Journal of the Royal Statistical Society Series A: Statistics in Society, 2025, **00**, 1–25
https://doi.org/10.1093/jrsssa/qnaf122

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series A
Statistics in Society

Original Article

Bayesian safe policy learning with chance constrained optimization: application to military security assessment during the Vietnam War

Zeyang Jia[1], Eli Ben-Michael[2] and Kosuke Imai[3]

[1]Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, MA 02138, USA
[2]Department of Statistics & Data Science and Heinz College of Information Systems & Public Policy, Carnegie Mellon University, 4800 Forbes Avenue, Hamburg Hall, Pittsburgh, PA 15213, USA
[3]Department of Government and Department of Statistics, Harvard University, Institute for Quantitative Social Science, 1737 Cambridge Street, Cambridge, MA 02138, USA

Address for correspondence: Kosuke Imai, Department of Government and Department of Statistics, Harvard University, Institute for Quantitative Social Science, 1737 Cambridge Street, Cambridge, MA 02138, USA. Email: imai@harvard.edu

Abstract

Algorithmic decisions are increasingly used in high-stake domains such as criminal justice and medicine. We examine whether a security assessment algorithm deployed during the Vietnam War could have been improved using outcomes observed shortly after its rollout in late 1969. This empirical setting highlights methodological challenges common in real-world algorithmic decision-making. First, a new algorithm must be carefully evaluated to avoid worsening outcomes relative to an existing algorithm. Second, because the existing algorithm is deterministic, learning improvements require extrapolation that is both transparent and credible. Third, the use of discrete decision tables complicates optimization. We introduce the Average Conditional Risk (ACRisk), which quantifies the risk that a new algorithm performs worse for subgroups of units and then averages this risk across the population. We also develop a Bayesian policy learning framework that maximizes the posterior expected outcome while constraining the expected ACRisk. This approach separates treatment effect estimation from policy optimization, allowing flexible modelling and tractable search over complex policy classes. We show that this leads to a constrained linear programming problem. Applying our method, the learned algorithm rates most regions as more secure and assigns greater emphasis on economic and political factors over military ones.

Keywords: algorithmic decisions, Bayesian nonparametrics, Conditional Average Treatment Effect, decision tables, risk

1 Introduction

Algorithmic decisions and recommendations have long been used in areas as diverse as credit markets (Lauer, 2017) and war (Daddis, 2012). They are now increasingly integral to many aspects of today's society, including online advertising (e.g. Li et al., 2010; Schwartz et al., 2017; Tang et al., 2013), medicine (e.g. Kamath et al., 2001; Nahum-Shani et al., 2018), and criminal justice (e.g. Greiner et al., 2020; Imai et al., 2023). A primary challenge when applying data-driven policies to consequential decision-making is to characterize and control the risk associated with any new policies learned from the data. Stakeholders in fields such as medicine, public policy, and the military may be concerned that the adoption of new data-derived policies could inadvertently lead to worse outcomes for some individuals in certain settings.

In this article, we consider a particularly high-stake setting, analyzing a United States (US) military security assessment policy that saw active use in the Vietnam War. During the war, the US

Received: December 14, 2024. **Revised:** July 7, 2025. **Accepted:** July 8, 2025

military developed a data-driven scoring system called the Hamlet Evaluation System (HES) to produce a security score for each region (PACAF, 1969); commanders used these scores to make air strike decisions. A recent analysis based on a regression discontinuity design shows that the airstrikes had significantly negative effects on development outcomes including regional safety, economic, and civic society measures, and so were broadly counter-productive (Dell & Querubin, 2018).

We consider whether it would have been possible for modern tools to recognize this at the time, and adjust the HES using contemporaneous data collected by the US military and related agencies to optimize for various military, economic, and social objectives, while practically controlling the risk of worsening these objectives in too many regions. In particular, the original HES was composed of various 'sub-model scores' that measured different aspects of each region (e.g. economic variables, local administration, enemy military presence) based on HES survey responses. It then combined these into a single security score through a three-level hierarchical aggregation using predefined decision tables. The security scores were presented to Air Force commanders, who used them to make targeting decisions.

Our focus is on modifying these underlying decision tables while maintaining the transparency and interpretability of the original HES. Even though the system was in use over a half-century ago, this analysis serves three purposes. First, as discussed below, we show how to address methodological problems that are also common in other algorithmic decisions and recommendations. Second, because we limit to using data that were available at the time, our analysis serves as a case study on the development of algorithm-assisted decision-making tools in high-stake settings. Finally, investigating the potential for improvement gives insight into ways that the HES fell short—even with regards to its stated objectives—providing a historical lesson.

This empirical analysis poses several methodological challenges that are commonly encountered in high-stake data-driven decision-making settings. First, we want to characterize and control the risk that a new learned decision, classification, or recommendation policy may lead to worse outcomes for some cases. Second, the HES is a deterministic function of the input data, implying that extrapolation is necessary to learn new policies. Third, the security score is produced via a series of aggregations using decision tables, which is discrete and difficult to optimize over. Indeed, such decision tables are widely used in many public policy and medical decision-making settings (e.g. risk scores in the US criminal justice system Ben-Michael et al., 2024a; Greiner et al., 2020; Imai et al., 2023).

To address these challenges, we introduce a risk measure, the Average Conditional Risk (ACRisk), that first quantifies the risk of a given policy for groups of individual units with a specified set of covariates and then averages this conditional risk over the distribution of the covariates. In contrast to existing risk measures that characterize the uncertainty around the average performance of the policy (e.g. Bai et al., 2022; Delage & Mannor, 2010; Vakili & Zhao, 2015), the ACRisk measures the extent to which a learned policy negatively affects subgroups. This allows us to better characterize the potential heterogeneous risks of applying a new policy. With this risk measure in hand, we propose a Bayesian safe policy learning framework that maximizes the posterior expected value given the observed data while controlling the posterior expected ACRisk. We formulate this as a chance-constrained optimization problem and show how to convert it into a linear constraint, which can be solved as a standard optimization problem.

The primary advantages of the proposed framework are its flexibility and practical applicability. Because the chance-constrained optimization problem only relies on posterior samples, one can use popular Bayesian nonparametric regression models such as BART and Gaussian Process regression (Chipman et al., 2010; Rasmussen & Williams, 2006), while efficiently finding an optimal policy within a complex policy class. This is especially helpful in settings such as ours with limited or no covariate overlap, where our framework allows for flexible extrapolation through the Bayesian prior. In contrast, frequentist notions of safe policy learning rely on robust optimization and require solving a minimax optimization problem over both the class of potential models and the class of potential policies, making it difficult to consider nonparametric models and complex policy classes at the same time (Ben-Michael et al., 2024b; Kallus & Zhou, 2021; Pu & Zhang, 2020; Zhang et al., 2022).

We show through simulation studies that controlling the posterior expected ACRisk effectively limits the ACRrisk across various scenarios, reducing the risk of harming certain subgroups of units. We also find that although the proposed methodology is designed to be conservative, under some settings with a low signal-to-noise ratio, it yields a new policy whose average value is higher than a policy learned without the safety constraint. This is evidence that the proposed safety constraint can effectively regularize the policy optimization problem.

In our empirical analysis, we apply the proposed methodology to find adaptations to the HES that directly optimize for military, economic, and social objectives, while limiting the posterior probability that these objectives worsen under the hypothetical new system relative to the original HES. We consider two policy learning problems—one where we only change the last layer of the hierarchical aggregation that combines military, political and social economic sub-model scores, and another where we modify all of the decision tables used in the three-level hierarchical aggregation at the same time. To deal with the latter complex case, we develop a stochastic optimization algorithm, based on random walks on partitions of directed acyclic graphs, that is generally applicable to decision tables.

Our analysis consistently shows that the original HES is too pessimistic—assessing regions as too insecure—and places too much emphasis on military factors, even when targeting military objectives. In contrast, our data-derived adaptations to the HES assess regions as more secure and rely more on economic and social factors to produce regional security scores.

1.1 Related literature

Recent years have witnessed a growing interest among statisticians and machine learning researchers in finding optimal policies from randomized experiments and observational studies (e.g. Athey & Wager, 2021; Beygelzimer & Langford, 2009; Dudík et al., 2011; Kallus, 2018; Kitagawa & Tetenov, 2018; Luedtke & Van Der Laan, 2016; Qian & Murphy, 2011; Swaminathan & Joachims, 2015; Zhang et al., 2012; Zhao et al., 2012; Zhou et al., 2017, 2022). These works typically consider the following two steps under a frequentist framework—first characterize the average performance, or value, of a given policy via the Conditional Average Treatment Effect (CATE), and then learn an optimal policy by maximizing the estimated value based on the observed data.

In contrast, we adopt a Bayesian perspective—first obtain the posterior distribution of the CATE given the observed data, and then learn an optimal policy by maximizing the posterior expected value. Bayesian methods have been widely used for causal inference (see Li et al., 2022b, for a recent review). In particular, BART and Gaussian processes are often used to flexibly estimate the CATE (Branson et al., 2019; Hahn et al., 2020; Hill, 2011; Taddy et al., 2016). However, it appears that the Bayesian approach has been rarely applied to policy learning (Kasy, 2018). Our proposed framework takes advantage of these popular Bayesian nonparametric methods for safe policy learning.

There is also a growing literature on policy learning in scenarios where the CATE is unidentifiable (Ben-Michael et al., 2024b, 2025; Kallus & Zhou, 2021; Manski, 2007; Pu & Zhang, 2020; Yata, 2021; Zhang et al., 2022). These include observational studies with unmeasured confounders (Kallus & Zhou, 2021), studies with noncompliance or an instrumental variable (Pu & Zhang, 2020), studies that lack overlap due to deterministic treatment rules (Ben-Michael et al., 2025; Zhang et al., 2022), and utility functions that involve the joint set of potential outcomes (Ben-Michael et al., 2024b). These works first partially identify the value of a given policy then find the policy that maximizes the worst-case value using robust optimization. Our approach differs in that we decouple estimation and policy optimization by only relying on posterior samples for policy learning.

In the Reinforcement Learning (RL) literature, various notions of safety have been studied under different names (e.g. safe reinforcement learning, risk averse reinforcement learning, pessimistic reinforcement learning; see Garcıa & Fernández, 2015). For example, Geibel and Wysotzki (2005) control the risk of the agent visiting a 'dangerous state' by explicitly imposing a risk constraint when finding an optimal policy. In contrast, Sato et al. (2001) and Vakili and Zhao (2015) use the variance of the return as a penalty term in the objective when finding an optimal policy with a high expected return and low variance. While this RL literature focuses on online settings where the algorithm is designed to avoid risks during exploration, we study the risk of applying data-driven policies in offline settings.

We also extend existing work by developing the notion of ACRisk and using it as a constraint in optimizing the posterior expected value of a new policy. A related literature is *pessimistic offline RL*, which quantifies the risk of a given policy using the lower confidence bound (LCB) of the value, and finds a policy that has the best LCB (Bai et al., 2022; Buckman et al., 2020; Chen & Jiang, 2022; Jin et al., 2022, 2021; Rashidinejad et al., 2021; Shi et al., 2022; Uehara & Sun, 2021; Xie et al., 2021; Yan et al., 2022; Yin & Wang, 2021; Zanette et al., 2021). In contrast, the proposed ACRisk measures the extent to which a new policy negatively affects some groups of individuals when compared to the baseline policy.

Finally, our work is also related to chance constrained optimization, which is widely used in the analysis of decision making under uncertainty (e.g. Delage & Mannor, 2007, 2010; Farina et al., 2016; Filar et al., 1995; Schwarm & Nikolaou, 1999). For example, Delage and Mannor (2010) consider chance constrained control for Markov Decision Processes. They assume a Gaussian model for the reward distribution and use chance constrained optimization to find a policy that achieves low regret with high posterior probability. In contrast, our method considers a more general setup beyond the Gaussian model and uses the posterior expected value of the ACRisk as a constraint.

1.2 Outline of the article

The remainder of this article is organized as follows. Section 2 describes US military security assessment in the Vietnam War, the HES, and the related empirical policy learning problem. Section 3 introduces a formal setup, and Section 4 describes the Bayesian safe policy learning framework and the chance-constrained optimization procedure, as well as implementation via Gaussian Processes and Bayesian Causal Forests. Section 5 presents numerical experiments evaluating our proposal. Section 6 applies the Bayesian safe policy learning method to the Military security assessment problem. Section 7 concludes and discusses limitations and future directions.

2 Military security assessment during the Vietnam War

During the Vietnam War, the United States Air Force (USAF) conducted numerous air strike campaigns. One factor guiding USAF commanders in making targeting decisions during these missions was a data-driven scoring system called the HES whose goal was to provide a metric for regional security based on survey data (PACAF, 1969). We briefly describe the HES and its aggregation rules, as well as the policy learning problem we consider. Specifically, we focus on the second version of the HES, which was developed in 1969 and was originally analyzed by Dell and Querubin (2018).

