

EDUCATION

Fairness-Constrained Distributionally Robust Causal Policy Optimization for Adaptive Instruction in Higher Education

Sanjeev Vijay

Department of Education, Jagannath University, Jaipur, Rajasthan, India

Published October 2026

Abstract

Adaptive instructional policies in higher education often optimize short-term engagement at the expense of durable learning gains, and they frequently neglect heterogeneity across student subpopulations. We propose a fairness-constrained distributionally robust causal policy optimization framework that replaces standard deep reinforcement learning agents with a causally grounded decision-making pipeline. A structural causal model of the educational environment is constructed first, and doubly robust estimation is then applied to observational log data so that confounding bias in the counterfactual evaluation of instructional actions is mitigated. To guard against distribution shifts and adversarial perturbations in student populations, the policy optimization problem is formulated as a distributionally robust objective that maximizes worst-case expected utility over an ambiguity set centered on the empirical distribution. The policy itself is represented by reduced-rank policy trees, which preserve interpretability while remaining computationally tractable, and multi-objective Bayesian optimization is employed to balance average performance, cross-subgroup variance, and explicit disparity constraints. Fairness is enforced through Lagrangian relaxation, while causal effect decomposition separates direct instructional effects from those mediated by motivation or cognitive load. Periodic off-policy evaluation and counterfactual subgroup audits provide closed-loop validation, triggering retraining or constraint tightening when inequities emerge. The main contribution is an integrated methodology that unites causal inference, distributional robustness, and subgroup fairness within a single policy learning loop for higher education. This approach is significant because it offers a principled path toward instructional policies that remain effective, equitable, and auditable throughout their deployment lifecycle.

Keywords: distributionally robust optimization; causal policy learning; fairness-aware adaptive instruction; higher education; personalized learning

1 Introduction

Adaptive learning systems in higher education have evolved from rule-based engines into data-driven frameworks that personalize instructional strategies at scale [1]. Foundational work in Intelligent Tutoring Systems established the value of tailored scaffolding and feedback, while Bayesian Knowledge Tracing and Item Response Theory provided the probabilistic machinery for modeling student proficiency [2]. The integration of Reinforcement Learning (RL) into educational contexts marked a significant shift, where methods such as Deep Q-Networks and Policy Gradient algorithms were applied to optimize long-term learning outcomes by treating instruction as a sequential decision-making process [3]. However, standard RL often suffers from sample inefficiency and instability when applied to sparse educational interaction logs, which has motivated interest in Model-Based and Offline RL techniques that learn policies from historical data without active exploration [4].

A critical limitation in applying these models to education is the presence of confounding variables. Observed correlations between instructional actions and learning outcomes do not imply causation, because students self-select into activities and instructors adapt to perceived student states [5]. Causal inference methodologies, particularly Potential Outcomes Frameworks and Structural Causal Models, have therefore been adopted to correct for selection bias using Inverse Propensity Scoring and Doubly Robust Estimation [6]. These estimators are essential for moving beyond predictive accuracy toward prescriptive analytics, ensuring that recommended actions actually cause improvements rather than merely correlating with them [7].

Simultaneously, the field has recognized that optimizing solely for test scores or completion rates can lead to inequitable outcomes for diverse student populations [8]. This concern has driven the development of fairness constraints in machine learning, drawing from concepts such as Demographic Parity, Equalized Odds, and Counterfactual Fairness [9]. In policy optimization, Multi-Objective Reinforcement Learning has emerged as a framework to balance competing goals, such as maximizing engagement while minimizing performance disparities across subgroups [10]. Algorithms like Pareto Front Optimization and Constrained Markov Decision Processes allow policymakers to define strict bounds on disparity metrics, ensuring that efficiency gains do not come at the cost of equity [11].

