Beyond Ads: Sequential Decision-Making Algorithms in Law and Public Policy

Peter Henderson
Stanford University
Stanford, California, USA

Brandon Anderson
Stanford University
Stanford, California, USA

Ben Chugg
Stanford University
Stanford, California, USA

Daniel E. Ho
Stanford University
Stanford, California, USA

ABSTRACT
We explore the promises and challenges of employing sequential decision-making algorithms – such as bandits, reinforcement learning, and active learning – in law and public policy. While such algorithms have well-characterized performance in the private sector (e.g., online advertising), the tendency to naively apply algorithms motivated by one domain, often online advertisements, can be called the “advertisement fallacy.” Our main thesis is that law and public policy pose distinct methodological challenges that the machine learning community has not yet addressed. Machine learning will need to address these methodological problems to move “beyond ads.” Public law, for instance, can pose multiple objectives, necessitate batched and delayed feedback, and require systems to learn rational, causal decision-making policies, each of which presents novel questions at the research frontier. We discuss a wide range of potential applications of sequential decision-making algorithms in regulation and governance, including public health, environmental protection, tax administration, occupational safety, and benefits adjudication. We use these examples to highlight research needed to render sequential decision making policy-compliant, adaptable, and effective in the public sector. We also note the potential risks of such deployments and describe how sequential decision systems can also facilitate the discovery of harms. We hope our work inspires more investigation of sequential decision making in law and public policy, which provide unique challenges for machine learning researchers with potential for significant social benefit.

CCS CONCEPTS
• Applied computing → Law; E-government; • Computing methodologies → Sequential decision making; Active learning settings; Batch learning.

KEYWORDS
Sequential Decision-making, Reinforcement Learning, Bandits, Active Learning, AI and Society, Law and AI, Responsible AI

1 INTRODUCTION
Sequential decision-making (SDM) algorithms have been explored under multiple (overlapping) paradigms, such as bandits, reinforcement learning (RL), and active learning. Together, these have been successfully applied to content recommendation and ad placement [35, 98, 143], clinical trials [15, 55], robotics [135], and control of power systems [127, 137]. The core challenge of SDM is carefully balancing the explore-exploit trade-off.

Resource-constrained problems commonly faced in law and public policy, particularly the public sector, might seem to be a natural but under-explored application for SDM algorithms. Should government agencies focus their resources more on taking advantage of historical information (“exploitation”), or gathering new information (“exploration”)? Too much of the former can result in missing opportunities and trends (e.g. discovering new tax shelters or missing emerging pandemic risk), while too much of the latter results in wasted resources. Sometimes, agencies already use what is effectively an SDM system without a formalization of these trade-offs. Formalization and improved methods may help to address the efficiency, transparency, and accountability of the process.

We present a survey of potential applications and challenges of deploying SDM algorithms in the public sphere.1 The primary audience for this paper is machine learning researchers and engineers. Our central thesis is that off-the-shelf deployment of existing SDM models to the public sector will not work, due to distinct challenges that law and public policy pose. Our primary purpose is to highlight these limitations, and provide motivation to the machine learning and computer science communities to work on these frontier problems. To this end, Section 4 provides examples of distinct problems faced by various public agencies that may be solved, or partially addressed, with further research into SDM algorithms.

We emphasize that SDM systems, especially those that interact with sensitive data, should be deployed with care. The purpose of this paper is not to suggest SDM algorithms as a one-size-fits-all solution to resource-allocation problems in the public sector. To the contrary, our work spells out the limitations of naively applying methods developed in other domains. Researchers and practitioners may be tempted to take methods that worked well in online advertising, for example, and apply them without modification to government services. We call this “the advertisement fallacy.” Algorithm designers are not the only ones susceptible to the advertisement fallacy. By procuring off-the-shelf algorithmic solutions, public agencies have likewise fallen victim to the same error (and been criticized for it [112]). We argue that for effective uses in the public sector, machine learning will need to evolve to move beyond ads by solving a number of unique technical challenges. That said, we also illustrate some of the promises of SDMs and we aim to spur more research on specific problems given the potential to increase

1We focus primarily on governance in the United States due to the expertise of authors, but most if not all of our examples have parallels in countries and regions globally.
the efficiency, transparency, and adaptability of government. We hope that by discussing specific problems faced by various agencies, and taking the first steps towards formalizing these problems in the language familiar to computer scientists, we can encourage the development of algorithms suited specifically for these challenges. Our contributions then are:

- A call to shift machine learning research “beyond ads” and to law and public policy, where the social gains may be substantial;
- A survey of potential SDM applications in law and public policy, across a wide range of areas, including public health, environmental protection, tax administration, occupational safety, and social welfare;
- An enumeration of distinct methodological and technical challenges posed by the public sector that render off-the-shelf application of SDM inappropriate, including multi-objective decision making, batched and delayed feedback, corrupted labels, feedback loops, and rationality requirements; and
- An initial formalization of such challenges to motivate further machine learning research to move beyond ads.

The remainder of the paper is organized as follows. Section 2 discusses the motivations of using SDM algorithms for law and public policy. Section 3 provides some brief background on the problem formulation. Section 4 then provides examples of problems faced by various agencies which are potential applications for SDM algorithms and highlights open methodological issues that the machine learning community can address to render such techniques policy-compliant, effective, and more widely applicable. We emphasize that while these examples are meant to illustrate interesting research avenues for SDM algorithms, they are not meant to indicate that SDMs will certainly solve these problems. Instead, the call is for machine learning researchers to expand the scope of problems they consider when developing algorithms, and to lay some of the foundations for formalizing some of the problems encountered in the public sector. Section 5 addresses concerns regarding the use of these algorithms, and provides some guidelines to discern when they are appropriate.

2 MOTIVATION

As noted, our article highlights frontier challenges that the machine learning research community must address to overcome the advertisement fallacy and address difficult technical challenges faced by the public sector. The examples are meant to illustrate interesting research avenues for SDM algorithms, and not meant to indicate that SDMs will certainly solve these problems. Instead, the call is for ML researchers to expand the scope of problems they consider when developing algorithms, and to lay some of the foundations for formalizing some of the problems encountered in the public sector. These problems demonstrate the broad range of challenges and needs posed by the public sector (e.g., improved data collection). Before we embark on that discussion, however, we hope to briefly discuss several reasons why administrative agencies might want to use SDM algorithms in the first place.

Many public policy processes are already de facto SDM algorithms, without formalization. Throughout the subsequent parts of this work, we will provide a number of examples to demonstrate this. We offer several motivations for why formalizing these processes as SDM algorithms may be welfare enhancing, but note that these must be weighed against potential harms (see Section 5).

**Efficiency.** Improving efficiency of government processes can improve welfare. By formalizing SDM processes, governments can identify bottlenecks in the delivery of important services, identify potential corruption and fraudulent outflows, and improve trust in government through transparency. For example, it has been suggested that investments in information technology at the Internal Revenue Service could lead to 10-to-1 returns on investments by recovering underreported tax payments [134].

Consider the case of enforcement prioritization at the Environmental Protection Agency (EPA). The EPA must allocate investigative resources to determine who might be in violation of regulations. This naturally requires an explore-exploit trade-off. Without formalization, the EPA may consider allocation via random sampling, or via some other implicit algorithm (perhaps a mix of investigator intuition and other factors). By formalizing the process, the EPA may be able to allocate investigative resources more optimally (e.g., by reducing the amount of random sampling or identifying useful sources of information).