2.1 The hamlet evaluation system

The HES was based on a total of 169 military, political, and socioeconomic indicators collected quarterly by Civil Operations and Revolutionary Development Support (CORDS), a joint civilian-military agency. These indicators were then categorized and summarized as 20 continuous 'sub-model' scores.[1] Each sub-model score ranges from 1 to 5 and measures a different aspect of a given region, such as *Enemy Military Activity*, *Economic Activity*, and *Public Health*. The HES aggregates these 20 continuous-valued sub-model scores into a single integer-valued security score ranging between 1 and 5. The sub-model scores are all semantically ordered so that lower values indicate that a region is 'worse' in the metric. The final HES score is ordered so that regions with a score of 1 should get the highest priority from USAF commanders and those with a score of 5 should get the lowest. After calculating the security score for each region at the beginning of a quarter, USAF commanders used these scores to make air strike decisions during that quarter.

Although several other factors contributed to the determination of eventual air strike targets, Dell and Querubin (2018) show that regions which fell just above a security score threshold—and hence had a higher security score—were less likely to be subject to air strikes than those regions

[1] The HES actually produces 19 different sub-model scores from the 169 indicators, but one sub-model score is used twice in the later aggregation. For simplicity and clarity, consider these to be 20 scores.

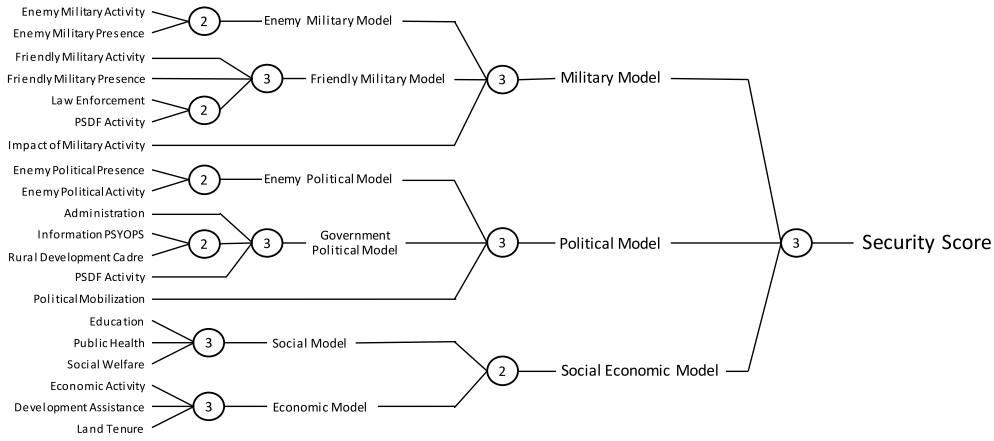

Figure 1. Aggregation of 20 sub-model scores. The HES uses 20 sub-model scores as inputs, and aggregates them using two-way and three-way decision tables. Each circle corresponds to one aggregation based on the two-way or three-way decision table, and the decision tables used in different circles are the same.

just below the threshold. Using a regression discontinuity design, the authors find that on average, the air strike campaign was largely counter-productive, increasing insurgency activities and decreasing civic engagements.

Despite the overall negative impact of the air strikes found in Dell and Querubin (2018), the effects at different decision boundaries vary (see supplementary material of Dell & Querubin, 2018), suggesting potential heterogeneity in treatment effects. Before looking at the data, it is in principle possible for changing the security score to have positive or negative effects. For instance, a lower security score may lead to more airstrikes, which may directly harm the regional economy and civic engagement, but it may also potentially reduce the level of insurgent activities and improve the level of regional safety, which may improve overall stability. Therefore, we investigate whether it is possible to improve the original HES security evaluation with the contemporaneous data available during the war to achieve better military, economic, and social objectives. As we will see, the broad negative impacts of air strikes appear to hold across many regions, and so the data-derived HES algorithms we learn generally increase the security score for most regions. Specifically, we focus on the first quarter that the HES was deployed (Q4 1969) and use the military impact assessments taken at the beginning of the succeeding quarter (January 1970) as the outcome to measure the efficacy of the HES.

2.2 Aggregation via decision tables

The HES aggregates 20 continuous sub-model scores to a single integer-valued security score that take a value in {1, 2, 3, 4, 5}, as shown in Figure 1. This hierarchical aggregation process starts by rounding the 20 continuous sub-model scores to the nearest integer in {1, 2, 3, 4, 5}. The security score then is produced through the following three steps. First, 18 out of these 20 rounded sub-model scores are aggregated to six integer-valued model scores relating to Enemy Military, Friendly Military, Enemy Political, Government Political, Social, and Economic model scores. Second, the resulting six model scores, along with two remaining inputs (impact of military activity and political mobilization), are aggregated to three higher tier integer-valued scores—Military, Political, and Social Economic model scores. Finally, these three scores are further aggregated to the integer-valued single security score.

At each step of aggregation, the HES uses a two-way or three-way decision table indicated by the number shown in each circle of Figure 1. These tables take two or three integer-valued scores, ranging from 1 to 5, and return a new score, which takes a value in {1, 2, 3, 4, 5}. Figures S1 and S2 in the Appendix show the two-way and three-way decision tables used in the HES, respectively. For example, the two-way decision table maps an input of (2, 3) to an output of 2. In the HES, the same two-way and three-way tables are used across all levels of aggregation.

In our analysis, we modify the underlying two-way and three-way decision tables. We focus on this, rather than modifying the hierarchical structure itself or using black-box machine learning tools, because the hierarchical structure contains substantial domain knowledge, and we wish to retain the transparency of the original system. This also facilitates easier comparisons between our data-derived modifications and the original HES system. As we discuss and explore in Section 6.3, altering the underlying decision tables alters the implicit influence of each of the 20 factors in constructing the overall HES score. We note, however, that our methodological development in the next sections can also be used to adjust the hierarchy itself or to use black box methods, although there may be additional computational considerations.

2.3 Evidence-based policy learning

To measure the impact of air strikes and learn a new decision policy, we analyze three binary outcomes that reflect regional security, economy, and civic society. These outcomes are based on the HES survey data collected in January 1970, after the period of air strikes in the final quarter of 1969 (Dell & Querubin, 2018). During this period, USAF commanders targeted 1024 of the 1954 regions defined by the HES.[2] Our objective is to develop a new decision rule that generates appropriate security scores based on the 20 continuous sub-model scores. It is worth noting that our decision variable is the output security score from the HES rather than air strike decisions, which are made by commanders. Therefore, our analysis assumes that a new scoring system would affect the decisions only through changes in the score. Under this assumption, we optimize the overall outcomes in regional security, economy, and civic society by constructing the new security score based on the 20 sub-model scores while marginalizing over the airstrike decisions made by commanders.

This application yields several methodological challenges. First, the stakes of this decision are high and there is potential heterogeneity across regions. This motivates a safe policy learning approach that limits the possibility of generating worse outcomes under a new, learned policy for some regions than under the existing policy. Second, the existing policy is deterministic, rather than stochastic. Therefore, due to the lack of overlap between regions that received different HES scores, we must extrapolate when estimating the potential outcomes under a new policy. Such extrapolation is usually computational intensive under a frequentist robust optimization framework, as it requires solving an inner optimization problem to partially identify the key parameters. Finally, the existing policy is based on multiple decision tables. Thus, to keep the existing structure of the policy, we must solve a complex optimization problem. We now develop a new Bayesian safe policy framework that addresses these challenges.

3 Preliminaries

Before we develop our Bayesian policy learning framework, we describe the problem setup and notation. In addition, we provide a brief review of Bayesian policy learning.

3.1 Setup and notation

We consider an individualized treatment rule or policy, which deterministically maps a set of p pretreatment covariates $\boldsymbol{X} \in \mathcal{X} \subseteq \mathbb{R}^p$ to a multivalued decision $D \in \mathcal{D} = \{0, 1, \ldots, K-1\}$ where K denotes the total number of decision categories. Formally, a policy is defined as a function $\delta : \mathcal{X} \to \mathcal{D}$. We have a simple random sample of n observations from the population of interest $\mathcal{P}$, which is characterized by the joint distribution of $\{\boldsymbol{X}, Y(0), Y(1), \ldots, Y(K-1)\}$ where $Y(k) \in \mathcal{Y}$ represents a potential outcome under decision $D = k$ with $k \in \mathcal{D}$ and the observed outcome is given by $Y_i = Y_i(D_i)$ (Neyman, 1990; Rubin, 1974). This notation implicitly assumes no interference between units (Rubin, 1980).

We also assume that the decision is unconfounded given the covariates, i.e. $\{Y_i(k)\}_{k=0}^{K-1} \perp\!\!\!\perp D_i \mid \boldsymbol{X}_i$ (Rosenbaum & Rubin, 1983). However, we do not assume that there exists covariate overlap for treated and controlled units—i.e. $0 < \Pr(D_i = k \mid \boldsymbol{X}_i = \boldsymbol{x}) < 1$ for all $k \in \mathcal{D}$ and $\boldsymbol{x} \in \mathcal{X}$. The lack of

[2] See Table S1 and Figure S3 of the online supplementary material for other summary statistics as well as the average values of the security score and various sub-model scores.

overlap is motivated by the fact that in many settings, the data are collected under an existing deterministic policy (Ben-Michael et al., 2025; Zhang et al., 2022).

Suppose that a policy-maker defines a utility function $u:\{0, 1, \ldots, K-1\} \times \mathcal{Y} \to \mathbb{R}$, which maps every decision-outcome pair to a real-valued utility. Then, the expected utility or value of policy δ in a policy class Δ is given by,

$$V(\delta) := \mathbb{E}[u(\delta(\boldsymbol{X}), Y(\delta(\boldsymbol{X})))]. \tag{1}$$

Our goal is to find an optimal policy within a policy class Δ that has a high value relative to the existing policy $\tilde{\delta}$, which also belongs to the same policy class Δ (Ben-Michael et al., 2025; Wei et al., 2022). In our application, the baseline policy is a deterministic rule that produces the security score in the HES. Our goal is to derive a new policy whose average value is higher than that of the original policy subject to a safety constraint that will be introduced in Section 4.

3.2 Bayesian policy learning

We next briefly introduce the basic elements of Bayesian policy learning (see Ding & Li, 2018; Li et al., 2022b, for reviews of Bayesian causal inference). Under the Bayesian framework, all parameters are regarded as random variables. Following the convention, we will use upper and lower case characters to represent parameters and their realizations, respectively.

Suppose that the marginal distribution of potential outcomes is characterized by a possibly infinite dimensional parameter $\boldsymbol{\Theta}$. Bayesian policy learning proceeds in two steps. First, we specify a prior distribution $\pi(\boldsymbol{\Theta})$ on the parameter $\boldsymbol{\Theta}$ and compute its posterior distribution using Bayes' rule,

$$\pi(\boldsymbol{\Theta} \mid \{\boldsymbol{X}_i, D_i, Y_i\}_{i=1}^n) \propto \prod_{i=1}^n p(Y_i \mid \boldsymbol{\Theta}, \boldsymbol{X}_i, D_i)\pi(\boldsymbol{\Theta}). \tag{2}$$

Under Bayesian policy learning, given the posterior distribution of $\boldsymbol{\Theta}$, the optimal policy maximizes the posterior expected value,

$$\delta_{opt} = \underset{\delta \in \Delta}{\operatorname{argmax}}\ \mathbb{E}\big[V(\delta; \boldsymbol{\Theta}) \mid \{D_i, \boldsymbol{X}_i, Y_i\}_{i=1}^n\big], \tag{3}$$

where $V(\delta; \boldsymbol{\Theta}) := \mathbb{E}[u(\delta(\boldsymbol{X}), Y(\delta(\boldsymbol{X}))) \mid \boldsymbol{\Theta}]$. We explicitly use the notation $V(\delta; \boldsymbol{\Theta})$ to emphasize its dependence on the parameter $\boldsymbol{\Theta}$ as well as the policy δ. Specifically, $V(\delta; \boldsymbol{\Theta})$ averages over both the marginal distribution of $\boldsymbol{X}$, and the conditional distribution of $Y(\delta(\boldsymbol{X}))$ given $\boldsymbol{X}$ and $\boldsymbol{\Theta}$. The expectation in equation (3) is taken over the posterior distribution of $\boldsymbol{\Theta}$ given the observed data.

To solve the optimization problem in equation (3), we can rewrite the expectation over parameter $\boldsymbol{\Theta}$, covariates $\boldsymbol{X}$, and potential outcome $Y(\delta(\boldsymbol{X}))$ as,

$$\mathbb{E}\big[V(\delta; \boldsymbol{\Theta}) \mid \{\boldsymbol{X}_i, D_i, Y_i\}_{i=1}^n\big] = \int_{\mathcal{X}} \sum_{k=0}^{K-1} \mathbb{E}\big[u(k, Y(k)) \mid \boldsymbol{X} = \boldsymbol{x}, \{\boldsymbol{X}_i, D_i, Y_i\}_{i=1}^n\big] I(\delta(\boldsymbol{x}) = k) dF(\boldsymbol{x}), \tag{4}$$

where the inside expectation is taken over the posterior distribution of $\boldsymbol{\Theta}$ given the observed data, and the distribution of the potential outcomes given $\boldsymbol{\Theta}$ and $\boldsymbol{X}$. The outside expectation is taken over the distribution of $\boldsymbol{X}$, where F is the CDF of $\boldsymbol{X}$. For simplicity, we will assume the distribution of X in the target population is known.