To ensure these complex policies are deployable in real-world educational settings, Explainable AI techniques have become indispensable. Methods such as SHAP values, LIME, and Attention Mechanisms provide transparency into why a system recommends a particular hint or difficulty adjustment [12]. Furthermore, interpretable models, including Rule Lists and Decision Trees, have gained traction because they offer inherent explainability compared to black-box neural networks [13]. Recent advances in Reduced-Rank Regression and Low-Rank Matrix Factorization have been utilized to simplify high-dimensional policy spaces, making it computationally feasible to learn robust rules that generalize well under distribution shifts [14]. Distributionally Robust Optimization, borrowed from operations research and control theory, ensures that policies perform well even under worst-case scenario distributions, which is crucial in education where student demographics and curriculum standards may change over time [15]. Additionally, Long-Term Feedback Simulation and Off-Policy Evaluation methods, such as Self-Normalized Importance Sampling, are used to validate policies before deployment, reducing the risk of negative impacts on current students [16].

We propose a fairness-constrained distributionally robust causal policy optimization framework for adaptive instruction in higher education. Unlike conventional RL approaches that may suffer from confounding bias or neglect subgroup equity, our method models adaptive instruction as a constrained multi-objective decision process. The core innovation lies in the use of reduced-rank policy trees guided by multi-objective Bayesian optimization, which learns action rules under strict disparity and distribution-shift constraints. Furthermore, the method incorporates causal effect decomposition and long-term feedback simulation to penalize policies that amplify historical biases, ensuring deployable and equitable instruction through periodic off-policy and counterfactual subgroup audits. The proposed framework differs from existing work in three key respects: it unifies causal estimation, distributional robustness, and subgroup fairness within a single policy learning loop; it replaces opaque neural policies with interpretable reduced-rank trees; and it provides a closed-loop audit mechanism that triggers retraining or constraint tightening when inequities emerge.

The remainder of this paper reviews related work on adaptive learning, causal inference and fair policy optimization; presents the preliminaries and problem formulation; details the proposed

fairness-constrained distributionally robust causal policy learning method; describes the experimental setup; reports the results and analysis; and discusses implications and future work before concluding.

2 Related work

Research on adaptive instruction in higher education spans several intertwined strands: reinforcement learning for sequential pedagogical decision-making, causal inference for observational educational data, distributionally robust optimization under population shift, and fairness-aware policy learning. We organize the discussion around these strands and then position our contribution relative to them.

2.1 Reinforcement learning for instructional decision-making

The treatment of instruction as a sequential decision problem has a long history in intelligent tutoring and has more recently been absorbed into the reinforcement learning paradigm. Early systems relied on hand-crafted pedagogical rules, whereas contemporary approaches learn policies that trade off immediate scaffolding against long-term mastery. Deep Q-networks and policy gradient methods have been applied to problem sequencing and hint generation, and offline reinforcement learning has been proposed to learn from historical interaction logs without risky live exploration [4]. A recurring difficulty is that educational reward signals are sparse, delayed, and frequently proxied by engagement rather than learning, which biases learned policies toward short-term behavioral signals [3]. Model-based variants partially alleviate sample inefficiency, yet they inherit the same reward misspecification problem. Our framework departs from this line by defining rewards through causally estimated learning gains rather than raw engagement, and by constraining the resulting policy with explicit disparity bounds.

2.2 Causal inference in educational data

Observational logs are confounded because students self-select into activities and instructors adapt to perceived student states [5]. Potential outcomes frameworks and structural causal models have therefore been introduced to educational analytics, with inverse propensity scoring and doubly robust estimation used to correct selection bias [6]. These estimators support prescriptive rather than merely predictive conclusions, which is essential when the goal is to recommend actions that genuinely cause improvement [7]. Nevertheless, most existing applications estimate average causal effects and stop there; they rarely embed the estimator inside a policy optimization loop, and they seldom decompose effects into direct and mediated pathways. Our method closes this gap by using doubly robust scores as the reward signal for policy learning and by explicitly separating direct instructional effects from those mediated by motivation and cognitive load.