**Transparency.** Formalization of processes as SDM algorithms allows for increased transparency, if deployed properly. The data gathered to formally verify SDM models can identify dangerous feedback loops and biases in existing systems. And government oversight organizations (such as the Office of Inspector General) can leverage this data to assess and adjust government services.

**Adaptive Policymaking.** It is often not clear whether to deploy an algorithm in a new context. Algorithmic impact assessments have been suggested as a solution to the problem of how and if algorithms should be deployed, aiming to foster reflection about the risks of adoption [109]. They have been adopted by several governmental organizations, such as the EU [85] and Canada [105]. Such reflection may rule out the deployment of algorithms with inherent harm. However, as noted by Moss et al. [111], there is disagreement concerning when algorithmic impact assessments should be conducted. Some advocate for ex ante assessments, i.e., occurring before the adoption of the proposed system, while others stress the importance of ex post reviews,2 which take place after the system has been implemented. The former emphasize that systems adopted without heavy scrutiny could cause avoidable harm, while the latter point out that insufficient information exists at the time of the adoption decision. Ex ante assessments, in other words, require predicting the consequences of algorithmic decision-making. Costs and benefits are not precisely known.

SDM algorithms can help narrow the divide between the ex post and ex ante camps. Because they are sequentially implemented, they enable real-time tailoring to adapt their behavior. For instance, we might increase the amount of random exploration conducted by the algorithm, or reduce the weight of previous time period which is unrepresentative of the larger population. Much like in the case of adaptive clinical trials – where sequential assignment enables researchers to limit the number of patients assigned to

2 A recent example is Facebook’s study of its role in Myanmar [151].
We also consider a basic hypothetical example and then modify $i$ $X$ At each time $t$ ($\neq t_1$) These actions then yield rewards. Typically, information pertaining to each project, including the results of any other agency which must plan the maintenance and rehabilitation of public infrastructure. Each time period, say yearly in this case, other agency, the agency has a limited budget with which to inspect and perform an audit, or the number of food safety violations at a restaurant by inspection. A policy decides which set of $m_t$ actions to take, with actions typically requiring selecting $m_t$ observations. These actions then yield rewards. Typically, $m_t \ll N_t$ though the set of observations and selection budget may change size over time ($N_t \neq N_{t+1}$ and $m_t \neq m_{t+1}$).

To clarify the presentation of the three main approaches to sequential decision-making under consideration – active learning, bandits, and reinforcement learning – we introduce the running example of civil infrastructure management. Consider a city or bandits, and reinforcement learning – we introduce the running paradigm [136]. The goal of active learning is to maximize accuracy of the underlying model while making as few queries as possible. Given a pool of unlabeled observations ($X_t$) (infrastructure projects) the policy would decide which label ($r_{it}$) to reveal (which inspection to perform) to maximize the model’s performance across the pool. Thus the reward is the reduction in generalization error

$$\text{reward}(t) \equiv \sum_{t \in T} \mathbb{E} \left[ \sum_{i \in S_t} r_{it} \right] \quad (1)$$

and

$$\text{regret}(t) \equiv \sum_{t \in T} \mathbb{E} \left[ \sum_{i \in S_t^*} r_{it} - \sum_{i \in S_t} r_{it} \right] \quad (2)$$

where $S_t$ is the (possibly singleton) set of actions we take at time $t$, and $S_t^*$ is set of actions with the highest total reward. The expectation is taken over any random choices made by the algorithm, as well as over stochastic rewards. In the infrastructure example, reward might be whether, or to what extent, the project requires maintenance. In this case, regret measures how many sites were visited that don’t require timely attention.

### 3.1 Bandits

A natural goal for the agency is to allocate infrastructure inspections to those projects that are most in need of them. Perhaps the most natural framework in this case is the multi-armed bandit formulation. Here, there are $N$ “arms” to choose from, each associated with its own reward distribution. At each time step, we choose an arm (or set of arms given the batched setting, see e.g., Perchet et al. [122]). In our example, arms might be specific projects, or they could be specific classes of projects (e.g., bridges, pipes, office buildings, apartments, hospitals, etc.).

If incorporating prior knowledge – observations – of each project (or class of projects), the problem can be cast as a batched, contextual bandit, in which the observations comprise the context.3

If we additionally assume that rewards can be approximated as a functional form of observations, we can interpret the problem naturally as a structured bandit [12, 108]. In the bandit setting, one of potentially multiple objectives (see Section 4.1) is maximizing cumulative reward over time (equivalently, minimizing cumulative regret). Cumulative reward and cumulative regret at time $t$ are defined as

$$\text{reward}(t) \equiv \sum_{t \in T} \mathbb{E} \left[ \sum_{i \in S_t^*} r_{it} - \sum_{i \in S_t} r_{it} \right] \quad (2)$$

where $S_t$ is the (possibly singleton) set of actions we take at time $t$, and $S_t^*$ is set of actions with the highest total reward. The expectation is taken over any random choices made by the algorithm, as well as over stochastic rewards. In the infrastructure example, reward might be whether, or to what extent, the project requires maintenance. In this case, regret measures how many sites were visited that don’t require timely attention.

### 3.2 Active Learning

Another possible objective in our example is to train a fixed model to predict maintenance requirements while meeting a strict inspection budget. This falls more naturally into an active learning paradigm [136]. The goal of active learning is to maximize accuracy of the underlying model while making as few queries as possible. Given a pool of unlabeled observations ($X_t$) (infrastructure projects) the policy would decide which label ($r_{it}$) to reveal (which inspection to perform) to maximize the model’s performance across the pool. Thus the reward is the reduction in generalization error

3Alternatively, observations might be grouped along shared features, in which case we can treat the problem as a contextual bandit in which arms are groups of observations [142].
from selecting a given arm. In some active learning formulations (like continuous active learning), the goal can instead be framed as successfully identifying all the labels of a given type [47]. In this case \( r_{it} = 1 \) if \( y_{it} \) is of the desired type.

### 3.3 Reinforcement Learning

If the agency would like to develop an explicit future maintenance schedule then reinforcement learning is an appropriate paradigm. RL enables us to develop a policy which minimizes future costs,\(^4\) and is particularly valuable when there is an underlying model for how various “states” relate to one another.

We define the reinforcement learning problem as an infinite-horizon discounted Markov Decision Process (MDP) [19, 124]. An MDP is defined as \( M = (S, \mathcal{A}, \mathcal{R}, T, \beta, \gamma) \) where \( S \) is the state space, \( \mathcal{A} \) is the action space, \( \mathcal{R} : S \times \mathcal{A} \rightarrow \mathbb{R} \) is the reward function, \( T : S \times \mathcal{A} \rightarrow \Delta(S) \) is the transition function, a kernel mapping state-action pairs to a probability distribution over \( S \), \( \beta \in \Delta(S) \) is the initial state distribution, and \( \gamma \in (0, 1) \) is the discount factor. In the applications we study, more often than not, multiple actions will need to be taken per time-step and the transition function will not necessarily be known \emph{a priori}. The state space will be constructed from \( X_i \), the set of arms available for selection. Furthermore, in many cases the true problem formulation may be better suited for the partially observable MDP framework [31], as the state space is unlikely to be fully observable. Unlike the bandit setting which maximizes myopic reward, the RL setting maximizes the expected discounted future return at each timestep,

\[
V^\pi(s) = \mathbb{E}\left[ \sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) | s_0 = s \right].
\]

Madanat [104] casts the infrastructure management problem as a MDP, where the states correspond to a project’s condition, and transition probabilities capture deterioration over time and the effect of performing various kinds of maintenance. Andriotis and Papakonstantinou [8] use RL to optimize a reward function that would minimize total discounted future costs, including cost of providing maintenance versus costs from deterioration.