Thus, the original problem in equation (3) can consist of two steps: (i) estimate the posterior expected utility of decision d given a covariate profile $\boldsymbol{x}$, $\mathbb{E}[u(d, Y(d))|\boldsymbol{X} = \boldsymbol{x}, \{\boldsymbol{X}_i, D_i, Y_i\}_{i=1}^n]$, and (ii) find the optimal policy δ by solving the optimization (4) with the estimated values. This formulation separates the problem of computing the posterior distribution from that of finding an optimal policy.

4 The proposed methodology

We now describe our Bayesian safe policy learning framework. We begin by introducing a new risk metric, the ACRisk, that represents the probability that a new policy yields a worse expected utility conditional on covariates than the baseline policy. We then propose to maximize the posterior expected value of the new policy while limiting the posterior expected ACRisk. Our methodology consists of two steps: first estimating the CATE using a flexible Bayesian model, and then finding a policy within the policy class that maximizes the posterior expected value while controlling the posterior expected ACRisk. Finally, we show that this chance-constrained optimization problem can be written as a standard Bayesian policy learning problem with linear constraints, avoiding additional computational complexity.

4.1 Average Conditional Risk

In the existing literature, the risk of a policy is typically measured through uncertainty about its value. We then seek a policy that has a high value with a small estimation uncertainty. One limitation of this common approach is its failure to account for potential heterogeneity in risk across different groups of individuals. A policy that performs well on average may not benefit everyone. In our empirical application, a security assessment rule that is effective on average may negatively impact some regions. While the risk due to estimation uncertainty will become small as the sample size increases, this inherent risk due to heterogeneous treatment effects will remain, even in large samples.

To address this, we introduce a new risk measure, the ACRisk, which represents the probability that, conditional on the covariates, a policy yields a worse expected utility than the baseline policy.

Definition 1 (ACRisk). For a given policy δ, the ACRisk with respect to a baseline policy $\tilde{\delta}$ is

$$R(\delta, \tilde{\delta}; \boldsymbol{\theta}) := \mathbb{P}\big(\mathbb{E}[u(\delta(\boldsymbol{X}), Y(\delta(\boldsymbol{X}))) \mid \boldsymbol{X}, \boldsymbol{\Theta} = \boldsymbol{\theta}] < \mathbb{E}\big[u(\tilde{\delta}(\boldsymbol{X}), Y(\tilde{\delta}(\boldsymbol{X}))) \mid \boldsymbol{X}, \boldsymbol{\Theta} = \boldsymbol{\theta}\big]\big),$$

where $\boldsymbol{\Theta}$ represents the possibly infinite-dimensional parameter that governs the marginal distributions of potential outcomes.

The inner expectation in the above definition is taken over the conditional distribution of $Y(\delta(\boldsymbol{X}))$ given $\boldsymbol{X}$ and $\boldsymbol{\Theta}$ while the outer probability is defined with respect to the marginal distribution of $\boldsymbol{X}$. We use the above notation to make it explicit that the ACRisk depends on the parameter $\boldsymbol{\Theta}$, which governs the conditional distribution of potential outcomes given the covariates $\boldsymbol{X}$. Note that under this definition, $R(\tilde{\delta}, \tilde{\delta}; \boldsymbol{\theta}) = 0$ for any value of $\boldsymbol{\theta}$.

The ACRisk first evaluates whether a group of individuals with a certain set of covariates has a lower expected utility under policy δ than under the baseline policy $\tilde{\delta}$. It then computes the proportion of such at-risk groups in the population by averaging over the distribution of covariates. The ACRisk is dependent on the covariates we choose, and using different set of covariates can result in different at-risk groups, potentially leading to different ACRisk values.

Table 1 provides a numerical example where both Policies I and II outperform the baseline policy. While the improvement of Policy I comes at the expense of group B, Policy II performs equally well for both groups. As a result, the ACRisk of Policy I is higher, indicating that 50% of the population is at risk. The ACRisk minimizes the proportion of the population that is negatively influenced by the new policy, leading to safe policy learning. This guarantees that less than a prespecified proportion of the population would be hurt by the new policy. In addition, the ACRisk can be configured with a user-specified threshold $c > 0$ to limit the performance degradation of the new policy δ relative to the baseline δ_0:

$$\mathbb{P}\big(\mathbb{E}[Y(\delta(X))|X] - \mathbb{E}[Y(\delta_0(X))|X] < -c\big) \leq \epsilon.$$

Fortunately, all the method discussed later will still be applicable to this case.

The conservative nature of the ACRisk makes it suitable for many high-stake decision-making settings. For example, in education, we can learn a new policy that assigns a subset of students to

Table 1. A numerical example that illustrates the ACRisk

	Conditional expected utility		Value	ACRisk
	Group A (50%)	Group B (50%)		
Baseline policy	0	0	0	0
Policy I	4	−2	1	0.5
Policy II	1	1	1	0

Note. In this example, both Policies I and II outperform the baseline policy on average, but the improvement of Policy I comes at the expense of the group B. The ACRisk shows that 50% of the population, which is the size of the group B in this example, is at risk.

new teaching materials while ensuring that most students are not negatively affected by this change (e.g. Kizilcec & Lee, 2022; Kundu et al., 2021). Such policy learning can also be done in precision medicine, where new treatments are being considered (e.g. Javaid et al., 2022; Smith, 2005; Wiens et al., 2019), and in criminal justice, where changes to the pretrial risk assessment instrument may be proposed (e.g. Ben-Michael et al., 2025; Greiner et al., 2020). In these applications, a priority is to minimize the risk of yielding a worse outcome than the status quo.

We note that the ACRisk is a risk measure based on groups (defined by covariates $\boldsymbol{X}$). In particular, the ACRisk differs from the following individual measure of risk.

Definition 2 (Average Individual Risk (AIRisk)). For a given policy δ, its AIRisk with respect to a baseline policy $\tilde{\delta}$ is defined as,

$$R^*(\delta, \tilde{\delta}; \lambda) := \mathbb{P}\big(u(\delta(\boldsymbol{X}), Y(\delta(\boldsymbol{X})) < u(\tilde{\delta}(\boldsymbol{X}), Y(\tilde{\delta}(\boldsymbol{X}))) \mid \Lambda = \lambda\big),$$

where Λ represents a (possibly infinite dimensional) parameter that governs the joint distribution of potential outcomes.

Unlike the ACRisk, the AIRisk is not identifiable because it depends on the joint distribution of potential outcomes. One possible strategy is to control an upper bound of the AIRisk (Ben-Michael et al., 2024b; Kallus, 2022; Li et al., 2022a). We leave the development of safe policy learning in terms of the AIRisk to future research and focus on the ACRisk for the remainder of this article.

To control the ACRisk, we must estimate the parameter $\boldsymbol{\Theta}$. Under our Bayesian safe policy learning framework, we consider the posterior average of the ACRisk, which we call the Posterior Average Conditional Risk or PACRisk. This measure allows us to directly incorporate estimation uncertainty into our analysis.

Definition 3 (PACRisk). For a given policy δ, the PACRisk with respect to a baseline policy $\tilde{\delta}$ is

$$R_p(\delta, \tilde{\delta}) := \mathbb{E}\big[R(\delta, \tilde{\delta}; \boldsymbol{\Theta}) \mid \{\boldsymbol{X}_i, D_i, Y_i\}_{i=1}^n\big],$$

where the expectation is taken over the posterior distribution of $\boldsymbol{\Theta}$ given the observed data $\{\boldsymbol{X}_i, D_i, Y_i\}_{i=1}^n$.

4.2 Bayesian safe policy learning with chance constrained optimization

We now propose a Bayesian safe policy learning procedure that limits the PACRisk introduced above (Definition 3). Specifically, we find a policy within a policy class Δ that maximizes the posterior average value while limiting the PACRisk below a specified threshold $\epsilon \in [0, 1]$,

$$\begin{aligned} \delta_{\text{safe}} \;&= \operatorname*{argmax}_{\delta \in \Delta} \mathbb{E}\big[V(\delta; \boldsymbol{\Theta}) \mid \{\boldsymbol{X}_i, D_i, Y_i\}_{i=1}^n\big] \\ \text{subject to} \quad &\mathbb{E}\big[R(\delta, \tilde{\delta}; \boldsymbol{\Theta}) \big| \{\boldsymbol{X}_i, D_i, Y_i\}_{i=1}^n\big] \le \epsilon, \end{aligned} \tag{5}$$

The expectation in the optimization target is taken over the posterior distribution of the parameter $\boldsymbol{\Theta}$. We typically choose a policy class Δ that contains the baseline policy so that a solution to equation (5) always exists. The constraint in equation (5) ensures that for a randomly selected group of individuals, δ_{safe} has an at most ϵ posterior probability of yielding a worse expected value for the group than the baseline policy $\tilde{\delta}$. Thus, a smaller value of ϵ yields a safer policy that is more conservative, limiting the extent of changes to the baseline policy.

A standard Bayesian policy maximizes the posterior expected utility, modifying the baseline policy for subgroups where the posterior expected CATE is large. However, the additional PACRisk constraint in equation (4) prioritizes changes to the baseline policy for those subgroups whose posterior expected value of the CATE is large and its posterior uncertainty is small. Consider an illustrative example, in which $\boldsymbol{X}$ is a binary variable and $\boldsymbol{\Theta} = (\theta_0, \theta_1)$ where θ_0 and θ_1 are the CATEs for $\boldsymbol{X} = 0$ and $\boldsymbol{X} = 1$, respectively. The baseline policy in this case is to treat nobody. In addition, suppose that the posterior distributions of θ_0 and θ_1 are $N(10, 100)$ and $N(1, 0.1)$, respectively. Under this setting, the Bayesian optimal policy δ_{opt} will treat everyone even though the treatment effect for the subgroup $\boldsymbol{X} = 0$ is highly uncertain. In contrast, the PACRisk constraint in our methodology considers the posterior probability of the group-level positive treatment effect, and will be more conservative when changing the baseline policy may negatively affect some subgroups due to a high degree of posterior uncertainty of the CATE.

Our simulation and empirical findings show that controlling the PACRisk often leads to conservative inference, especially when the data are collected under a deterministic baseline policy. This is because our methodology is unlikely to chage the baseline policy whenever the posterior uncertainty of the CATE is high (even if the posterior expected CATE is large). The proposed method is most suitable for high-stake applications where it is crucial to avoid negatively affecting subgroups of a target populatoin. If the goal is to simply maximize the overall utility without considering the risk of such negative impact, then constraining the PACRisk may result in overly conservative policies.

One potential drawback of chance constraints such as the one shown in equation (5) is its computational difficulty due to their complex dependency on both the parameter $\boldsymbol{\Theta}$ and the policy δ. This makes it challenging to determine how the value of the constraint changes as a function of the policy δ. For this reason, researchers typically impose strong assumptions on the marginal distribution of potential outcomes and the parameter $\boldsymbol{\Theta}$ to obtain a closed-form solution instead (e.g. Delage & Mannor, 2010; Mowbray et al., 2022; Vitus et al., 2015).

We overcome this challenge by formulating equation (5) as a deterministic optimization problem with a linear constraint that separates the estimation of CATE from policy optimization in the manner similar to equation (4) under the standard Bayesian policy learning. The following theorem formally establishes the equivalence between the chance constraint optimization in equation (5) and a constrained linear programming problem.

Theorem 1 (Control of the PACRisk as a Linear Constraint). Define the posterior conditional benefit and risk of decision k relative to the existing policy $\tilde{\delta}$ as,

$$b_k(\boldsymbol{x}) := \mathbb{E}\big[\tau_k(\boldsymbol{x}, \boldsymbol{\Theta}) \mid \{D_i, \boldsymbol{X}_i, Y_i\}_{i=1}^n\big],$$
$$r_k(\boldsymbol{x}) := \mathbb{P}\big(\tau_k(\boldsymbol{x}, \boldsymbol{\Theta}) < 0 \mid \{D_i, \boldsymbol{X}_i, Y_i\}_{i=1}^n\big),$$

where $\tau_k(\boldsymbol{x}, \boldsymbol{\theta}) := \mathbb{E}[u(k, Y(k)) - u(\tilde{\delta}(\boldsymbol{x}), Y(\tilde{\delta}(\boldsymbol{x}))) \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{\Theta} = \boldsymbol{\theta}]$. Then, the chance-constrained optimization defined in equation (5) is equivalent to the following deterministic optimization problem,

$$\begin{aligned} \delta_{\text{safe}} = \underset{\delta \in \Delta}{\text{argmax}} \quad & \sum_{k=0}^{K-1} \int I(\delta(\boldsymbol{X}) = k) b_k(\boldsymbol{X}) dF(\boldsymbol{X}) \\ \text{subject to} \quad & \sum_{k=0}^{K-1} \int I(\delta(\boldsymbol{X}) = k) r_k(\boldsymbol{x}) dF(\boldsymbol{X}) \leq \epsilon \end{aligned} \tag{6}$$

where $F(\boldsymbol{x})$ is the cumulative distribution function of covariates $\boldsymbol{X}$.