2.3 Distributional robustness and fairness in policy learning

Distributionally robust optimization has been developed to protect decisions against ambiguity in the underlying data-generating distribution, typically by maximizing worst-case expected utility over an ambiguity set defined through Wasserstein or Kullback-Leibler divergence [15]. In parallel, fairness-aware machine learning has produced criteria such as demographic parity, equalized odds, and counterfactual fairness, together with constrained Markov decision process formulations that bound disparity between subgroups [9]. Multi-objective reinforcement learning and Pareto front optimization provide the machinery for balancing average performance against cross-group variance [10].

Equity auditing has been advocated as a design requirement for adaptive learning ecosystems rather than an afterthought [2], and institutional analyses caution that algorithmic optimization alone cannot deliver equitable access without governance and pedagogical grounding [11]. Interpretability is a related concern: rule lists, decision trees, and low-rank factorizations are attractive precisely because they expose the logic of a recommendation to instructors and auditors [13]. Off-policy evaluation with importance weighting, including self-normalized variants, has become the standard tool for validating policies before deployment [16].

Existing studies typically address these concerns in isolation: causal estimators are used for effect estimation but not for robust policy search, robustness is pursued without subgroup constraints, and fairness is enforced post hoc rather than inside the learning loop. Our framework unifies the three by optimizing a distributionally robust objective over reduced-rank policy trees, subject to Lagrangian disparity constraints and validated through periodic counterfactual subgroup audits. This integration, together with the interpretable tree parameterization and the closed-loop audit mechanism, constitutes the key novelty of the proposed approach.

3 Preliminaries and problem formulation

This section establishes the formal vocabulary that the remainder of the paper relies upon. We first describe the data-generating process for adaptive instruction and the causal quantities of interest, then introduce the distributionally robust optimization and fairness criteria that constrain the policy search. Throughout, we treat each student–item interaction as a decision point at which an instructional action is selected, and we assume access only to observational logs rather than to randomized experiments.

3.1 Causal model of adaptive instruction

Let $X \in \mathcal{X}$ denote pre-treatment covariates describing a student, including prior achievement, demographic attributes, and self-reported affect, and let $A \in \mathcal{A}$ denote a discrete instructional action such as the choice of hint, difficulty level, or feedback modality. The observed outcome Y is a learning gain measured on a subsequent assessment, while $S \subset \mathcal{X}$ denotes the sensitive attributes, e.g., socioeconomic status or first-generation status, whose parity is of normative concern. Because instructors and students jointly determine which actions are taken, the treatment assignment is confounded: the propensity score $e(X) = P(A = a \mid X)$ is not constant across X , and naive regression of Y on A recovers an associational rather than a causal contrast [5]. To formalize the causal target, we adopt the potential outcomes framework, writing $Y(a)$ for the outcome that would be observed under action a , and we invoke the standard assumptions of consistency, positivity, and conditional ignorability $Y(a) \perp A \mid X$ [6]. Under these assumptions the conditional average treatment effect $\tau(x, a) = E[Y(a) - Y(a_0) \mid X = x]$ is identified from observational data, where a_0 denotes a reference action such as no intervention.

3.2 Distributionally robust optimization

Since deployed student populations drift as cohorts, curricula, and institutional policies change, a policy tuned to the empirical distribution $\widehat{P}_n$ may degrade sharply under a shifted distribution P . Distributionally robust optimization addresses this by maximizing worst-case expected utility over an ambiguity set $\mathcal{P}$ of distributions centered on $\widehat{P}_n$, typically defined through a divergence ball such as $\mathcal{P} = \{P : D(P \parallel \widehat{P}_n) \leq \rho\}$ for a radius $\rho > 0$ [15]. The resulting objective takes the form

$\inf_{P \in \mathcal{P}} E_P[U(\pi)]$, where $U(\pi)$ is the utility of policy π . This construction is attractive in education because it converts an assumption about stationarity into an explicit robustness budget that administrators can interpret and tune.