### 4 APPLICATIONS & OBSTACLES

Here we discuss various applications of SDM in law and public policy. Along with the examples, we highlight specific problems to

| Framework          | Goal                                                      | Example Use Case                                           |
|--------------------|-----------------------------------------------------------|------------------------------------------------------------|
| Bandit             | Minimize regret by selecting rewarding arms               | Allocate inspections to buildings most in need of repair    |
| Active learning    | Minimize generalization error by selecting examples which most improve the model | Construct, at lowest cost, a maximally accurate model of building inspection need |
| Reinforcement learning \((0 < \gamma < 1)\) | Maximize future discounted reward by constructing explicit policy | Learn an optimal maintenance schedule by minimizing future maintenance costs |

Table 1: A stylized breakdown of the three frameworks and their goals, along with an example use case from infrastructure management.

\(^4\)We note that contextual bandits can be considered a form of reinforcement learning, but for simplicity we distinguish between the two here.
| Technical Challenge | Public Policy Example | Example Organization(s) | Example Legal and Policy Constraint(s) |
|---------------------|----------------------|-------------------------|---------------------------------------|
| Multi-objective decision-making | eDiscovery: maximizing the retrieval of responsive documents in discovery, while estimating the prevalence of responsive documents | U.S. Courts | Federal Rules of Civil Procedure Rule 26(g) (creating "reasonable inquiry" standard for the production of documents during discovery, often taken to mean a high recall rate for retrieval models [68, 70]). Internal Revenue Code § 7602; Internal Revenue Manual 1.2.1.5.10 (stating policy around use of "statistics indicating the probability of substantial error"); Improper Payments Elimination and Recovery Act (requiring a "statistically valid estimate" of improper payments); Office of Management and Budget, Requirements for Payment Integrity Improvement (spelling out methodological considerations for estimation) |
|                      | Identifying Tax Misreporting: maximizing discovery of tax evasion while estimating population prevalence of income misreporting | Internal Revenue Service | |
| Batched and delayed feedback | Anti-corruption: anticorruption audits can be risk-selected, but feedback can be non-randomly delayed by years | Brazilian Controladoria Geral da União (CGU) | Law №10683 of 28 May 2003 (Brazilian law creating the CGU which conducts anticorruption audits) |
| Distribution shifts in the small data regime | Public health pandemic response: incorporating exogenous and endogenous changes in disease rates into predictions | Local Public Health Departments, U.S. Centers for Disease Control and Prevention | Executive Order 13,994 (requiring agencies to investigate new data collection mechanisms and data-driven responses to public health threats); Executive Order 13,996 (establishing a board to allocate resources for testing, while prioritizing at-risk communities); [89] (discussing additional legal constraints). Food Safety and Modernization Act § 201 (Regulations regarding the methods for "Targeting Of Inspection Resources For Domestic Facilities, Foreign Facilities, and Ports Of Entry") |
|                      | Food Safety: maximizing the discovery of food safety violations while estimating the prevalence of non-compliance in import inspections | U.S. Food and Drug Administration | |
| Learning with and identifying corrupted labels | Occupational safety: Allocating inspectors to visit facilities to ensure occupational health and safety requirements are met | Occupational Safety and Health Administration | Occupational Health and Safety Act, 29 U.S.C. §§ 657, 667 (providing for inspections and requiring that state enforcement must be at least as effective as federal enforcement); Administrative Procedure Act, 5 U.S.C § 706 (providing that a reviewing court shall set aside agency action that is arbitrary and capricious); [74, 110] (discussing inconsistencies in inspection process and state enforcement) |
| Feedback Loops | Environmental compliance: Determine which facilities to inspect, and the method of inspection, to ensure environmental regulations are upheld | U.S. Environmental Protection Agency | Clean Water Act § 308 (providing for inspection authority); National Compliance Initiative (aiming for reduction in overall noncompliance rate); [67] (documenting data gaps to monitor overall levels of compliance) |
| Rational and Causal Decision-Making | Disability adjudication: Determining the subset of claims likely to be approved so that they can be flagged for speedy review | Social Security Administration | 20 CFR §§ 404.1619, 416.1019 (providing for quick disability determination if there is a "high degree of probability that the individual is disabled"); Motor Vehicle Mfrs. Ass’n v. State Farm Mutual Automobile Ins. Co., 463 U.S. 29 (1983) (articulating standard for whether agency has considered the relevant factors and reached a reasoned decision) |
| Validation | Wildfire prevention targeting: Identify forest regions to actively manage, including conducting prescribed burns, to reduce wildfire impacts. | U.S. Department of the Interior | John D. Dingell, Jr. Conservation, Management, and Recreation Act (2019); Executive Order 13,855: Promoting Active Management of America’s Forests, Rangelands, and Other Federal Lands To Improve Conditions and Reduce Wildfire Risk |

Table 2: A subset of the examples discussed throughout the paper, listed alongside (one of) their core technical challenges and any legal requirements which place constraints on the problem. We note that other legal constraints also apply to all of the presented algorithms, including privacy and non-discrimination standards. This table is meant as a non-exhaustive illustration of the distinct challenges posed by law and public policy.
Population estimation is related to so-called active exploration, which has been studied both in bandits [9, 28] and in reinforcement learning and MDPs [140, 144, 146]. Active exploration is focused on efficiently exploring an unknown space to make accurate predictions about unlabelled observations. Erraqabi et al. [59] study how to trade-off active exploration with reward maximization in the multi-armed bandit setting. Developing algorithms to trade off these two objectives in other settings remains an open problem.

Agencies may have multiple secondary objectives. For instance, educators wish both to determine the effectiveness of various educational policies, but also maximize students’ learning in the meantime [100]. That is, spending resources on exploring new policies might disrupt the classroom and hamper learning. Attempts to handle generic multi-objective problems often involve extending the reward scalar to a vector [53], and searching for Pareto optimal solutions. While such approaches are useful for generic problems, they discard the possibility of taking advantage of specific objectives that agencies commonly face.

Example 4.1 (Optimizing tax audits at the IRS). Every year, the Internal Revenue Service (IRS) audits between 0.5% and 1.1% of individual taxpayer returns, out of a potential pool of hundreds of millions [134]. In addition to maximizing revenue, they must generate reliable population estimates (the “tax gap”). Moreover, these estimates are often subject to accuracy standards set by the Office of Management and Budget (OMB) [118].

**Formulation.** This was recently modeled as a structured bandit problem [73]. Each year the IRS selects a batch of taxpayers to audit from the population based on their factorized tax returns $x_{it}$. The reward received, $r_{it}$, is the difference between the taxpayer reported amount of taxes owed and the true adjusted amount after audit. The reward is related to taxpayer’s attributes $x_{it}$ via some unknown function $f$ and parameter $\theta_i$: $r_{it} = f(x_{it}, \theta_i)$. One objective is to maximize reward over time, i.e., Equation (1). Another is to estimate the average adjustment each year, $\bar{\mu} = \sum_i r_{it}$.

There are several ways in which we might formalize the objective. One is to minimize a convex combination of these objectives, e.g.,

$$\min_{S_1, \ldots, S_r} \left( \text{regret}(t) + \alpha \sum_r \left| \hat{\mu}_r - \mu_r \right| \right),$$

where $\alpha$ is some normalizing parameter which weights the importance of accurate population estimation against regret. A separate approach is to treat it as a constrained optimization problem in which we maximize reward subject to limits on the expected percent difference and/or the variance, in adherence with OMB guidelines for precision of estimates:

$$\max_{S_1, \ldots, S_r} \text{reward}(t) \quad \text{s.t.} \quad |\mathbb{E}[\hat{\mu}_r] - \mu_r|/\mu_r \leq \beta, \quad \forall (\hat{\mu}_r) \leq \kappa \quad \text{for all } r \leq t,$$

for some $\beta, \kappa > 0$.