Algorithm 1 Bayesian safe policy learning based on posterior draws

Input: Data $\{\boldsymbol{X}_i, D_i, Y_i\}_{i=1}^n$; A total of M posterior draws, i.e. $\{\boldsymbol{\Theta}^{(m)}\}_{m=1}^M$; Covariate distribution function $F(x)$

Output: Bayesian safe policy $\hat{\delta}_{\text{safe}} : \mathcal{X} \longrightarrow \mathcal{D} = \{0, 1, \ldots, K-1\}$

1 Approximate the posterior conditional benefit of decision k, $b_k(\boldsymbol{x}) = \mathbb{E}[\tau_k(\boldsymbol{x}, \boldsymbol{\Theta}) \mid \{\boldsymbol{X}_i, D_i, Y_i\}_{i=1}^n]$, by the sample average of posterior draws for $\boldsymbol{x} \in \mathcal{X}$:

$$\hat{b}_k(\boldsymbol{x}) = \frac{1}{M} \sum_{m=1}^{M} \tau_k(\boldsymbol{x}, \boldsymbol{\Theta}^{(m)})$$

2 Approximate the posterior conditional risk of decision k. $r_k(\boldsymbol{x}) = \mathbb{P}(\tau_k(\boldsymbol{x}, \boldsymbol{\Theta}) < 0 \mid \{D_i, \boldsymbol{X}_i, Y_i\}_{i=1}^n)$, by the sample average of posterior draws for $\boldsymbol{x} \in \mathcal{X}$:

$$\hat{r}_k(\boldsymbol{x}) = \frac{1}{M} \sum_{m=1}^{M} I\{\tau_k(\boldsymbol{x}, \boldsymbol{\Theta}^{(m)}) < 0\}$$

3 Solve this approximated optimization problem:

$$\hat{\delta}_{\text{safe}} = \underset{\delta \in \Delta}{\text{argmax}} \sum_{k=0}^{K-1} \int I(\delta(\boldsymbol{X}) = k) \hat{b}_k(\boldsymbol{X}) dF(\boldsymbol{X})$$
$$\text{subject to} \sum_{k=0}^{K-1} \int I(\delta(\boldsymbol{X}) = k) \hat{r}_k(\boldsymbol{x}) dF(\boldsymbol{X}) \leq \epsilon$$

Proof is given in the Appendix. In the theorem, $\tau_k(x, \boldsymbol{\theta})$ is the improvement in the conditional expected utility when changing the decision $\tilde{\delta}(x)$ to k where $\{\tau_k(x, \boldsymbol{\Theta})\}_{k=0}^{K-1}$ is fully determined by $K-1$ pairwise CATEs between K decisions. In practice, we use the empirical CDF for $F(\boldsymbol{X})$.

Theorem 1 enables us to explicitly separate the posterior of $\boldsymbol{\Theta}$ and the policy δ in the chance-constrained optimization. We note that this formulation is possible because the ACRisk of a policy δ is a weighted average of its conditional risk given $\boldsymbol{X}$, which can be represented by $\{r_k(\boldsymbol{x})\}_{k=0}^{K-1}$. We can first calculate the conditional improvement and risk, i.e. $b_k(\boldsymbol{x})$ and $r_k(\boldsymbol{x})$, using either closed-form calculation or samples from posterior draws. Then, like in standard Bayesian policy learning, we can solve the constrained policy optimization problem based on $b_k(\boldsymbol{x})$ and $r_k(\boldsymbol{x})$. Algorithm 1 summarizes the procedure of our proposed Bayesian safe policy learning based on the posterior draws of $\boldsymbol{\Theta}$.

If the number of posterior draws is large, the policy obtained by solving the optimization problem with posterior samples, $\hat{b}_k(\boldsymbol{x})$ and $\hat{r}_k(\boldsymbol{x})$, will be close to the one based on the true posterior mean of $b_k(\boldsymbol{x})$ and $r_k(\boldsymbol{x})$. For example, if the approximation error $\sup_{\boldsymbol{x} \in \mathcal{X}, 1 \leq k \leq K} |b_k(\boldsymbol{x}) - \hat{b}_k(\boldsymbol{x})|$ is $O(\frac{1}{M})$, then the posterior expected value of the obtained policy will be at most worse than the optimal policy that maximizes posterior expected value by $O(\frac{1}{M})$.

4.3 Bayesian CATE estimation

A primary advantage of the proposed Bayesian safe policy learning framework introduced above is separating the tasks of estimating the CATE and optimizing the policy. This means that our framework can readily accommodate the use of flexible Bayesian nonparametric models for CATE estimation.

As an example that mirrors our empirical application, suppose that we have an ordered decision, $k \in \{0, 1, \ldots, K-1\}$ and a continuous outcome. We begin by modelling the marginal

distribution of the baseline potential outcome $Y_i(0)$ and then the $K-1$ CATEs between two adjacent treatment levels. Specifically, we assume the following model,

$$Y_i(k) = \sum_{d=0}^{k} f_d(X_i) + \epsilon_i, \quad \forall\, k \in \{0, 1, \ldots, K-1\}$$

where ϵ_i is an i.i.d Gaussian random variable with mean 0 and variance σ^2, $f_0(\boldsymbol{x}) = \mathbb{E}[Y(0) \mid \boldsymbol{X} = \boldsymbol{x}]$, and $f_k(\boldsymbol{x}) = \mathbb{E}[Y(k) - Y(k-1) \mid \boldsymbol{X} = \boldsymbol{x}]$ for $1 \leq k \leq K-1$. Here, we have an infinite dimensional parameter $\boldsymbol{\Theta} = (\sigma^2, \{f_k(\cdot)\}_{k=0}^{K-1})$ that determines the marginal distribution of each potential outcome $Y(k)$ given covariates $\boldsymbol{X}$. If we have binary or categorical outcomes, we use a link function to transform the mean outcome to a continuous scale, e.g.

$$Y_i(d) \mid \boldsymbol{X}_i \sim \text{Bernoulli}(p_d(\boldsymbol{X}_i)) \quad \text{where } p_d(\boldsymbol{X}_i) = g^{-1}\left(\sum_{k=1}^{d} f_{k-1}(\boldsymbol{X}_i)\right) \tag{7}$$

where $g^{-1}(\cdot)$ is the inverse link function.

In principle, any Bayesian method can be applied for modelling $\{f_k(\cdot)\}_{k=0}^{K-1}$. In Section S3 of the online supplementary material, we discussed two popular methods, Bayesian Additive Regression Trees (BART) and Gaussian Processes (GPs) to demonstrate how such models can be accommodated.

5 A simulation study

In this section, we conduct a simulation study to examine the empirical performance of the proposed Bayesian safe policy learning methodology. To highlight the key concepts, we consider a simple setup with a binary decision.

5.1 Setup

We consider the following data generating processes with the number of observations n varying from 50 to 500, i.e. $n \in \{50, 100, 200, 500\}$. There are two covariates $\boldsymbol{X} = (X_1, X_2)$ that are independently and identically distributed according to a uniform distribution $X_1, X_2 \overset{i.i.d}{\sim} \text{Uniform}(-1, 1)$. There are two scenarios, one with covariate overlap and the other without it:

- **Scenario I (with covariate overlap):** The data are collected through a randomized experiment with $\mathbb{P}(D = 1 \mid \boldsymbol{X} = \boldsymbol{x}) = 0.5, \forall \boldsymbol{x} \in \mathcal{X}$. Here, the baseline policy is to treat nobody, i.e. $\tilde{\delta}(\boldsymbol{x}) = 0, \forall \boldsymbol{x} \in \mathcal{X}$.
- **Scenario II (without covariate overlap):** The data are collected under a deterministic policy $\tilde{\delta}(x) = I(x_1 > 0.5)$, which serves as the baseline policy. Thus, the treatment assignment is also deterministic, i.e. $\mathbb{P}(D = 1 \mid \boldsymbol{X} = \boldsymbol{x}) = \tilde{\delta}(\boldsymbol{x})$.

Within each simulation scenario, we specify the outcome model as,[3]

$$Y \mid \boldsymbol{X} \sim N\big(X_1 + X_2 + \{4I(X_1 > 0, X_2 > 0) - 2\}D|X_1||X_2|, \sigma^2\big)$$

where we consider a high signal-to-noise ratio case ($\sigma = 1$), a medium signal-to-noise ratio case ($\sigma = 2$), and a low signal-to-noise ratio case ($\sigma = 3$). In this setup, treatment is effective for individuals with $X_1 > 0$, $X_2 > 0$. Finally, we set the utility to be equal to the outcome, i.e. $u(d, y) = y$.

[3] We also consider a data generating process with binary outcomes in Section S4 of the online supplementary material. The results are broadly similar to the continuous case.

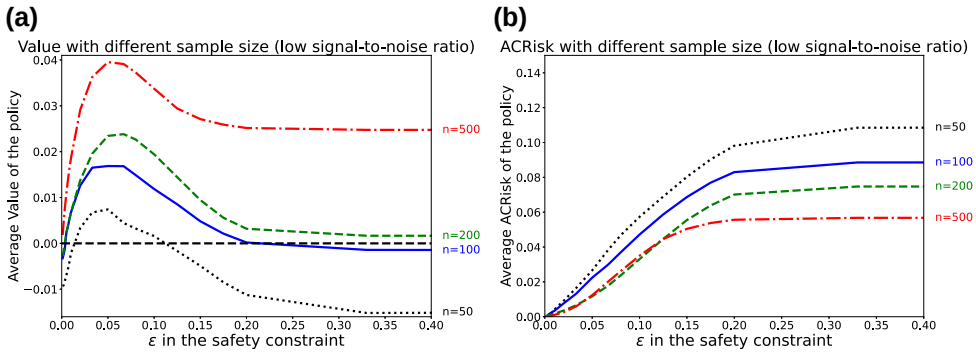

Figure 2. The average value (left panel) and ACRisk (right panel) of the learned policies using the data with covariate overlap, varying the safety constraint ϵ and sample size n. (a) Average Value (b) Average ACRisk.

We consider policies based on a linear separation of the covariate space $\mathcal{X}$:

$$\Delta := \left\{\delta(\cdot\,,\,\cdot) : \delta(x_1, x_2) = I(ax_1 + bx_2 + c > 0);\, a, b, c \in \mathbb{R}\right\} \tag{8}$$

Notice that this policy class does not contain the oracle treatment rule—giving the treatment to individuals with $X_1 > 0$ and $X_2 > 0$. We apply the proposed Bayesian safe policy learning methods, and use the empirical distribution of the covariates $\boldsymbol{X}$ as an approximate CDF of $\boldsymbol{X}$.

For Scenario I with covariate overlap, we use Bayesian Causal Forests (BCF) to estimate the CATE and the expected outcome under the control condition, i.e. $\mathbb{E}(Y(0))$, with the default hyperprior parameter specification suggested by Hahn et al. (2020). For modelling $\mathbb{E}[Y(0)|\boldsymbol{X}]$, we use 200 trees and hyperparameters of $\beta = 2$, $\eta = 0.95$. For modelling the CATE, we impose a stronger prior and use a BART model with 50 trees, and $\beta = 3$, $\eta = 0.25$. Here, β is the penalization hyperparameter for the depth of trees (larger values lead to more shallow trees) and η is the splitting probability hyperparameter (larger values lead to deeper trees).

For Scenario II without covariate overlap, we use GP regression to model the expected outcome under the control condition as well as the CATE. We set the kernel of the Gaussian process as a Matern kernel with parameter $l = 0.5$, 1 or 2, $\nu = 3/2$ (see equation (S1)). We use a noninformative prior where the mean function of the GP, $m(x)$, is set to 0. For other hyperparameters in the kernel function, we set the length scale to $\{0.5, 1, 2\}$ which correspond to a *nonsmooth*, *moderately smooth*, and *highly smooth* CATE. We set σ_0^2 to $\{1, 4, 16\}$ which corresponds to a *strong*, *medium*, and *weak* prior.

For both methods, we use `Stan` (Carpenter et al., 2017) to compute 2,000 posterior samples from two chains with a burn-in period of 500 samples.

5.2 Findings

Since we know the true data generating process, we can calculate the ACRisk for the learned policy under different policy learning setups. Since the proposed methodology controls the PACRisk, it is of interest to investigate how the ACRisk of the learned policy as well as the policy value change as functions of the safety constraint ϵ in equation (5). We also examine the influence of the prior distribution on the performance of the proposed methodology when there is no covariate overlap.

