–407. https://doi.org/10.1214/aos/1016120463

Fukumoto, E., & Bozeman, B. (2019). Public Values Theory: What Is Missing? *The American Review of Public Administration*, *49*(6), 635–648. https://doi.org/10.1177/0275074018814244

Guo, H., Tang, R., Ye, Y., Li, Z., He, X., & Dong, Z. (2018). *DeepFM: An End-to-End Wide & Deep Learning Framework for CTR Prediction*.

Heckman, J. (2020, March 4). AI as 'ultimate auditor?' Congress praises IRS' adoption of emerging tech. *Federal News Network*. https://federalnewsnetwork.com/artificial-intelligence/2020/03/ai-as-ultimate-auditor-congress-praises-irss-adoption-of-emerging-tech/

Huang, H., Kim, K.-C. (Casey), Young, M. M., & Bullock, J. B. (2021). A matter of perspective: Differential evaluations of artificial intelligence between managers and staff in an experimental simulation. *Asia Pacific Journal of Public Administration*, 1–19. https://doi.org/10.1080/23276665.2021.1945468

Ingrams, A., Kaufmann, W., & Jacobs, D. (2021). In AI we trust? Citizen perceptions of AI in government decision making. *Policy & Internet*, *n/a*(n/a). https://doi.org/10.1002/poi3.276

Jain, N., & Srivastava, V. (2013). Data mining techniques: A survey paper. *IJRET: International Journal of Research in Engineering and Technology*, *2*(11), 2319–1163. https://doi.org/10.15623/ijret.2013.0211019

Janssen, M., Brous, P., Estevez, E., Barbosa, L. S., & Janowski, T. (2020). Data governance: Organizing data for trustworthy Artificial Intelligence. *Government Information Quarterly*, 101493. https://doi.org/10.1016/j.giq.2020.101493

Janssen, M., & Kuk, G. (2016). The Challenges and Limits of Big Data Algorithms in Technocratic Governance. *Government Information Quarterly*, *33*, 371–377.

Jørgensen, T. B., & Bozeman, B. (2007). Public Values: An Inventory. *Administration & Society*, *39*(3), 354–381. https://doi.org/10.1177/0095399707300703

Juan Liu, Eric Bier, Aaron Wilson, John Alexis Guerra-Gomez, Tomonori Honda, Kumar Sricharan, Leilani Gilpin, & Daniel Davies. (2016). Graph Analysis for Detecting Fraud, Waste, and Abuse in Healthcare Data. *AI Magazine*, *37*(2). https://doi.org/10.1609/aimag.v37i2.2630

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, *30*, 3146–3154.

Kim, N., & Hong, S. (2021). Automatic classification of citizen requests for transportation using deep learning: Case study from Boston city. *Information Processing & Management*, *58*(1), 102410. https://doi.org/10.1016/j.ipm.2020.102410

King, G., & Zeng, L. (2001). Logistic Regression in Rare Events Data. *Political Analysis*, *9*(2), 137–163. Cambridge Core. https://doi.org/10.1093/oxfordjournals.pan.a004868

Kingson, E. R., & Levin, M. (1984). Local administrative practice and AFDC error in Maryland. *Journal of Social Service Research*, *7*(3), 41–57.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). HUMAN DECISIONS AND MACHINE PREDICTIONS. *The Quarterly Journal of Economics*, *133*(1), 237–293. https://doi.org/10.1093/qje/qjx032

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *ArXiv Preprint ArXiv:1609.05807*.

Le, Q. V., Ranzato, M., Monga, R., Devin, M., Chen, K., Corrado, G. S., Dean, J., & Ng, A. Y. (2013). *Building High-level Features using Large Scale Unsupervised Learning*. 8595–8598.

Levy, K., Chasalow, K., & Riley, S. (2021). Algorithms and Decision-Making in the Public Sector. *Annual Review of Law and Social Science*, *17*(1), 309–334. https://doi.org/10.1146/annurev-lawsocsci-041221-023808

Matt, G. E., & Cook, T. D. (1993). The war on fraud and error in the food stamp program: An evaluation of its effects in the Carter and Reagan administrations. *Evaluation Review*, *17*(1), 4–26.

Mendeloff, J. (1977). Welfare procedures and error rates: An alternative perspective. *Policy Analysis*, 357–374.

Mikalef, P., Lemmer, K., Schaefer, C., Ylinen, M., Fjørtoft, S. O., Torvatn, H. Y., Gupta, M., & Niehaves, B. (2021). Enabling AI capabilities in government agencies: A study of determinants for European municipalities. *Government Information Quarterly*, 101596. https://doi.org/10.1016/j.giq.2021.101596

Nabatchi, T. (2018). Public Values Frames in Administration and Governance. *Perspectives on Public Management and Governance*, *1*(1), 59–72. https://doi.org/10.1093/ppmgov/gvx009

O'Neil, C. (2017). *Weapons of Math Destruction: How big Data Increases Inequality and Threatens Democracy*. Broadway Books.