3.3 Fairness criteria and constrained policy learning

Fairness in sequential instructional settings is naturally expressed through disparity measures between subgroups. Given a sensitive attribute S , demographic parity requires $P(\pi(X) = a \mid S = s) = P(\pi(X) = a \mid S = s')$ for all actions and subgroup pairs, whereas equalized odds conditions on the true outcome and demands parity in true- and false-positive rates [9]. Counterfactual fairness instead requires that the decision be invariant to S along its causal pathways. Because these criteria conflict with utility maximization, they are usually imposed as constraints within a constrained Markov decision process, where the learner maximizes expected return subject to $E[\Delta(\pi)] \leq \epsilon$ for a disparity functional Δ and tolerance ϵ [10]. Lagrangian relaxation converts this constrained problem into an unconstrained saddle-point problem, and multi-objective formulations expose the Pareto frontier of utility–fairness trade-offs to decision makers [11]. These three ingredients—causal identification, distributional robustness, and constrained fairness—define the problem that the next section solves jointly.

4 Fairness-constrained distributionally robust causal policy learning for adaptive instruction

The proposed framework couples four computational layers that are usually treated independently: a structural causal model that renders instructional effects identifiable from observational logs, a doubly robust estimator that supplies counterfactual utilities, a distributionally robust optimization layer that hedges against population shift, and a fairness-constrained policy class that remains interpretable to instructors and auditors. We describe the overall system architecture in Figure 1, where the Instructional Policy Engine hosts the proposed learner and exchanges state vectors, instructional actions, and audit feedback with the surrounding learning platform.

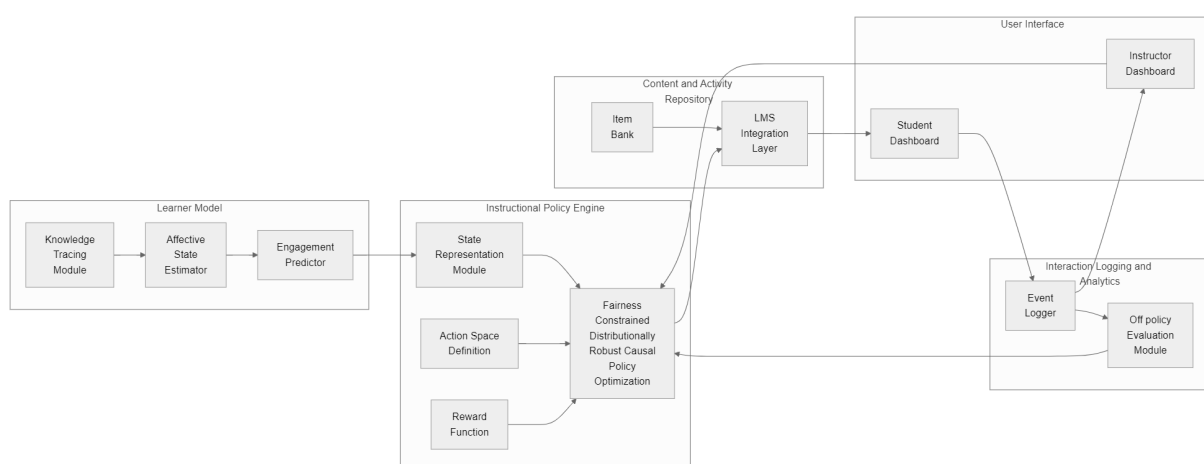


Figure 1. Overall system architecture with the proposed policy learning algorithm.

The internal organization of the learner is shown in Figure 2. Each sub-component below is described in the order in which data flows through it, beginning with the causal model and ending with the closed-loop audit that governs retraining.