5.2.1 Sample size and signal strength

Figure 2 presents the average value (left panel) and average ACRisk (right panel) of learned policies under different values of the safety constraint ϵ (x-axis) with various sample sizes, using the continuous outcome model. Similarly, Figure S5 in the appendix shows the results with different values of the safety constraint ϵ and signal-to-noise ratio. The data generating process follows

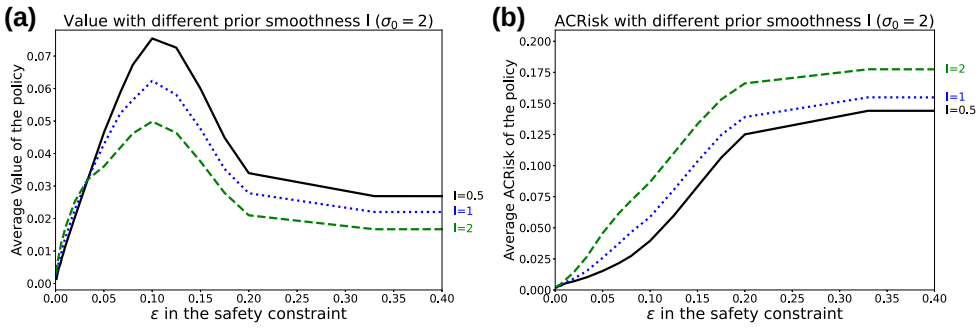

Figure 3. Average Value (left panel) and ACRisk (right panel) for learned policies using data without covariate overlap, varying the safety constraint ϵ and prior smoothness for the CATE l (a greater value corresponds to a greater degree of prior smoothness). (a) Average Value (b) Average ACRisk.

Scenario I, and there is no extrapolation. As expected, we find that a weaker of safety constraint (i.e. a greater value of ϵ) leads to a greater ACRisk of the learned policy, reaching a plateau that corresponds to the ACRisk obtained by maximizing the posterior expected utility with no constraint. Note that for small sample sizes, the average value of learned policies can be negative due to finite sample error even though the baseline policy has value 0.

On average, a smaller sample size and lower signal-to-noise ratio lead to a greater ACRisk. Interestingly, an appropriate level of the safety constraint (around 0.05 in this simulation) achieves a greater average value compared to maximizing the posterior expected value with no constraint. This is especially true when the sample size is small and/or the signal-to-noise ratio is low. In such settings, the safety constraint regularizes the policy optimization, only modifying the existing policy for those who are expected to benefit from a new policy with a high degree of certainty. The ACRisk constraint can reduce the areas that are incorrectly assigned to the new policy due to random noise.

To see how the ACRisk constraint can have a regularization effect, consider the following toy example, in which our baseline policy is the control condition ($d = 0$) and the potential outcome model is $Y(d) = -0.1d + \epsilon$ with $\epsilon \sim N(0, 1)$. In this case, the optimal policy should be the same as the baseline policy for everyone. However, since the treatment effect is quite small relative to the noise, we may incorrectly modify the policy for some cases especially when the sample size is small. The ACRisk constraint enables us to take into account the uncertainty of treatment effect estimation and therefore reduce these incorrect decisions. In Section S4.1 of the online supplementary material, we further explore the regularization impact of the ACRisk constraint in our main simulation specification.

5.2.2 Prior in extrapolation

Next, we investigate the influence of prior parameters in GPs on the performance of the proposed methodology under Scenario II where there is no covariate overlap. Specifically, we consider the scale parameter l and the variance parameter σ_0^2, which determine the smoothness of CATE and prior strength, respectively. Recall that a greater value of l implies a smoother function, while a smaller value of σ_0^2 corresponds to a stronger prior belief. We fix the sample size to 200 and use a low signal-to-noise ratio when generating the data.

Figure 3a shows the average value of the learned policy under different prior smoothness. In general, the value of learned policy tends to decrease as the prior smoothness increases. However, when the safety constraint is strong (i.e. ϵ is small), the value of the learned policy with stronger prior smoothness is slightly greater. A small value of ϵ implies that we only wish to change the baseline policy for a set of covariates that correspond to large CATEs. A smoother prior leads to a more conservative estimate of the CATE and can better identify such covariates.

Furthermore, Figure 3b shows that as the prior for the CATE becomes smoother, we extrapolate further, leading to an increase in the average ACRisk. In Figure S6 of the appendix, we observe a similar pattern when varying the prior strength for estimating the CATE.

To summarize, our simulation study shows that the safety constraint on the PACRisk can effectively control the true ACRisk in policy learning, reducing the risk of harming specified subgroups. In addition, a moderately binding safety constraint can induce beneficial regularization and improve the average value of the learned policy, especially when the sample size is small and the signal-to-noise ratio is low.

6 Empirical analysis of the HES security assessment

With our methodological development in hand, we now return to the analysis of the military security assessment described in Section 2. We will focus on learning a new rule to aggregate the 20 sub-model scores into one overall security score, while keeping the original structure of the HES and only modifying the two-way and three-way tables used for the aggregation.

6.1 Setup

To formalize our problem, let $X^{(k)} \in [1, 5]$ denote the k-th sub-model scores with $k \in \{1, 2, \ldots, 20\}$, where $\boldsymbol{X} \in \mathcal{X}$ denotes the whole 20-dimensional input vector. We use $S^{(i,j)} \in \{1, 2, 3, 4, 5\}$ to represent the jth score at the ith level where $j \in \{1, 2, \ldots, n_i\}$ and $i \in \{1, 2, 3\}$ with n_j representing the total number of scores aggregated at the j-th level. We note that $X^{(k)} \neq S^{(1,k)}$ because the raw input score $X^{(k)}$ is rounded to obtain the first level score $S^{(1,k)}$. To apply the Bayesian safe policy learning method, we use the empirical CDF of $\mathbf{X}$ as an approximate of its population CDF.

The output security score is denoted by $D \in \mathcal{D} = \{1, 2, 3, 4, 5\}$. We use $Y \in \{0, 1\}$ to represent one of three binary regional development outcomes measured in January, 1970: *regional safety*, *regional economy*, and *regional civic society*. These variables are calculated by Dell and Querubin (2018) based on a latent class analysis of data from the HES surveys, using a mixture model to estimate the posterior probability that each Hamlet belongs to one of two latent groups associated with 'high' and 'low' outcome. We analyze each outcome separately. This metric for utility is far from comprehensive in characterizing the complex outcomes for each region, but it is a practical metric that summarizes many survey questions.

We consider a policy δ, which is a deterministic function that maps the inputs $\boldsymbol{x} \in \mathcal{X}$ (the 20 sub-model scores) to an output score $d \in \mathcal{D}$. We define two-way and three-way decision tables as the following functions that map integer-valued input scores to an integer-valued output score. Recall that each input/output score takes an integer value, ranging from 1 to 5.

$$\begin{aligned} T_2(\cdot, \cdot) &: \{1, 2, 3, 4, 5\}^2 \to \{1, 2, 3, 4, 5\}, \\ T_3(\cdot, \cdot, \cdot) &: \{1, 2, 3, 4, 5\}^3 \to \{1, 2, 3, 4, 5\}. \end{aligned} \tag{9}$$

Let $\widetilde{T}_2$ and $\widetilde{T}_3$ represent the baseline decision tables used in the existing HES. These existing rules are *monotonic*, meaning that a decision table with a greater input score never yields a smaller output score, holding the other input scores constant. Formally, a two-way decision table T_2 is monotonic if and only if we have $T_2(i_1, j_1) \leq T_2(i_2, j_2)$ for all $i_1, j_1, i_2, j_2 \in \{1, 2, 3, 4, 5\}$ and $i_1 \leq i_2, j_1 \leq j_2$. Similarly, a three-way decision table T_3 is monotonic if and only if we have $T_3(i_1, j_1, k_1) \leq T_3(i_2, j_2, k_2)$ for all $i_1, j_1, k_1, i_2, j_2, k_2 \in \{1, 2, 3, 4, 5\}$ and $i_1 \leq i_2, j_1 \leq j_2, k_1 \leq k_2$. Recall that the sub-model scores are ordered so that lower values are 'worse.' Thus, the monotonicity condition is reasonable because higher sub-model scores should not make a region less secure. We will require that our learned decision tables also satisfy this monotonicity condition.

6.2 Learning to aggregate military, economic, and political sub-models

We begin by learning a single decision table while keeping the other tables of the HES. In particular, we consider learning a new third-level three-way decision table T_3 that aggregates three scores—the Military, Political, and Social Economic models—and produces the final output security score D (see Figure 1).

We use separate GPs with a logit link function g to model the expected outcome under different decisions (see Section 4.3). Specifically, we assume the model given in equation (7) where

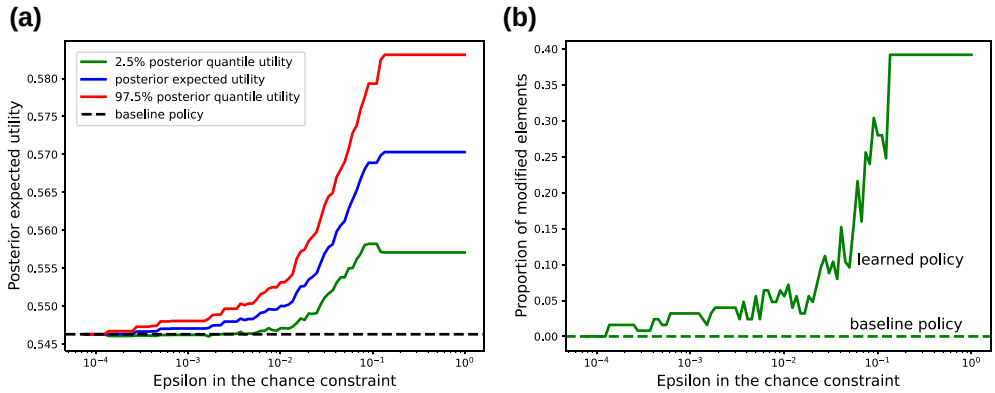

Figure 4. The posterior expected utility of the learned policy (left panel) and the proportion of elements in the three-way table changed by the learned policy (right panel) under different values of ϵ, when regional *safety* development is the outcome. A weaker safety constraint (i.e. a greater value of ϵ) leads to a greater difference between the baseline and learned policies. The posterior expected utility also becomes greater. (a) Posterior expected utility (b) Proportion of changed elements.

$d = 1, 2, 3, 4, 5$. We use f_0 to model the expected outcome under the decision $D = 1$, whereas f_k with $k = 1, 2, 3, 4$ is used to model the effect between two adjacent decisions.

We set the prior mean function for all the GPs to zero, i.e. $m(x) = 0$, implying that all potential outcomes are distributed as Bernoulli with probability 1/2. For the kernel function of the GP, we use a Matern kernel with $\nu = 3/2$, $l = 1$, $\sigma_0^2 = 4$ (see equation (S1)). This corresponds to a weak prior, implying that with ¿90% probability, $f_k(\cdot)$ is no less smooth than a Lipschitz function with a Lipschitz constant of 10.95. Section S5.2 of the online supplementary material presents a sensitivity analysis where we use larger/smaller values of l to induce more/less extrapolation. Overall, these results are consistent with our main findings that are presented below.

Optimizing a single decision table in our application is not computationally demanding. Therefore, we use the Gurobi solver to optimize equation (5) over three-way monotonic decision tables. We use the military impacts on regional *safety*, *economy*, and *civic society* as separate binary outcomes and set the utility function to $u(D, Y) = Y$. Our utility only considers the outcome and does not penalize the use of a lower or higher security score.

Figure 4 shows how the results of our Bayesian safe policy learning procedure change with the safety constraint ϵ when the outcome is regional *safety*. Figure 4a shows the posterior expected utility of the learned policy, while Figure 4b presents the proportion of changed elements in the three-way table. We find that a weaker safety constraint (i.e. a greater value of ϵ) leads to a greater difference between the learned and baseline policies in terms of the changed elements of the three-way table. The learned policy also has a higher posterior expected utility when the safety constraint is weaker. A stronger posterior ACRisk constraint also reduces the width of the credible interval for the posterior expected utility under the policy, since fewer individuals are assigned to the new decision. In addition, Figure 4a shows the posterior expected utility rather than the true utility. This is why we do not observe a reguarlization effect similar to the one shown in Figure 2.

Table 2 shows how the learned security score differs from the original score across regions. Our setup for this result is the same as the one used in Figure 4 with regional safety as the outcome. No region has a change in the security score greater than 1 in its absolute magnitude. This is due to the structure of the HES: changing the original score by ¿1 involves a high degree of extrapolation. Because the prior mean of the effect between two adjacent scores is 0, the further we move away from the existing score, the closer the posterior mean of the effect moves towards 0. In addition, more extrapolation leads to a higher posterior uncertainty and so a higher level of PACRisk. These properties of the posterior combine to limit changes to the original score.