Example 4.2 (Technology Assisted Review during civil litigation). eDiscovery, short for electronic discovery, entails identifying relevant documents during the discovery process in legal proceedings. Since the body of potential evidence is often overwhelming, current state-of-the-art eDiscovery mechanisms use active learning to identify relevant documents [37, 47, 70]. The use of technology assisted review is especially common in antitrust litigation by the Department of Justice and Federal Trade Commission. We also note that identifying documents for FOIA requests uses a similar process in some cases.

**Formulation.** The policy sees a set of documents (arms) $X_t$, with $x_{it}$ consisting of standard natural language processing features (e.g., TF-IDF vectors [82]). The policy selects a set of documents for lawyers to review and receives a reward if the document was responsive ($r_{it} = 1$) or non-responsive ($r_{it} = 0$). This is repeated across multiple rounds until there is confidence that all responsive documents have been identified. Here we’re trying to minimize the number of non-responsive documents, i.e., minimizing $\sum_i r_{it} - \sum_{i \in S_t} r_{it}$, while simultaneously estimating $\mu_t$. In current implementations, a random sampling mechanism is often employed to estimate $\mu_t$ [97]. Combining these two processes into one multi-objective system might help to reduce labeling costs.

4.2 Batched and Delayed Feedback

In contrast to online advertisements, it is rare to witness the result of each action before having to make subsequent decisions in law and policy. Instead, decision-making must be batched and could potentially be based on outdated information. This leads to ask whether we can incorporate such features into SDM algorithms. There has been work on both batched and delayed feedback in the bandit setting [60, 64, 77, 92], as well as the active learning setting [41, 71]. In multi-armed bandits in particular, concept drift is sometimes modelled as discrete, abrupt changes in the mean reward of the arms [4, 107], sometimes called switching bandits.

Unfortunately, such algorithms typically make unrealistic assumptions, such as knowing an a priori bound on the amount of drift or the total number of time steps in order to optimally partition the actions in pre-defined time blocks [32, 132, 158]. The existing work on batched feedback in bandits suffers similar drawbacks [64, 122].

In law and policy settings, delays can be stochastic, biased, and dependent on arm features, relationships which aren’t considered in prior work. For example, an audit of a small facility will likely be finished more quickly than a large one. Variance in delays can also be heteroskedastic: more complicated audits could have a wider range of delays. Moreover, concept drift and batched actions have not been studied in conjunction with one another, whereas they often appear together in policy settings.
4.3 Distribution Shifts in the Small Data Regime

The distributions of both policy-relevant features and their relationship with reward are rarely fixed in reality—behaviors adapt, policies change, exogenous economic shocks occur, etc. This concept drift emphasizes the exploration aspect of SDM, requiring careful selection just to keep up. Algorithms in public policy settings are doubly hampered as the amount of available samples is often small, potentially leaving little budget for reliably rewarding selections.

Recent work in both the bandit and active learning spaces focuses on either detecting concept drift or designing algorithms which are robust to it [32, 90, 142]. Such work, however, develops algorithms that are purely reactive. That is, they do not anticipate concept drift. Meanwhile, in law and public policy settings, there is often specific domain information which can act as an a priori indicator of concept drift (e.g., a statutory revision). Incorporating this structured drift information into existing models could improve resilience and efficiency, particularly in small-data regimes. More research is needed to create reliable methods for incorporating such structural priors or external sources of information into SDM policies. Moreover, it is an open question as to which algorithms retain performance and fairness guarantees under such large discontinuous sources of drift.

The scarcity of data in such conditions makes deploying reinforcement learning algorithms more difficult. RL algorithms often require a large amount of data to function properly, with many successful real-world applications relying on simulations to train them [18, 50]. Some efforts have been made to make agent-based simulations for public policy problems [149, 159]. But it is unclear whether such simulations have enough accuracy or fidelity to prove useful for training RL policies in simulation and applying them to real public policy or law tasks.

Example 4.3 (Auditing local government corruption at the CGU). In 2003, the government of Brazil introduced the Controladoria Geral da Unão (CGU), an autonomous federal agency to conduct anti-corruption audits of local governments [61]. While the CGU audited local governments at random, Ash et al. [10] demonstrated that selection informed by tree-based classifier would catch nearly twice as much corruption. Auditing, however, will change the subsequent behavior of departments and can take years to complete. Therefore, all audits for the next fiscal year have to be chosen before receiving any feedback and will deal with outdated information. Time to completion for an audit will also likely correlate with the local government’s size and composition. The heterogeneity of the delay schedule might affect algorithmic biases if not corrected for appropriately.

**Formulation.** Each year, the policy can select a set of local governments to audit given features consisting of tax revenue, size of government, history of previous corruption, etc. The reward consists of successfully identified sources of corruption, but may take several years (time steps) to return and update the policy. In particular, when an audit is conducted it takes time \( d_{it} \sim D_t \) to receive reward \( r_{it} \), where \( D \) is some unknown delay distribution. While previous work assumes a single delay distribution, or one that varies with time (e.g., Bistritz et al. [22] and Vernade et al. [150]), here the distribution is most likely affected by the government itself. We can model this as a family of distributions \( \mathcal{D} = \{ D_\theta \} \) parameterized by latent variables \( \theta \). In addition to identifying corruption, one should learn the parameters \( \theta \) which give rise to delays and the model must account for censoring by marginalizing over audits that have not yet cleared. To complicate things further, we must select all audits in year \( t \) as a batch, without receiving feedback from any of them. We might also expect partial feedback of the audits, in which we receive some signal each year until we receive the true result [69].

Example 4.4 (Allocating public health resources during a pandemic). Infectious disease outbreaks can cause severe resource allocation issues. Effective distribution of tests and vaccines are of paramount importance for curtail the spread of the disease. But conditions in this setting drift rapidly: one population could quickly become vaccinated and less vulnerable, or a new variant could quickly make another population more susceptible to infection. Could the incorporation of external information like trends from other regions help adjust to drift quickly?

**Formulation.** Recently, this problem has been formulated as a multi-armed bandit [40] whereby testing resources were allocated to different neighborhoods—each constituting an arm—rewarded by the number of COVID-19 positive individuals identified. Testing of incoming populations at the border has been similarly formulated as an RL problem, leading to more efficient detection of potentially infectious persons [16]. However, we might augment such strategies by incorporating the output of disease models into the predicted reward. Traditional bandit strategies such as Thompson sampling [133] or UCB sampling [12] predict subsequent rewards \( \hat{r}_t \) as some function of the history of past rewards \( \hat{r}_t = f(\mathcal{M}_t) \). But a disease model \( M(\hat{\theta}) \) can forecast the effects current case rates on future case rates, and take into account possible effects of new strains and policy changes. In this way, predicted rewards can be anticipative instead of reactive, i.e., \( \hat{r}_t = f(M(\hat{\theta})) \) where parameters \( \hat{\theta} \) are learned from \( \mathcal{M}_t \). Further, if \( \hat{\theta}_i \) are the model parameters of location \( i \) then, in this context, the explore-exploit tradeoff becomes a tradeoff between approximating the true parameters \( \theta_i \) versus allocating resources the most at-risk areas, i.e., that \( i \) was where \( M(\hat{\theta}_i) \) makes the most severe predictions (i.e., the health equity objective noted above).
4.4 Learning with and Identifying Corrupted Labels

In many law and public policy settings there is often a “human in the loop,” responsible for providing the labels or collecting the reward (e.g., auditors, inspectors, judges). This can inject variance and subjectivity into the information received by the algorithm, which at best can make learning difficult and at worst can make an algorithmic approach infeasible (see Section 5). This leads to a general warning, illustrated by the example below: when we receive information from heterogeneous and subjective sources, algorithmic design must pay particular attention to the signal-to-noise ratio. In some public sector settings, noise may swamp the signal and make naive machine learning a poor vehicle for solving the problem [84].