Pallasch, J. (2020). *Addressing Fraud in the Unemployment Insurance (UI) System and Providing States with Funding to Assist with Efforts to Prevent and Detect Fraud and Identity Theft and Recover Fraud Overpayments in the Pandemic Unemployment Assistance (PUA) and Pandemic Emergency Unemployment Compensation (PEUC) Programs* (Unemployment Insurance Program Letter No. 28-20, p. 17) [Memo]. Employment and Training Administration Advisory System. https://wdr.doleta.gov/directives/attach/UIPL/UIPL_28-20.pdf

Perrault, R., Shoham, Y., Brynjolfsson, E., Clark, J., Etchemendy, J., Grosz, B., Lyons, T., Manyika, J., Mishra, S., & Niebles, J. C. (2019). *The AI Index 2019 Annual Report* (A. I. S. Committee, Trans.). Human-Centered AI Institute, Stanford University.

Pugh, D. L. (1991). The origins of ethical frameworks in public administration. *Ethical Frontiers in Public Management*, 9–34.

Russell, S. J., & Norvig, P. (2015). *Artificial Intelligence: A Modern Approach* (Third). Pearson India.

Selbst, A. D., & Barocas, S. (2018). The Intuitive Appeal of Explainable Machines. *Fordham Law Review*, *3*, 1085–1140.

Stone, D. (2011). *Policy Paradox: The art of political decision making* (3rd ed.). W.W. Norton & Company.

Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. Information Fusion, 81, 84-90.

The Economist. (2020, June). Technology Quarterly: Artificial intelligence and its limits. *The Economist*, *435*, 19–22.

van der Voort, H. G., Klievink, A. J., Arnaboldi, M., & Meijer, A. J. (2019). Rationality and politics of algorithms. Will the promise of big data survive the dynamics of public decision making? *Government Information Quarterly*, *36*(1), 27–38. https://doi.org/10.1016/j.giq.2018.10.011

Vredenburgh, K. (2022). The Right to Explanation*. *Journal of Political Philosophy*, *30*(2), 209–229. https://doi.org/10.1111/jopp.12262

Vrendenburgh, K. (Forthcoming). Fairness. In J. B. Bullock, Y.-C. Chen, J. Himmelreich, V. M. Hudson, A. Korinek, M. M. Young, & B. Zhang (Eds.), *Oxford Handbook of the Governance of AI*. Oxford University Press.

Wang, R., Fu, B., Fu, G., & Wang, M. (2017). Deep & Cross Network for Ad Click Predictions. *Proceedings of the ADKDD'17*. https://doi.org/10.1145/3124749.3124754

Wirtz, B. W., & Müller, W. M. (2019). An integrated artificial intelligence framework for public management. *Public Management Review*, *21*(7), 1076–1100. https://doi.org/10.1080/14719037.2018.1549268

Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2019). Artificial Intelligence and the Public Sector—Applications and Challenges. *International Journal of Public Administration*, *42*(7), 596–615. https://doi.org/10.1080/01900692.2018.1498103

Wirtz, B. W., Weyerer, J. C., & Sturm, B. J. (2020). The Dark Sides of Artificial Intelligence: An Integrated AI Governance Framework for Public Administration. *International Journal of Public Administration*, *43*(9), 818–829. https://doi.org/10.1080/01900692.2020.1749851

Witten, I., Frank, E., Hall, M., & Pal, C. (2016). *Data Mining: Practical Machine Learning Tools and Techniques* (4th ed.). Morgan Kaufmann.

Young, M. M., Bullock, J. B., & Lecy, J. D. (2019). Artificial Discretion as a Tool of Governance: A Framework for Understanding the Impact of Artificial Intelligence on Public Administration. *Perspectives on Public Management and Governance*, *2*(4), 301–313. https://doi.org/10.1093/ppmgov/gvz014

Young, M. M., Himmelreich, J., Bullock, J. B., & Kim, K.-C. (2021). Artificial Intelligence and Administrative Evil. *Perspectives on Public Management and Governance*, *4*(3), 244–258. https://doi.org/10.1093/ppmgov/gvab006

Young, M. M., Himmelreich, J., Honcharov, D., & Soundarajan, S. (2021). The Right Tool for The Job? Assessing the Use of Artificial Intelligence for Identifying Administrative Errors.

In *DG.O2021: The 22nd Annual International Conference on Digital Government Research* (pp. 15–26). Association for Computing Machinery. https://doi.org/10.1145/3463677.3463714