Without constraining the PACRisk (i.e. when $\epsilon = 1$), the learned policy increases the security score for all regions whose baseline security score is either 2 or 3. This finding is consistent with that of Dell and Querubin (2018) that a higher security score improves regional safety. We

Table 2. Change in the security score. Each row corresponds to the original security score. Each cell shows the proportion of regions with their learned security score change where '−1' and '+1' indicate the learned score is one point less and greater than the original score, respectively

Original score	$\epsilon = 1$			$\epsilon = 0.1$			$\epsilon = 0.01$		
	−1	no change	+1	−1	no change	+1	−1	no change	+1
1		100%			100%			100%	
2			100%		53.5%	46.5%		86.2%	13.8%
3			100%		5.5%	94.5%	3.4%	95.2%	1.4%
4		98.4%	1.6%		100%			100%	
5		100%			100%			100%	

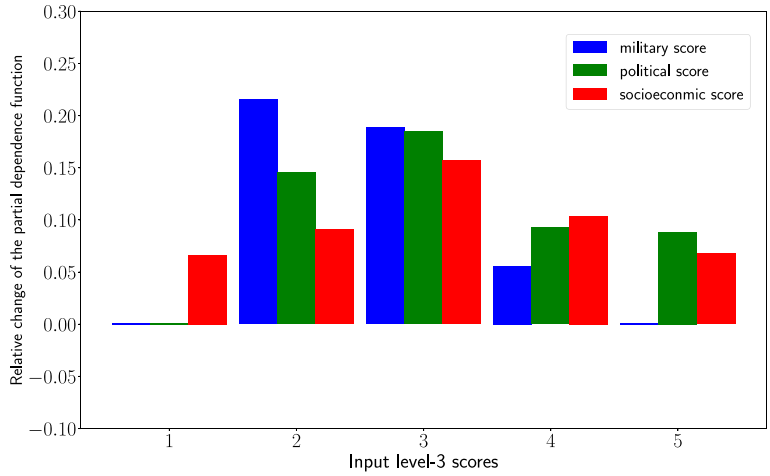

Figure 5. The relative change of the PD function from the baseline policy to the learned policy for $\epsilon = 0.1$. Each block corresponds to a different input of the PD function, and different colours corresponds to different level-3 scores. For example, the first blue bar in the first block corresponds to $\{I_{military}(1;\, T_3) - I_{military}(1;\, \tilde{T}_3)\}/I_{military}(1;\, \tilde{T}_3)$, where T_3 and $\tilde{T}_3$ are the learned and baseline policies, respectively.

find that as the risk constraint ϵ decreases, fewer regions have learned scores different from the original scores. When $\epsilon = 0.01$, the learned policy even decreases the security score for a small proportion of regions whose baseline security score is 3. We note that this happens in part due to the monotonicity constraint imposed on the decision table.

Figure 5 presents the relative difference in the partial dependence (PD) of the baseline and learned policies with respect to the three level-3 scores. The PD function measures the dependence of the output security score on the level-3 scores, i.e. $S^{(3,1)}$, $S^{(3,2)}$, $S^{(3,3)}$, by computing the marginal expected output of the policy given certain input scores (Friedman, 2001; Greenwell et al., 2018). For example, the PD function for policy $\delta(\,\cdot\,;\, T_3)$ with respect to level-3 military score $S^{(3,1)}$ is computed as,

$$I_{military}(x;\, T_3) = \frac{1}{n}\sum_{i=1}^{n} T_3(x,\, S_i^{(3,2)},\, S_i^{(3,3)}).$$

We find the percent change in the PD function between the learned and baseline policies are mostly positive. That is, the expected output security score under the learned policy is generally higher than the baseline policy for all three level-3 input scores. Recall that a higher security score signifies that a region should not be targeted, so increasing the scores would be likely to decrease the frequency of airstrikes, and potentially improve the regional outcome. This finding is consistent with that of Dell and Querubin (2018): airstrikes increased the military and political activities of the communist insurgency, negatively impacting regional safety and the regional economy.

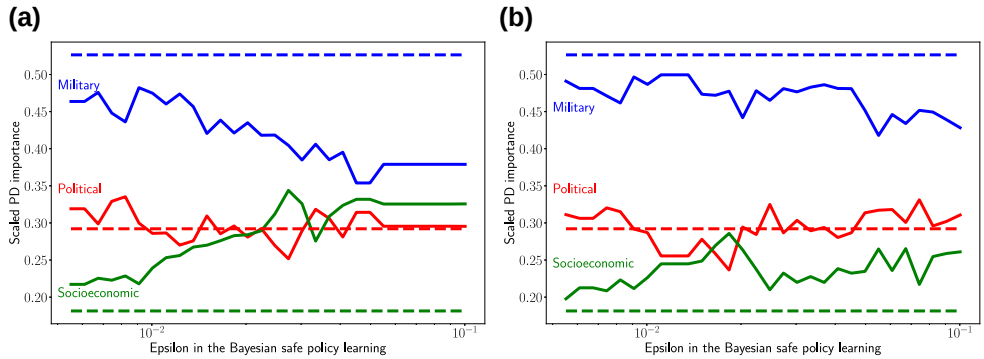

Figure 6. The scaled Partial Dependence (PD) importance of level-3 scores of the learned policy, as a function of the ϵ. The solid line corresponds to the learned policy, and the dashed line indicates the baseline policy. Lines with different colours shows the PD importance of different level-3 scores. (a) Regional economy as outcome (b) Regional safety as outcome.

We further compute the PD importance (Greenwell et al., 2018) of the military, political, and socioeconomic level-3 scores under the learned policy as a function of the safety constraint ϵ, using the three different outcomes. The PD importance measures the degree to which a PD function is sensitive to the input. A small PD importance means that the output score does not strongly depend on the input. For ease of interpretation, we scale the PD importance of level-3 scores for each policy so that they sum to one, allowing us to compare the PD importance of level-3 scores across different policies.

Figure 6 compares the scaled PD importance of each factor (denoted by different colours) between the baseline (dotted lines) and learned (solid lines) policies as a function of the strength of the safety constraint ϵ. The figure shows a consistent pattern: when both the regional economy and regional safety are used as outcomes, the learned policies up-weight the socioeconomic model and down-weight the military model. Figure S12a of the online supplementary material further investigates the PD importance of the different factors when learning policies targeting civic society outcomes. Overall, our analysis suggests that the HES over-emphasized the importance of the military model for all three types of outcomes.

6.3 Learning the entire scoring system

Next, we consider learning both the two-way and three-way tables through all three levels of aggregation. In other words, we consider a policy class with the same aggregation structure as in the original HES, but also allowing the two-way and three-way tables to be modified. Formally, we consider the following policy class, $\Delta_3 = \{\delta(\cdot\,; T_2, T_3) : T_2, T_3 \text{ is monotonic}\}$. We use the same posterior CATE draws obtained in the previous analysis.

Unlike Section 6.2, this complex policy class makes it difficult to directly optimize the objective. In Section S6 of the online supplementary material, we develop an optimization method based on directed acyclic graph (DAG) partitioning to solve this specific problem. As in the previous analysis, we compute the PD importance measure for each sub-model score.

As the policy class Δ_3 is large and contains many different policies, it is difficult to directly interpret the obtained policy through the learned decision tables. Therefore, we use the PD importance measure to understand how the data-derived policy differs from the HES. Figure 7 presents the relative importance of the sub-model scores under the baseline and optimal policies with a safety constraint $\epsilon = 0.1$. Each plot shows the scaled PD importance of the 19 different sub-model scores under the baseline and learned policies for the three different outcomes: regional safety, economy, and civic society. Each arrow shows how the scaled PD importance changes from the baseline policy to the learned policy.

We find that when the outcome is either regional economy or civic society (Figures 7a and 7b), the learned policy puts less weight on the security sub-model scores than the baseline policy, with particularly large declines in the relative importance of enemy military activity and presence.

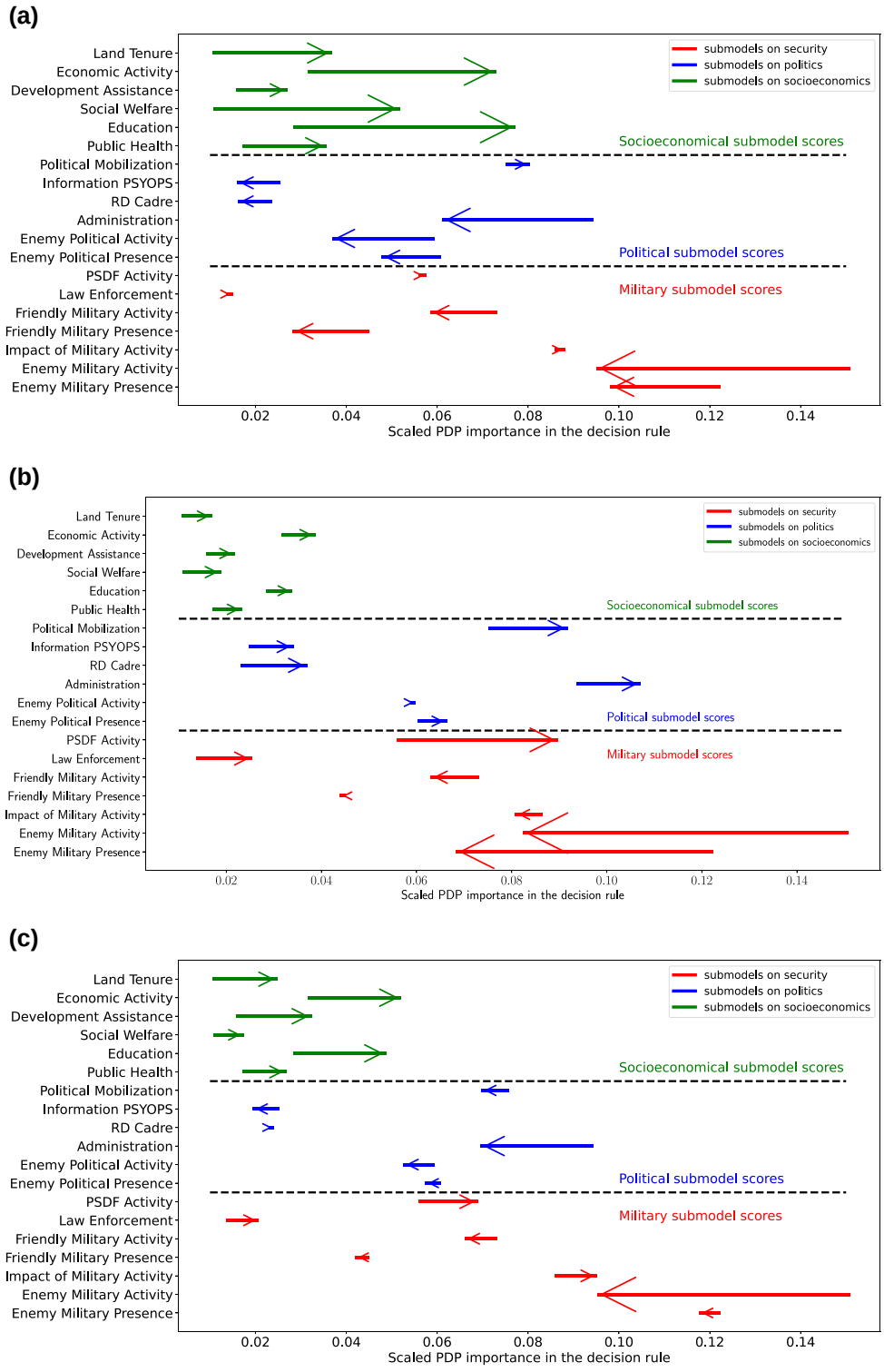

Figure 7. The scaled PD importance of 19 different sub-model scores in the learned policy ($\epsilon = 0.1$) and baseline policy with different outcomes. Each arrow indicates the change of the scaled PDP importance from the baseline policy to the learned policy. (a) Regional economy as outcome (b) Regional civic society as outcome (c) Regional safety as outcome.

Note, however, that this is partially offset by an increase in the relative importance of PSDF (People's Self-Defense Force) activity and law enforcement.

In contrast, the socioeconomic sub-model scores are more important when targeting all three outcomes, and particularly so when targeting the economic outcome. Finally, the change in the relative importance of the political sub-model scores is more mixed; their importance only changes slightly when targeting regional safety (Figure 7c). It generally declines when targeting the regional economy, and increases when targeting regional civic society.

7 Concluding remarks

In this article, we propose a new notion of the ACRisk, which represents the population proportion of groups that would be worse off under a new policy than under the baseline policy. We then develop a Bayesian safe policy learning methodology that limits this risk. We separate the estimation of heterogeneous treatment effects from the optimization of policies, enabling the use of flexible Bayesian nonparametric models while considering a complex policy class.

We apply the proposed methodology to the HES, a military security assessment used during the Vietnam War to guide airstrike decisions for USAF commanders. Substantively, our analysis shows that the HES—which saw active use during the war—attached too much importance to the military security scores, even when targeting military objectives. In contrast, the Bayesian safe policy assigns greater weights to socioeconomic factors. We also develop a stochastic optimization algorithm that is generally applicable to monotonic decision tables. Given the popularity of such decision rules in public policy and other settings, we believe that our optimization algorithm is of independent interest.