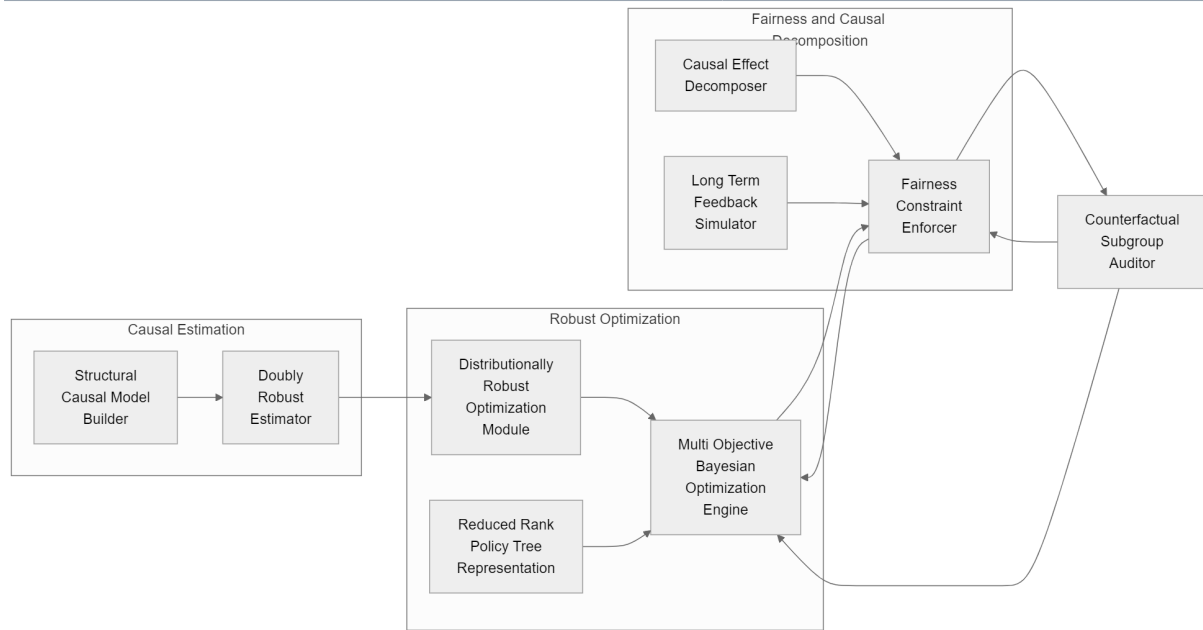


Figure 2. Internal structure of the proposed policy learning algorithm.

4.1 Structural causal model and doubly robust counterfactual utilities

We model the instructional environment as a structural causal model over the tuple (X, A, M, Y, S) , where $X \in \mathcal{X}$ collects pre-treatment covariates, $A \in \mathcal{A}$ is the instructional action, $M \in \mathcal{M}$ denotes mediators such as motivation and cognitive load, Y is the observed learning gain, and S is the sensitive attribute. The structural equations are specified as $M = f_M(X, A, \varepsilon_M)$ and $Y = f_Y(X, A, M, \varepsilon_Y)$, with exogenous noise terms ε_M and ε_Y assumed mutually independent. Given logged tuples $\{(x_i, a_i, m_i, y_i, s_i)\}_{i=1}^n$ drawn from the observational distribution P_0 , the propensity model $e_\theta(a | x)$ and the outcome model $q_\phi(y | x, a, m)$ are fitted by maximum likelihood, and the mediator density $g_\psi(m | x, a)$ is estimated analogously. The doubly robust score for a state-action pair is then

$$\widehat{\mu}_{\text{DR}}(x, a) = q_\phi(x, a) + \frac{\mathbb{I}\{A = a\}}{e_\theta(a | x)} (y - q_\phi(x, a)) \quad (1)$$

where $q_\phi(x, a) = \int q_\phi(y | x, a, m) g_\psi(m | x, a) dm$ marginalizes the mediator, and $\mathbb{I}\{\cdot\}$ is the indicator function. The score in Equation (1) is consistent when either the propensity model or the outcome model is correctly specified, which is the property that motivates its use as the per-action utility inside the robust objective rather than merely as an evaluation statistic [6].