There has been a substantial amount of work on learning in the presence of corrupted data and noisy labels, mostly in the classification setting work on active learning with corrupted labels [36, 113, 152, 154], but more recently in regression as well [34]. Such work, however, generally ignores the structure of the noisy labels, attempting to only learn the true distribution. In policy settings, however, the structure of the corrupted data may be of interest. We may want to learn the noisy distribution, develop algorithms to determine when a noisy label is likely to occur, or differentiate multiple sources of noise from one another.

4.5 Feedback Loops

SDMs must be designed to avoid runaway feedback loops—a mis-calibrated algorithm may focus only on the areas (or companies, individuals, etc.), at the expense of learning more about others. Ensign et al. [58], for instance, show that predictive policing algorithms could repeatedly focus on the same neighbourhoods, even when underlying crime rates are random. Similar effects have also been observed in health care and recommender systems [2, 80, 141]. Recent work has also proposed methods to tackle feedback loops in the context of the bank loan problem [120]. Open research questions remain. The most basic is perhaps the question of how to formalize

---

**Example 4.5 (Inspecting food safety at the FDA).** The U.S. imports nearly 15 million yearly shipments of food. The Food Safety Modernization Act (FSMA) provides for inspections of imported food products and American-based food production facilities to ensure compliance with FDA standards. As acts are amended, however, compliance standards change as do the inspection requirements (see the Tester-Hagan amendment to the FSMA [23], for instance) It may be required to report new information, or to report old information in new ways. This causes distribution shift, and may cause problems for algorithms trained on older inspections data.

**Formulation.** Let the arms $X_t$ be the set of facilities or shipments in year $t$, where each observation containing relevant features and operational details. The hidden label $r_t$ for the facility or shipments indicates non-compliance (whether as a binary variable or some measure of the severity). There are various kinds of distribution shift of which we should be wary. The first is simply a change in the relationship of features and labels caused by factors such as price changes, or updated standards in facilities themselves. Models trained to recognize certain attributes as being indicative of unsafe food production might begin to under perform. The second originates from statutory amendments or rule changes, which occur periodically. These can pose a new kind of distribution shift that we might call feature shift: The feature space changes abruptly. A model trained on $X_t$ might not even be well-defined on $X_{t+1}$ (e.g., due to a new reporting standard, or the elimination of features). This raises the question of how to take advantage of the information in $X_t$ for a model which must make predictions on $X_{t+1}$. Feature engineering and dimensionality reductions are options, but lack theoretical guarantees in the SDM setting.

---

**Example 4.6 (Prioritizing inspections).** Recent work has suggested that machine learning could aid in targeting health and safety inspections [11, 66, 81]. Like auditing at other agencies, this process could in theory be improved by SDM. But inspections can be highly stochastic. Recent evidence shows that two health inspectors could give drastically different scores to the same restaurant [73]. Occupational safety and health, mining, nursing home, and nuclear safety inspections each exhibit such inspector variability [74]. As a result, any algorithms using these noisy (or incorrect) labels may allocate more resources to areas where inspectors are most strict, perpetuating biases and undermining regulatory goals. Until SDM algorithms incorporate mechanisms to identify and correct for such noisy labels, such algorithms should not be deployed blindly in inspection contexts.

**Formulation.** Each timestep, we have inspectors $1, \ldots, m$ and facilities or features $x_1, \ldots, x_n$. While each facility $x_i$ admits a true reward $r_i$, we receive a noisy signal $r_i + \epsilon_j$ when inspector $j$ inspects the facility. We attempt to learn both the true rewards and the noise distribution for each inspector. Note that if we are able to select which inspector inspects which facility, this adds a layer of complexity on top of the classic bandit model, as there are now two actions to choose but we receive a single confounded reward. We might incorporate this by introducing more terms into our objective function:

$$\min_{S_1, \ldots, S_N} \text{regret}(t) + \alpha \sum_{j=1}^{m} D_{KL}(\hat{D}_j||D_j),$$

where $D_j$ is the true noise distribution for inspector $j$, $\hat{D}_j$ is our approximate distribution, and $\alpha$ is some normalizing parameter. Here $D_{KL}(D_j||D_2)$ is the KL divergence between distributions $D_1$ and $D_2$. 

---

*a*https://www.fda.gov/media/120585/download  
*b*https://www.fda.gov/media/78021/download
unwanted feedback loops. Jiang et al. [80] define feedback loops in recommender systems as the condition \( \lim_{t \to \infty} \| \mu_t - \mu_0 \| = \infty \), where \( \mu_t \) is the user’s interest at time \( t \). In other words, the user’s interest diverges infinitely from their original preferences. We might change this somewhat for generic multi-armed bandit settings and write \( \lim_{t \to \infty} D_{KL}(D_t^{\mu} || D_t^{\mu_0}) > T \), for some finite threshold \( T \), where \( D_t^{\mu} \) is the award distribution of arm \( a \) at time \( t \), and, as before, \( D_{KL}(D_t || D_0) \) is the KL Divergence between distributions \( D_t \) and \( D_0 \). However, such a definition assumes that the feedback loop affects the reward distribution. This might not be the case, and we give an example of another definition in the box below. Moreover, should one adopt this definition, more needs to be done for the problem formulation. In particular, such a definition is coupled with the goal of maximizing reward may encourage feedback loops if they increase the mean of the reward distribution.

Even after properly defining feedback loops, it’s important to ask how we can best (i) detect feedback loops, and (ii) intervene to avoid them. How can we determine when to use an SDM system if there is a risk for such a feedback loop? Conversely, when does formalization of an existing feedback loop if there exist facilities \( t \) visited at the expense of \( j \). If such that \( j \) pollutes more often than \( i \), and yet \( i \) is visited at the expense of \( j \). This might be formalized as the three conditions

\[
\sum_{r \geq 0} 1(r_{jt} - r_{it} > 0) = \infty,
\]

and

\[
\frac{1}{T} \sum_{t=1}^{T} 1(i \in S_t) \xrightarrow{T \to \infty} 0, \quad \frac{1}{T} \sum_{t=1}^{T} 1(j \in S_t) \xrightarrow{T \to \infty} 0.
\]

where \( 1(\cdot) \) is an indicator. Note that \( S_t \) is a function of the particular SDM algorithm. We can thereby compare the propensity of different algorithms to generate feedback loops.

### Example 4.7 (Checking Environmental Compliance at the EPA)

The federal Environmental Protection Agency (EPA) and state EPAs must decide which facilities to inspect in order to ensure compliance with environmental laws such as the Clean Water Act, Clean Air Act, Safe Drinking Water Act, Toxic Substances Control Act, among others [20]. Most of these inspections involve physical visitations which are time intensive and costly. As a result, only a handful of investigation and testing resources are allocated to various facilities. The same risk exists in this setting as in predictive policing—algorithms might fall into a local maximum whereby they repeatedly allocate inspections to the same facilities.