There are several directions for further research. First, while we take a Bayesian approach, one may consider a frequentist approach to controlling the ACRisk in policy learning. Second, our simulation study suggests that the posterior ACRisk constraint plays a role of regularization in policy optimization and an appropriate level of risk constraint can improve the average value of the learned policy. Better understanding how regularization improves policy learning is an important next step. Third, it is of interest to extend the proposed Bayesian policy learning framework to other types of constraints such as fairness and budget constraints.

Conflicts of interest: None declared.

Funding

We acknowledge the partial support from Cisco Systems, Inc. (CG# 2370386), National Science Foundation (SES–2051196), and Sloan Foundation (Economics Program; 2020–13946).

Data availability

The replication code and data are available at Harvard Dataverse as Jia et al. (2025).

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series A*.

Appendix: Proof of Theorem 1

For simplicity, we will denote the triplet of $\{\boldsymbol{X}_i, D_i, Y_i\}$ as Z_i. Therefore, our goal is to show that the original chance-constraint optimization problem (see equation (5))

$$\begin{aligned} \delta_{\text{safe}} = \underset{\delta \in \Delta}{\text{argmax}} \quad & \mathbb{E}\big[V(\delta; \boldsymbol{\Theta}) \mid \{Z_i\}_{i=1}^n\big], \\ \text{subject to} \quad & \mathbb{E}\big[R(\delta, \tilde{\delta}; \boldsymbol{\Theta}) \mid \{Z_i\}_{i=1}^n\big] \leq \epsilon, \end{aligned} \tag{A1}$$

can be equivalently written as,

$$\begin{aligned}\delta_{\text{safe}} = \underset{\delta\in\Delta}{\text{argmax}} \quad & \sum_{k=0}^{K-1} \int I(\delta(\boldsymbol{x}) = k) b_k(\boldsymbol{x}) dF(\boldsymbol{x}) \\ \text{subject to} \quad & \sum_{k=0}^{K-1} \int I(\delta(\boldsymbol{x}) = k) r_k(\boldsymbol{x}) dF(\boldsymbol{x}) \le \epsilon,\end{aligned} \tag{A2}$$

where $F(\boldsymbol{x})$ be the CDF of the covariate $\boldsymbol{X}$, and

$$\begin{aligned}\tau_k(\boldsymbol{x}, \boldsymbol{\theta}) &:= \mathbb{E}\big[u(k, Y(k)) - u(\tilde{\delta}(\boldsymbol{x}), Y(\tilde{\delta}(\boldsymbol{x}))) \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{\Theta} = \boldsymbol{\theta}\big], \\ b_k(\boldsymbol{x}) &:= \mathbb{E}\big[\tau_k(\boldsymbol{x}, \boldsymbol{\Theta}) \mid \{Z_i\}_{i=1}^n\big], \\ r_k(\boldsymbol{x}) &:= \mathbb{P}\big(\tau_k(\boldsymbol{x}, \boldsymbol{\Theta}) < 0 \mid \{Z_i\}_{i=1}^n\big).\end{aligned}$$

We first show that the constraint given in equation (A1) is equivalent to equation (A2). By the definition of the PACRisk, we have

$$\begin{aligned}R(\delta, \tilde{\delta}; \boldsymbol{\Theta}) &= \mathbb{P}\big(\mathbb{E}[u(\delta(\boldsymbol{X}), Y(\delta(\boldsymbol{X}))) \mid \boldsymbol{X}, \boldsymbol{\Theta}] < \mathbb{E}\big[u(\tilde{\delta}(\boldsymbol{X}), Y(\tilde{\delta}(\boldsymbol{X}))) \mid \boldsymbol{X}, \boldsymbol{\Theta}\big] \,\big|\boldsymbol{\Theta}\big) \\ &= \mathbb{P}\big(\mathbb{E}\big[u(\delta(\boldsymbol{X}), Y(\delta(\boldsymbol{X}))) - u(\tilde{\delta}(\boldsymbol{X}), Y(\tilde{\delta}(\boldsymbol{X}))) \mid \boldsymbol{X}, \boldsymbol{\Theta}\big] < 0 \,\big|\boldsymbol{\Theta}\big) \\ &= \mathbb{P}\left(\sum_{k=0}^{K-1} \tau_k(\boldsymbol{X}, \boldsymbol{\Theta}) I(\delta(\boldsymbol{X}) = k) < 0 \,\middle|\, \boldsymbol{\Theta}\right)\end{aligned} \tag{A3}$$

Therefore, equation (A1) can be written as

$$\begin{aligned}&\mathbb{E}\big[R(\delta, \tilde{\delta}; \boldsymbol{\Theta}) \mid \{Z_i\}_{i=1}^n\big] \\ &= \mathbb{E}\left[\mathbb{P}\left(\sum_{k=0}^{K-1} \tau_k(\boldsymbol{X}, \boldsymbol{\Theta}) I(\delta(\boldsymbol{X}) = k) < 0 \,\middle|\, \boldsymbol{\Theta}\right) \middle| \{Z_i\}_{i=1}^n\right] \\ &= \mathbb{E}_{\boldsymbol{\Theta}}\left[\mathbb{E}_{\boldsymbol{X}}\left[I\left\{\sum_{k=0}^{K-1} \tau_k(\boldsymbol{X}, \boldsymbol{\Theta}) I(\delta(\boldsymbol{X}) = k) < 0\right\} \,\middle|\, \boldsymbol{\Theta}\right] \middle| \{Z_i\}_{i=1}^n\right]\end{aligned} \tag{A4}$$

$$\begin{aligned}&= \mathbb{E}_{\boldsymbol{X},\boldsymbol{\Theta}}\left[I\left\{\sum_{k=0}^{K-1} \tau_k(\boldsymbol{X}, \boldsymbol{\Theta}) I(\delta(\boldsymbol{X}) = k) < 0\right\} \,\middle|\, \{Z_i\}_{i=1}^n\right] \\ &= \mathbb{E}_{\boldsymbol{X}}\left[\mathbb{E}_{\boldsymbol{\Theta}}\left[I\left\{\sum_{k=0}^{K-1} \tau_k(\boldsymbol{X}, \boldsymbol{\Theta}) I(\delta(\boldsymbol{X}) = k) < 0\right\} \,\middle|\, \{Z_i\}_{i=1}^n, \boldsymbol{X}\right] \middle| \{Z_i\}_{i=1}^n\right]\end{aligned} \tag{A5}$$

$$= \mathbb{E}_{\boldsymbol{X}}\left[\sum_{k=0}^{K-1} I(\delta(\boldsymbol{X}) = k) \mathbb{E}_{\boldsymbol{\Theta}}\big[I\big\{\tau_k(\boldsymbol{X}, \boldsymbol{\Theta}) < 0\big\} \,\big|\, \{Z_i\}_{i=1}^n, \boldsymbol{X}\big] \,\middle|\, \{Z_i\}_{i=1}^n\right] \tag{A6}$$

$$\begin{aligned}&= \mathbb{E}_{\boldsymbol{X}}\left[\sum_{k=0}^{K-1} I(\delta(\boldsymbol{X}) = k) r_k(\boldsymbol{X}) \,\middle|\, \{Z_i\}_{i=1}^n\right] \\ &= \sum_{k=0}^{K-1} \int I(\delta(\boldsymbol{x}) = k) r_k(\boldsymbol{x}) dF(\boldsymbol{x}).\end{aligned}$$

Here I use the subscript of $\boldsymbol{X}$ and $\boldsymbol{\Theta}$ to explicitly denote the expectation being taken over the randomness of $\boldsymbol{X}$ of $\boldsymbol{\Theta}$. From equation (A4) to equation (A5), we use the law of iterated expectation to switch the order of the inner and outer expectations. From (A5) to (A6), we utilize the fact that there's only one nonzero term in the summation $\sum_{k=0}^{K-1} \tau_k(\boldsymbol{X}, \boldsymbol{\Theta}) I(\delta(\boldsymbol{X}) = k)$, which simplifies the indicator function.

Similarly, for the optimization target, we have

$$\begin{aligned} &\mathbb{E}\left[V(\delta; \boldsymbol{\Theta}) \mid \{Z_i\}_{i=1}^n\right] \\ &\quad = \mathbb{E}\left[u(\delta(\boldsymbol{X}),\ Y(\delta(\boldsymbol{X}))) \mid \{Z_i\}_{i=1}^n\right] \end{aligned} \tag{A7}$$

$$\begin{aligned} &= \mathbb{E}\left[\mathbb{E}\left\{u(\delta(\boldsymbol{X}),\ Y(\delta(\boldsymbol{X}))) \mid \{Z_i\}_{i=1}^n, \boldsymbol{X}\right\} \middle| \{Z_i\}_{i=1}^n\right] \\ &= \mathbb{E}\left[\mathbb{E}\left\{\sum_{k=0}^{K-1} u(\delta(\boldsymbol{X}),\ Y(k)) I(\delta(\boldsymbol{X}) = k) \middle| \{Z_i\}_{i=1}^n, \boldsymbol{X}\right\} \middle| \{Z_i\}_{i=1}^n\right] \\ &= \mathbb{E}_X\left[\mathbb{E}_{\boldsymbol{\Theta}}\left\{\sum_{k=0}^{K-1} \tau_k(\boldsymbol{X}, \boldsymbol{\Theta}) I(\delta(\boldsymbol{X}) = k) \middle| \{Z_i\}_{i=1}^n, \boldsymbol{X}\right\} \middle| \{Z_i\}_{i=1}^n\right] + \text{const.} \\ &= \mathbb{E}_X\left[\sum_{k=0}^{K-1} I(\delta(\boldsymbol{X} = k) b_k(\boldsymbol{X}) \middle| \{Z_i\}_{i=1}^n\right] + \text{const.} \end{aligned} \tag{A8}$$

$$= \sum_{k=0}^{K-1} \int I(\delta(\boldsymbol{x}) = k) b_k(\boldsymbol{x}) dF(\boldsymbol{x}) + \text{const.}$$

where const. represents the term that is not a function of δ. In equation (A7), the expectation is taken over the distribution of X, the posterior distribution of $\boldsymbol{\Theta}$, and the conditional distribution of Y given X, $\boldsymbol{\Theta}$. Then, we first compute the expectation over $\boldsymbol{\Theta}$ and the conditional distribution of Y given X. Taking the expectation over the distribution of X gives equation (A8). This completes the proof. □

References

Athey S., & Wager S. (2021). Policy learning with observational data. *Econometrica: Journal of the Econometric Society*, *89*(1), 133–161. https://doi.org/10.3982/ECTA15732

Bai C., Wang L., Yang Z., Deng Z., Garg A., Liu P., & Wang Z. (2022). Pessimistic bootstrapping for uncertainty-driven offline reinforcement learning. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Y4cs1Z3HnqL

Ben-Michael E., Greiner D. J., Huang M., Imai K., Jiang Z., & Shin S. (2024a). 'Does AI help humans make better decisions? A statistical evaluation framework for experimental and observational studies', arXiv, arXiv:2403.12108, preprint: not peer reviewed. https://doi.org/10.48550/arXiv.2403.12108

Ben-Michael E., Greiner D. J., Imai K., & Jiang Z. (2025). Safe policy learning through extrapolation: Application to pretrial risk assessment. *Journal of the American Statistical Association*, 1–14. https://doi.org/10.1080/01621459.2025.2489135

Ben-Michael E., Imai K., & Jiang Z. (2024b). Policy learning with asymmetric counterfactual utilities. *Journal of the American Statistical Association*, *119*(548), 3045–3058. https://doi.org/10.1080/01621459.2023.2300507

Beygelzimer A., & Langford J. (2009). The offset tree for learning with partial labels. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 129–138).

Branson Z., Rischard M., Bornn L., & Miratrix L. W. (2019). A nonparametric Bayesian methodology for regression discontinuity designs. *Journal of Statistical Planning and Inference*, *202*(2), 14–30. https://doi.org/10.1016/j.jspi.2019.01.003

Buckman J., Gelada C., & Bellemare M. G. (2020). 'The importance of pessimism in fixed-dataset policy optimization', arXiv, arXiv:2009.06799, preprint: not peer reviewed. https://doi.org/10.48550/arXiv.2009.06799

Carpenter B., Gelman A., Hoffman M. D., Lee D., Goodrich B., Betancourt M., Brubaker M., Guo J., Li P., & Riddell A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

Chen J., & Jiang N. (2022). Offline reinforcement learning under value and density-ratio realizability: The power of gaps. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence* (pp. 378–388).

Chipman H. A., George E. I., & McCulloch R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, *4*(1), 266–298. https://doi.org/10.1214/09-AOAS285

Daddis G. A. (2012). The problem of metrics: Assessing progress and effectiveness in the Vietnam War. *War in History*, *19*(1), 73–98. https://doi.org/10.1177/0968344511422312

Delage E., & Mannor S. (2007). Percentile optimization in uncertain Markov decision processes with application to efficient exploration. In *Proceedings of the 24th International Conference on Machine Learning* (pp. 225–232). Association for Computing Machinery.