**Formulation.** Each facility is an arm with features \( X_t \) indicating prior enforcement history, region, sensor metrics, calculated risk based on satellite imagery [39, 155], and \( r_{it} \) is either the amount of unpermitted pollution per facility (regression), or whether the facility was in noncompliance with specific regulations (classification). Let \( S_t \) be the set of facilities inspected at time \( t \). We might identify a positive feedback loop if there exist facilities \( i, j \) such that \( j \) pollutes more often than \( i \), and yet \( i \) is visited at the expense of \( j \). This might be formalized as the three conditions

\[
\sum_{r \geq 0} 1(r_{jt} - r_{it} > 0) = \infty,
\]

and

\[
\frac{1}{T} \sum_{t=1}^{T} 1(i \in S_t) \xrightarrow{T \to \infty} 0, \quad \frac{1}{T} \sum_{t=1}^{T} 1(j \in S_t) \xrightarrow{T \to \infty} 0.
\]

where \( 1(\cdot) \) is an indicator. Note that \( S_t \) is a function of the particular SDM algorithm. We can thereby compare the propensity of different algorithms to generate feedback loops.

### 4.6 Rational and Causal Decision-Making

Law and public policy often impose restrictions on permissible solutions to decision-making problems. Subject matter expertise may also rule out certain solutions. For instance, in the health setting, Caruana et al. [30] discovered that a model trained to predict pneumonia risk learned to associate an asthmatic condition with a lower risk of death. This relationship did exist in the training data, but only because asthmatics diagnosed with pneumonia were treated quickly and thoroughly due to their underlying medical condition. The model hence learned precisely the wrong relationship. While this was only a static classification setting, were resources to be dynamically allocated based on such a model, the resulting distribution would be sub-optimal.

Legal requirements can pose restrictions on solutions. Algorithmic fairness can be conceived of as the area most concerned with the consideration and impacts on protected attributes. But the law may also have requirements for the rationale. An oft-cited requirement of administrative decision-making is that it not be “arbitrary and capricious.” While the applicability to algorithms adopted by government agencies is open to debate, one interpretation is that decisions should be causally sound and not based on spurious correlations, even if such correlations give high predictive accuracy.

There have been several research avenues dedicated to such issues in SDM algorithms. There has been recent work on incorporating prior constraints in reinforcement learning [128], constrained linear bandits [6], and causal inference in bandits [51, 93, 95, 102, 119, 157]. However, there are some difficulties in applying such work to the kinds of problems we’ve outlined here. For one, causal bandits are focused on learning the effect of interventions on causal graphs. Each arm corresponds to assigning specific values to a set of nodes, after which we witness the effect on the outcome variable. This approach is useful for traditional questions of causal inference, such as determining the effect of school choice on educational outcomes, for instance. It may be less applicable, however, to the kinds of resource-constraint problems we’ve identified here. For these problems, we’re often interested in the causal connection between an observation’s features and its label. Whether this problem can be connected to the causal bandits literature in some way is an open question. Meanwhile, the constrained reinforcement learning approach of Roy et al. [128] relies on constrained MDPs [5] and uses cost functions to model constraints. The policy space is then restricted to those policies that do not exceed the cost function in expectation. While promising, meeting constraints in expectation only may be unsuitable for sensitive questions of law and public policy. Finally, constrained bandits restrict the action space to a “safe set” of arms. In our applications however, it is rare that constraints take the form of overt restrictions on what can be sampled. Rather, they pertain to why certain arms are sampled, or which arms can be sampled in conjunction with which others (e.g., enforcing limits on how many facilities from a given geographic can be audited). Moreover, the approaches outlined above, insofar as they are applicable to public policy, must also be adapted to deal with the difficulties of previous sections (distribution shifts, unknown causal structures, small data, multiple objectives, etc.).
4.7 Validation

Even when all of the above challenges are tackled, researchers working on SDMs for public policy face another not-entirely-solved hurdle: validation of the SDM algorithm.

Validation for SDM algorithms remains an ongoing and unsolved challenge. To be confident that a particular SDM method will function as expected, it would ideally provide confidence intervals on validation performance, theoretical guarantees, and some notion of robustness to distribution shifts. Researchers will likely need to do this a priori with logged data to be sure of an algorithm’s utility.

The right method for measuring and providing confidence intervals on the performance of SDM algorithms is still debated [3, 72]. Off-policy evaluation on retrospective data is also an ongoing research challenge [54, 114]. And much work has pointed to the consistently underpowered evaluations of machine learning algorithms [3, 27, 46]. Finally, providing statistical guarantees on algorithm performance often using relying on assumptions of linearity, with only some work proving guarantees for non-linear estimators or agnostic of estimators [52].

Policy problems and associated data do not conform to formats that are ideal for validation. Data is often high-dimensional and non-linear [73]. Though sometimes it is randomly collected, it may also come from data that could be highly off-policy. And more often than not, there is little data to create benchmarks with large statistical power. As such, it may be necessary to incorporate evaluation in the SDM algorithm itself to ensure sufficient exploration to assess algorithm performance in an iterative fashion. Recent work explores such challenges and may prove fruitful in tackling the validation difficulties in law and public policy [33, 153, 156].

The ability to evaluate an algorithm in a robust fashion has implications for ensuring compliance with legal standards – a unique challenge for public policy and law deployments. As noted earlier, the law may require population estimates. The 2018 Office of Management and Budget (OMB) guidelines, for example, recommended that compliance with regulations required misreporting estimates to be “statistically valid” (unbiased estimates of the mean) and have “±3% or better margin of error at the 95% confidence level for the improper payment percentage estimate.” [118, 145]. And new legal requirements for algorithmic robustness are being introduced (see, e.g., a survey of legislation and regulation efforts in [1]), that may introduce requirements on validation procedures.

Last, validation may be easier with simpler approaches [130], which may inform the complexity of modeling. While standard bandits, for instance, do not require modeling features or context, contextual bandits or reinforcement learning approaches typically do. Such models may require more investigation to validate (e.g., to assess model sensitivity, calibration, tuning). In part for that reason, one public health jurisdiction adopted a simpler bandit approach for allocating testing [40].

Example 4.8 (Identifying Claims at the SSA). To reduce processing times, the U.S. Social Security Administration (SSA) has developed mechanisms for quickly identifying disability claims likely to be approved [37]. The system makes its prediction based on information such as medical history and treatment protocols.

**Formulation.** While there may be some a priori identifiable constraints – e.g., perhaps that identification of disability claims should not depend on gender – others must be learned over time. Specifically, administrative law may require ascertaining whether the algorithm (i) has relied on impermissible factors, (ii) has incorporated required factors (e.g., those that the law mandates be taken into account), and (iii) has reached a reasoned decision. These raise challenging technical problems. Suppose we have a model \( f \) used to predict whether a disability claim \( x \) will be approved. At each round \( t \), the model receives a new claim and determines whether to flag it for review or not. If it is reviewed, a human will identify the true label and the model will be updated. At round \( t \), we discover that \( f \) has been relying on a correlation that it shouldn’t.

How do we modify \( f \) to adopt new constraints? Simply removing the confounds \( C \) from the feature space may be insufficient, as other features may be correlated with \( C \). One possibility involves adversarial selection, originally proposed for learning deconfounded lexicons in NLP [123]. The idea is that a decision rule for \( f \) should be unrelated to the confounds \( C \), thus have trouble predicting \( C \) given \( x \). We can implement this using a gradient reversal layer [63].