Delage E., & Mannor S. (2010). Percentile optimization for Markov decision processes with parameter uncertainty. *Operations Research*, *58*(1), 203–213. https://doi.org/10.1287/opre.1080.0685

Dell M., & Querubin P. (2018). Nation building through foreign intervention: Evidence from discontinuities in military strategies. *The Quarterly Journal of Economics*, *133*(2), 701–764. https://doi.org/10.1093/qje/qjx037

Ding P., & Li F. (2018). Causal inference: A missing data perspective. *Statistical Science*, *33*(2), 214–237. https://doi.org/10.1214/18-STS645

Dudík M., Langford J., & Li L. (2011). Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning*. Omnipress.

Farina M., Giulioni L., & Scattolini R. (2016). Stochastic linear model predictive control with chance constraints – a review. *Journal of Process Control*, *44*(2), 53–67. https://doi.org/10.1016/j.jprocont.2016.03.005

Filar J. A., Krass D., & Ross K. W. (1995). Percentile performance criteria for limiting average Markov decision processes. *IEEE Transactions on Automatic Control*, *40*(1), 2–10. https://doi.org/10.1109/9.362904

Friedman J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Garcıa J., & Fernández F. (2015). A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, *16*(1), 1437–1480. https://doi.org/10.5555/2789272.2886795

Geibel P., & Wysotzki F. (2005). Risk-sensitive reinforcement learning applied to control under constraints. *Journal of Artificial Intelligence Research*, *24*(1), 81–108. https://doi.org/10.1613/jair.1666

Greenwell B. M., Boehmke B. C., & McCarthy A. J. (2018). 'A simple and effective model-based variable importance measure', arXiv, arXiv:1805.04755, preprint: not peer reviewed. https://doi.org/10.48550/arXiv.1805.04755

Greiner J., Halen R., Stubenberg M., & Griffin C. L. (2020). Randomized control trial evaluation of the implementation of the PSA-DMF system in Dane county, WI.

Hahn P. R., Murray J. S., & Carvalho C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, *15*(3), 965–1056. https://doi.org/10.1214/19-BA1195

Hill J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, *20*(1), 217–240. https://doi.org/10.1198/jcgs.2010.08162

Imai K., Jiang Z., Greiner D. J., Halen R., & Shin S. (2023). Experimental evaluation of computer-assisted human decision-making: Application to pretrial risk assessment instrument (with discussion). *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, *186*(2), 167–189. https://doi.org/10.1093/jrsssa/qnad010

Javaid M., Haleem A., Singh R. P., Suman R., & Rab S. (2022). Significance of machine learning in healthcare: Features, pillars and applications. *International Journal of Intelligent Networks*, *3*(11), 58–73. https://doi.org/10.1016/j.ijin.2022.05.002

Jia Z., Ben-Michael E., & Imai K. (2025). Replication code for: Bayesian safe policy learning with chance constrained optimization: Application to military security assessment during the Vietnam War. *Harvard Dataverse V1*. https://doi.org/10.7910/DVN/SPR616.

Jin Y., Ren Z., Yang Z., & Wang Z. (2022). 'Policy learning "without" overlap: Pessimism and generalized empirical Bernstein's inequality', arXiv, arXiv:2212.09900, preprint: not peer reviewed. https://doi.org/10.48550/arXiv.2212.09900

Jin Y., Yang Z., & Wang Z. (2021). Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning* (pp. 5084–5096). PMLR.

Kallus N. (2018). Balanced policy evaluation and learning. *Advances in Neural Information Processing Systems*, *31*, 8909–8920. https://doi.org/10.5555/3327546.3327566

Kallus N. (2022). What's the harm? Sharp bounds on the fraction negatively affected by treatment. *Advances in Neural Information Processing Systems*, *35*, 15996–16009. https://doi.org/10.48550/arXiv.2205.10327

Kallus N., & Zhou A. (2021). Minimax-optimal policy learning under unobserved confounding. *Management Science*, *67*(5), 2870–2890. https://doi.org/10.1287/mnsc.2020.3699
Kamath P. S., Wiesner R. H., Malinchoc M., Kremers W., Therneau T. M., Kosberg C. L., D'Amico G., Dickson E. R., & Kim W. R. (2001). A model to predict survival in patients with end-stage liver disease. *Hepatology*, *33*(2), 464–470. https://doi.org/10.1053/jhep.2001.22172
Kasy M. (2018). Optimal taxation and insurance using machine learning—sufficient statistics and beyond. *Journal of Public Economics*, *167*(8), 205–219. https://doi.org/10.1016/j.jpubeco.2018.09.002
Kitagawa T., & Tetenov A. (2018). Who should be treated? Empirical welfare maximization methods for treatment choice. *Econometrica: Journal of the Econometric Society*, *86*(2), 591–616. https://doi.org/10.3982/ECTA13288
Kizilcec R. F., & Lee H. (2022). Algorithmic fairness in education. In *The Ethics of Artificial Intelligence in Education* (pp. 174–202). Routledge.
Kundu S. S., Sarkar D., Jana P., & Kole D. K. (2021). Personalization in education using recommendation system: An overview. *Computational Intelligence in Digital Pedagogy*, *197*, 85–111. https://doi.org/10.1007/978-981-15-8744-3_5
Lauer J. (2017). *Creditworthy: A history of consumer surveillance and financial identity in America*. Columbia University Press.
Li A., Jiang S., Sun Y., & Pearl J. (2022a). 'Unit selection: Learning benefit function from finite population data', arXiv, arXiv:2210.08203, preprint: not peer reviewed. https://doi.org/10.48550/arXiv.2210.08203
Li F., Ding P., & Mealli F. (2022b). Bayesian causal inference: A critical review. *Philosophical Transactions of the Royal Society A*, *381*(2247), 20220153. https://doi.org/10.1098/rsta.2022.0153
Li L., Chu W., Langford J., & Schapire R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web (WWW '10)* (pp. 661–670). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/1772690.1772758
Luedtke A. R., & Van Der Laan M. J. (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Annals of Statistics*, *44*(2), 713. https://doi.org/10.1214/15-AOS1384
Manski C. F. (2007). Minimax-regret treatment choice with missing outcome data. *Journal of Econometrics*, *139*(1), 105–115. https://doi.org/10.1016/j.jeconom.2006.06.006
Mowbray M., Petsagkourakis P., del Rio-Chanona E. A., & Zhang D. (2022). Safe chance constrained reinforcement learning for batch process control. *Computers & Chemical Engineering*, *157*(11), 107630. https://doi.org/10.1016/j.compchemeng.2021.107630
Nahum-Shani I., Smith S. N., Spring B. J., Collins L. M., Witkiewitz K., Tewari A., & Murphy S. A. (2018). Just-in-time adaptive interventions (JITAIs) in mobile health: Key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine*, *52*(6), 446–462. https://doi.org/10.1007/s12160-016-9830-8
Neyman J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, *5*(4), 465–472. https://doi.org/10.1214/ss/1177012031
PACAF H. Q. (1969). TACC fragging procedures. *Project CHECO Southeast Asia Report*.
Pu H., & Zhang B. (2020). Estimating optimal treatment rules with an instrumental variable: A partial identification learning approach. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *83*(2), 318–345. https://doi.org/10.1111/rssb.12413
Qian M., & Murphy S. A. (2011). Performance guarantees for individualized treatment rules. *Annals of Statistics*, *39*(2), 1180. https://doi.org/10.1214/10-AOS864
Rashidinejad P., Zhu B., Ma C., Jiao J., & Russell S. (2021). Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, *34*, 11702–11716. https://doi.org/10.1109/TIT.2022.3185139
Rasmussen C. E., & Williams C. K. I. (2006). *Gaussian processes for machine learning*. The MIT Press.
Rosenbaum P. R., & Rubin D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41
Rubin D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688. https://doi.org/10.1037/h0037350
Rubin D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, *75*(371), 591–593. https://doi.org/10.2307/2287653
Sato M., Kimura H., & Kobayashi S. (2001). TD algorithm for the variance of return and mean-variance reinforcement learning. *Transactions of the Japanese Society for Artificial Intelligence = Jinko Chino Gakkai Ronbunshi*, *16*(3), 353–362. https://doi.org/10.1527/tjsai.16.353
Schwarm A. T., & Nikolaou M. (1999). Chance-constrained model predictive control. *AIChE Journal*, *45*(8), 1743–1752. https://doi.org/10.1002/aic.v45:8
Schwartz E. M., Bradlow E. T., & Fader P. S. (2017). Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, *36*(4), 500–522. https://doi.org/10.1287/mksc.2016.1023

Shi L., Li G., Wei Y., Chen Y., & Chi Y. (2022). Pessimistic Q-learning for offline reinforcement learning: Towards optimal sample complexity. In *Proceedings of the 39th International Conference on Machine Learning* (pp. 19967–20025).

Smith C. M. (2005). Origin and uses of primum non nocere—above all, do no harm! *Journal of Clinical Pharmacology*, *45*(4), 371–377. https://doi.org/10.1177/0091270004273680

Swaminathan A., & Joachims T. (2015). Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research: JMLR*, *16*(1), 1731–1755. https://doi.org/10.5555/2789272.2886805

Taddy M., Gardner M., Chen L., & Draper D. (2016). A nonparametric bayesian analysis of heterogenous treatment effects in digital experimentation. *Journal of Business & Economic Statistics*, *34*(4), 661–672. https://doi.org/10.1080/07350015.2016.1172013

Tang L., Rosales R., Singh A., & Agarwal D. (2013). Automatic AD format selection via contextual bandits. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management* (pp. 1587–1594). Association for Computing Machinery. https://doi.org/10.1145/2505515.2514700

Uehara M., & Sun W. (2021). 'Pessimistic model-based offline reinforcement learning under partial coverage', arXiv, arXiv:2107.06226, preprint: not peer reviewed. https://doi.org/10.48550/arXiv.2107.06226

Vakili S., & Zhao Q. (2015). Mean-variance and value at risk in multi-armed bandit problems. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)* (pp. 1330–1335). IEEE.

Vitus M. P., Zhou Z., & Tomlin C. J. (2015). Stochastic control with uncertain parameters via chance constrained control. *IEEE Transactions on Automatic Control*, *61*(10), 2892–2905. https://doi.org/10.1109/TAC.2015.2511587

Wei D., Nair R., Dhurandhar A., Varshney K. R., Daly E. M., & Singh M. (2022). On the safety of interpretable machine learning: A maximum deviation approach. *Advances in Neural Information Processing Systems*, *35*(2022), 9866–9880. https://doi.org/10.5555/3600270.3600987

Wiens J., Saria S., Sendak M., Ghassemi M., Liu V. X., Doshi-Velez F., Jung K., Heller K., Kale D., Saeed M., *et al.* (2019). Do no harm: A roadmap for responsible machine learning for health care. *Nature Medicine*, *25*(9), 1337–1340. https://doi.org/10.1038/s41591-019-0548-6

Xie T., Cheng C.-A., Jiang N., Mineiro P., & Agarwal A. (2021). Bellman-consistent pessimism for offline reinforcement learning. *Advances in Neural Information Processing Systems*, *34*, 6683–6694. https://doi.org/10.5555/3540261.3540773

Yan Y., Li G., Chen Y., & Fan J. (2022). The efficacy of pessimism in asynchronous Q-learning. *IEEE Transactions on Information Theory*, *69*(11), 7185–7219. https://doi.org/10.1109/TIT.2023.3299840

Yata K. (2021). 'Optimal decision rules under partial identification', arXiv, arXiv:2111.04926, preprint: not peer reviewed. https://doi.org/10.48550/arXiv.2111.04926

Yin M., & Wang Y.-X. (2021). Towards instance-optimal offline reinforcement learning with pessimism. *Advances in Neural Information Processing Systems*, *34*, 4065–4078. https://doi.org/10.5555/3540261.3540572

Zanette A., Wainwright M. J., & Brunskill E. (2021). Provable benefits of actor-critic methods for offline reinforcement learning. *Advances in Neural Information Processing Systems*, *34*, 13626–13640. https://doi.org/10.5555/3540261.3541305

Zhang B., Tsiatis A. A., Davidian M., Zhang M., & Laber E. (2012). Estimating optimal treatment regimes from a classification perspective. *Stat*, *1*(1), 103–114. https://doi.org/10.1002/sta4.v1.1

Zhang Y., Ben-Michael E., & Imai K. (2022). 'Safe policy learning under regression discontinuity designs', arXiv, arXiv:2208.13323, preprint: not peer reviewed. https://doi.org/10.48550/arXiv.2208.13323

Zhao Y., Zeng D., Rush A. J., & Kosorok M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, *107*(499), 1106–1118. https://doi.org/10.1080/01621459.2012.695674

Zhou X., Mayer-Hamblett N., Khan U., & Kosorok M. R. (2017). Residual weighted learning for estimating individualized treatment rules. *Journal of the American Statistical Association*, *112*(517), 169–187. https://doi.org/10.1080/01621459.2015.1093947

Zhou Z., Athey S., & Wager S. (2022). Offline multi-action policy learning: Generalization and optimization. *Operations Research*, *71*(1), 148–183. https://doi.org/10.1287/opre.2022.2271