More specifically, we train a neural network to learn an encoding \( e(x) \) of the observation \( x \) whose loss is \( L_y - L_C \), where \( L_y \) is the loss of \( e(x) \) predicting the label \( y \), and \( L_C \) the loss of \( e(x) \) predicting values of \( C \). Thus, we learn an encoding which is correlated with \( y \) but not with \( C \). This approach has also been used to try and enforce fairness constraints in neural networks [125]. As this is a sequential decision making system, these selection mechanisms may need to be introduced into the sampling algorithm itself, as in the case of balanced bandits [51].

Example 4.9 (Wildfire Prevention Targeting at DOI). The U.S. Department of the Interior (DOI) is responsible for implementing wildfire management policies. A 2019 Executive Order tasks DOI with implementing more active management of forests to improve wildfire outcomes [148]. In particular, DOI takes several wildfire fuel treatment actions to reduce the magnitude of forest fires, including prescribed fires, removing brush, etc. DOI utilizes decision support systems to prioritize and track treatment of fuels, in addition to implementing a database to validate predicted outcomes of those treatments when a forest fire hits the treated fuel [131].

**Formulation.** Various SDM algorithms have been applied to wildfire management policy-making and others have reviewed a number of these applications, including reinforcement learning methods [78]. Lauer et al. [94], for example, propose a reinforcement learning for learning optimal strategies for forest management to reduce wildfire risk. Consider an RL problem, where an agent
at timestep $x$ is given the option to clear brush or conduct a prescribed burn from one of $A$ regions. It is difficult to reward the agent based on true wildfire reduction, so the agent is trained in simulation with a custom reward as prescribed by Lauer et al. [94] to reduce wildfire risk. There is limited data to evaluate the effectiveness of policy based on recorded wildfire intensities after treatments, provided by the Fuels Treatment Effectiveness Monitoring program [131]. Key open questions are how researchers can evaluate a program program based on limited off-policy data without overfitting.

5 ASSESSING AND MITIGATING SOCIAL HARMs

Ensuring the safety and reliability of machine learning systems used in government is of paramount importance due to the sensitivity of information being collected and the importance of the decisions being made. If automated decision making systems are relied upon at scale, small errors can be magnified in their effect size [24]. Algorithmic failure modes have been amply demonstrated, from bias and disparate impact [13, 25, 48, 88, 106, 117], lack of transparency [45, 79, 130], erosion of privacy [21, 147], and misplaced or non-existent accountability [43, 56, 91]. Such issues have spurred much discussion on the role and safety measures of AI in government [26, 96, 112, 129] in addition to the legality of employing these tools [44, 56]. We refrain from restating these known and serious potential harms and refer readers to extensive treatments on these topics [14, 87], but we note that addressing such harms will be critical in considering any SDM deployment. From a validity perspective [49], our work highlights the technical challenges required for successful public sector deployment. The advertising fallacy can lead to harms if these challenges aren’t catalogued and faced head on, and the motivations and potential benefits set forth in Section 2 may be irrelevant if harms are not addressed.

We do emphasize that SDM algorithms carry with them distinct risks of social harm, beyond those well-documented in non-sequential machine learning. SDM algorithms like reinforcement learning may learn to exploit reward functions in unpredictable ways. Carroll et al. [29], for example, found that there are situations where an algorithm can learn to induce preference shifts in its users to achieve its goals. In the context of a recommendation system this may mean radicalizing a user to get more clicks, as opposed to simply finding content that the user might like. The influences that these effects can have on policymakers and line officers at agencies, may be in conflict with core legal principles [56]. As illustrated in Section 4.5, the danger of feedback loops in SDM algorithms is substantial. The danger arises precisely because, as witnessed by the many applications we’ve discussed, an SDM algorithm is selecting its own training data. This results in statistical bias which, if unaccounted for, can lead the algorithm to begin self-reinforcing its own priors. Poor initial validation can lead to model overconfidence, and thus over-reliance on the predictions by policymakers. This problem can be exacerbated by small datasets which may cause overfitting and a tendency to read erroneous signals from noise.

High quality, ground truth datasets remain essential to overcome cold start problems with SDMs.

While the examples we discuss here may illustrate potential benefits of formulating existing processes as SDMs, the same techniques might be applied for more nefarious purposes. Efficient algorithms might, for instance, encourage government overreach, or be used by autocratic regimes. They might also falsely convince policymakers that better algorithms are a one-size-fits-all solution, but they cannot substitute for good governance and strong oversight.

Finally, when policies can be updated, bad actors might learn to exploit algorithmic failure modes. For example, they might inject poisoned data [99, 103] to steer the algorithm away from them or to manipulate outcomes. Consider for example the case of audit selection. If a set of entities wished to exploit the algorithm and avoid audits, entities with similar features could pool their resources together and vigorously fight any audits for their group. If the algorithm saw enough examples of audits that yielded no successful enforcement outcomes, it might de-prioritize the entire group in future rounds. This could then lead to a feedback loop where the group is never audited again.

Several implications follow from these harms. First, the deployment of an SDM algorithm should be accompanied by a cost-benefit analysis, rolled out incrementally, and robustly stress-tested before large-scale deployment. Machine learning developers must work with public sector partners to identify the right mechanisms for oversight.

Second, shifting from off-the-shelf one-time procurements to a continuous oversight is essential for SDM systems. Such auditing mechanisms can potentially be encoded directly into the SDM process (e.g., inducing exploration to yield proper estimates of validity at each timestep). We note that further research is needed into auditing and monitoring [42, 126]. For example, to date, relying on FOIA requests has been insufficient for full transparency [62]. In short, the adoption of SDMs in the public sector heighten the need for mechanisms of review, evaluation, and oversight.

6 CONCLUSION

Our work has reviewed the potential and challenges for SDM in the public sector. SDM algorithms have the promise to improve law and public policy, making the public sector more efficient, transparent, and accountable. But the research challenges posed by the public sector are considerable to render such approaches policy-compliant, trustworthy, and effective. We hope this review will inspire much more work to address these challenges at the intersection of law and computer science and identify the potential social gains and limitations of SDMs in law and policy.

REFERENCES

[1] Ada Lovelace Institute, AI Now Institute and Open Government Partnership. Algorithmic accountability for the public sector. 2021.

[2] George Alexandru Adam, Chun-Hao Kingsley Chang, Benjamin Haibe-Kains, and Anna Goldenberg. Hidden risks of machine learning applied to healthcare: Unintended feedback loops between models and future data causing model degradation. In Machine Learning for Healthcare Conference, pages 710–731. PMLR, 2020.

[3] Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. Advances in Neural Information Processing Systems, 34, 2021.
Beyond Ads: Sequential Decision-Making Algorithms in Law and Public Policy

CSLAW ’22, November 1–2, 2022, Washington, DC, USA

Piyush Jain, Sean CP Coogan, Sriram Ganapathi Subramanian, Mark Crowley, Dan Jurafsky and James H Martin. Speech and language processing. vol. 3.

Matthew S Johnson, David I Levine, and Michael W Toffel. Improving regulatory infrastructure engineering, 10(9):1176–1210, 2014.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by gret bounds for batched bandits. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 7340–7348, 2021.

Claudio Ferraz and Frederico Finan. Exposing corrupt politicians: the effects of brazil’s publicly released audits on electoral outcomes. The Quarterly journal of Economics, 123(2):703–745, 2008.

Katherine Fink. Opening the government’s black boxes: freedom of information and algorithmic accountability. Information, Communication & Society, 21(10):1453–1471, 2018.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by matching multi-armed bandits problem. arXiv preprint arXiv:1904.01763, 2019.

Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew J. Beam, Irene Y Chu, and Ronitt Rubinfeld. A review of challenges and opportunities in machine learning for health. AMIA Summits on Translational Science Proceedings, 2020:191, 2020.

Edward J. Glaeser, Andrew Hillsii, Scott Duke Kominers, and Michael D. L. Spatial social interactions and optimal management for a fire-threatened landscape. Forest Policy and Economics, 83:107–120, 2017.

Sanghack Lee and Elias Barreiro. Structural causal bandits: where to intervene? Advances in Neural Information Processing Systems 31, 2018.

Yangyi Lu, Amirhossein Meisami, and Ambuj Tewari. Causal bandits with unknown graph structure. In Proceedings of the 28th International Conference on Machine Learning, pages 818–825, 2015.

Samir Madanat. Optimal infrastructure management decisions under uncertainty. Transportation Research Part C: Emerging Technologies, (13):77–88, 1995.

Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: learning good interventions via causal inference. In Proceedings of the 30th International Conference on Neural Information Processing Systems, pages 1189–1197, 2016.

Fenwick McKelvey and Margaret MacDonald. Artificial intelligence policy infrastructure engineering, 10(9):1176–1210, 2014.

Karenna Gou and Dale Schuurmans. Discriminative batch mode active learning. In NIPS, pages 593–600. Citeseer, 2007.

Peter Hendersoll, Riahsat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meier. Deep reinforcement learning that matters. In Proceedings of the AAAI conference on artificial intelligence, volume 32, 2018.

Peter Hendersoll, Ben Chugg, Brandon Anderson, Kristen Altenburger, Alex Turk, John Guyton, Jacob Goldin, and Daniel E. Ho. Integrating Reward Maximisation and Population Estimation: Sequential Decision-Making for Internal Revenue Service Audit Selection. arXiv preprint arXiv:2004.11910, 2020.

Daniel E. Ho. Does peer review work? an experiment of experimentalism. Stanford Law Review, 60(6):121–137, 2008.

Daniel E. Ho. Equity in the bureaucracy. UC Irvine Law Rev., 7:401, 2017.

Samer Madanat. Optimal infrastructure management decisions under uncertainty. Transportation Research Part C: Emerging Technologies, (13):77–88, 1995.

Finnian Lattimore, Tor Lattimore, and Mark D. Reid. Causal bandits: learning good interventions via causal inference. In Proceedings of the 30th International Conference on Neural Information Processing Systems, pages 1189–1197, 2016.

Christopher J. Lauer, Claire A. Montgomery, and Thomas G Dietterich. Spatial social interactions and optimal management for a fire-threatened landscape. Forest Policy and Economics, 83:107–120, 2017.

Sanghack Lee and Elias Barreiro. Structural causal bandits: where to intervene? Advances in Neural Information Processing Systems 31, 2018.

Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: learning good interventions via causal inference. In Proceedings of the 30th International Conference on Neural Information Processing Systems, pages 1189–1197, 2016.

Yangyi Lu, Amirhossein Meisami, and Ambuj Tewari. Causal bandits with unknown graph structure. arXiv preprint arXiv:2004.11910, 2020.

Samir Madanat. Optimal infrastructure management decisions under uncertainty. Transportation Research Part C: Emerging Technologies, (13):77–88, 1995.

Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: learning good interventions via causal inference. In Proceedings of the 30th International Conference on Neural Information Processing Systems, pages 1189–1197, 2016.

Christopher J. Lauer, Claire A. Montgomery, and Thomas G Dietterich. Spatial social interactions and optimal management for a fire-threatened landscape. Forest Policy and Economics, 83:107–120, 2017.

Sanghack Lee and Elias Barreiro. Structural causal bandits: where to intervene? Advances in Neural Information Processing Systems 31, 2018.

Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: learning good interventions via causal inference. In Proceedings of the 30th International Conference on Neural Information Processing Systems, pages 1189–1197, 2016.

Christopher J. Lauer, Claire A. Montgomery, and Thomas G Dietterich. Spatial social interactions and optimal management for a fire-threatened landscape. Forest Policy and Economics, 83:107–120, 2017.

Sanghack Lee and Elias Barreiro. Structural causal bandits: where to intervene? Advances in Neural Information Processing Systems 31, 2018.

Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: learning good interventions via causal inference. In Proceedings of the 30th International Conference on Neural Information Processing Systems, pages 1189–1197, 2016.

Christopher J. Lauer, Claire A. Montgomery, and Thomas G Dietterich. Spatial social interactions and optimal management for a fire-threatened landscape. Forest Policy and Economics, 83:107–120, 2017.

Sanghack Lee and Elias Barreiro. Structural causal bandits: where to intervene? Advances in Neural Information Processing Systems 31, 2018.

Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: learning good interventions via causal inference. In Proceedings of the 30th International Conference on Neural Information Processing Systems, pages 1189–1197, 2016.

Christopher J. Lauer, Claire A. Montgomery, and Thomas G Dietterich. Spatial social interactions and optimal management for a fire-threatened landscape. Forest Policy and Economics, 83:107–120, 2017.

Sanghack Lee and Elias Barreiro. Structural causal bandits: where to intervene? Advances in Neural Information Processing Systems 31, 2018.

Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: learning good interventions via causal inference. In Proceedings of the 30th International Conference on Neural Information Processing Systems, pages 1189–1197, 2016.

Christopher J. Lauer, Claire A. Montgomery, and Thomas G Dietterich. Spatial social interactions and optimal management for a fire-threatened landscape. Forest Policy and Economics, 83:107–120, 2017.

Sanghack Lee and Elias Barreiro. Structural causal bandits: where to intervene? Advances in Neural Information Processing Systems 31, 2018.

Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: learning good interventions via causal inference. In Proceedings of the 30th International Conference on Neural Information Processing Systems, pages 1189–1197, 2016.

Christopher J. Lauer, Claire A. Montgomery, and Thomas G Dietterich. Spatial social interactions and optimal management for a fire-threatened landscape. Forest Policy and Economics, 83:107–120, 2017.

Sanghack Lee and Elias Barreiro. Structural causal bandits: where to intervene? Advances in Neural Information Processing Systems 31, 2018.

Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: learning good interventions via causal inference. In Proceedings of the 30th International Conference on Neural Information Processing Systems, pages 1189–1197, 2016.

Christopher J. Lauer, Claire A. Montgomery, and Thomas G Dietterich. Spatial social interactions and optimal management for a fire-threatened landscape. Forest Policy and Economics, 83:107–120, 2017.

Sanghack Lee and Elias Barreiro. Structural causal bandits: where to intervene? Advances in Neural Information Processing Systems 31, 2018.

Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: learning good interventions via causal inference. In Proceedings of the 30th International Conference on Neural Information Processing Systems, pages 1189–1197, 2016.

Christopher J. Lauer, Claire A. Montgomery, and Thomas G Dietterich. Spatial social interactions and optimal management for a fire-threatened landscape. Forest Policy and Economics, 83:107–120, 2017.

Sanghack Lee and Elias Barreiro. Structural causal bandits: where to intervene? Advances in Neural Information Processing Systems 31, 2018.

Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: learning good interventions via causal inference. In Proceedings of the 30th International Conference on Neural Information Processing Systems, pages 1189–1197, 2016.

Christopher J. Lauer, Claire A. Montgomery, and Thomas G Dietterich. Spatial social interactions and optimal management for a fire-threatened landscape. Forest Policy and Economics, 83:107–120, 2017.

Sanghack Lee and Elias Barreiro. Structural causal bandits: where to intervene? Advances in Neural Information Processing Systems 31, 2018.