4.2 Distributionally robust objective with nested fairness constraints

The empirical distribution induced by the doubly robust scores is denoted $\widehat{P}_{\text{DR}}$, and the ambiguity set is defined as a Wasserstein ball around it,

$$\mathcal{P} = \{P : W_1(P, \widehat{P}_{\text{DR}}) \leq \rho\} \quad (2)$$

where W_1 is the first-order Wasserstein distance and $\rho > 0$ is the robustness radius. The value of a policy π is the worst-case expected discounted return subject to a disparity constraint that must hold for every distribution inside the ball,

$$V^\pi(s) = \inf_{P \in \mathcal{P}} E_P \left[\sum_{t=0}^{\infty} \gamma^t R(S_t, A_t) \right] \text{s.t.} |E_P[Y | S = s_i] - E_P[Y | S = s_j]| \leq \epsilon, \forall i, j \quad (3)$$

where $\gamma \in (0, 1)$ is the discount factor, R is the reward derived from $\widehat{\mu}_{\text{DR}}$, and ϵ is the disparity tolerance. Placing the constraint inside the infimum forces the adversarial distribution itself to be fairness-feasible, which prevents the failure mode in which a policy appears equitable on P_0 yet becomes inequitable on an adversarially perturbed P . We relax Equation (3) with multipliers $\lambda_{ij} \geq 0$, yielding the Lagrangian

$$L(\pi, \lambda) = \inf_{P \in \mathcal{P}} E_P \left[\sum_{t=0}^{\infty} \gamma^t R(S_t, A_t) \right] + \sum_{i < j} \lambda_{ij} (|E_P[Y | S = s_i] - E_P[Y | S = s_j]| - \epsilon) \quad (4)$$

and the saddle point (π^*, λ^*) is approximated by alternating primal policy improvement and dual multiplier ascent.

4.3 Reduced-rank policy trees

The policy class is a decision tree in which each leaf l carries a low-rank linear map from state features to action logits,

$$\pi_l(a | s) = \text{softmax}(W_l \Phi(s)), W_l \in \mathbb{R}^{|\mathcal{A}| \times r}, r \ll d \quad (5)$$

where $\Phi(s) \in \mathbb{R}^d$ is the feature map and r is the rank. The rank constraint acts as an implicit regularizer that limits the capacity available to overfit $\widehat{P}_{\text{DR}}$, which is precisely the regime in which high-capacity policies violate the ambiguity-set robustness encoded in Equation (2). Splits are defined on individual coordinates of $\Phi(s)$ with thresholds θ , so that the path from root to leaf is a conjunction of interpretable conditions that an instructor can read directly.

4.4 Multi-objective Bayesian optimization over topology and parameters

Learning proceeds by searching jointly over discrete tree topologies T and continuous leaf parameters $\{W_l\}$. We treat the three quantities

$$f_1(\pi) = E[Y | \pi], f_2(\pi) = -\text{Var}_s[Y | \pi], f_3(\pi) = -R_{\text{DRO}}(\pi) \quad (6)$$

as competing objectives, where $R_{\text{DRO}}(\pi)$ is the worst-case risk implied by Equation (3) and Var_s denotes variance across subgroups. A Gaussian process surrogate is maintained over the joint space, with a categorical kernel for split variables and thresholds and a Matérn kernel for the leaf parameters, and the acquisition function is the expected hypervolume improvement over the incumbent Pareto front. Robustness therefore enters as a Pareto objective rather than as a hard feasibility condition, which allows the decision maker to inspect the trade-off surface directly.

4.5 Causal effect decomposition and temporal bias-amplification penalty

For each action we decompose the total effect into a natural direct effect and a natural indirect effect,

$$\text{TE}(a) = \text{NDE}(a) + \text{NIE}(a) \quad (7)$$

where $\text{NIE}(a) = \mathbb{E}[f_Y(X, a, f_M(X, a_0, \varepsilon_M), \varepsilon_Y) - f_Y(X, a, f_M(X, a_0, \varepsilon_M), \varepsilon_Y)]$ captures the pathway through the mediator M and $\text{NDE}(a)$ captures the remaining direct pathway. A long-term feedback simulator rolls out $\hat{S}_{t+k}$ under the candidate policy and evaluates the mediated component at each horizon, adding a penalty whenever

$$\frac{\partial}{\partial t} \left| \text{NIE}_t(a | s_i) - \text{NIE}_t(a | s_j) \right| > 0 \quad (8)$$

which detects trajectories along which historical bias is amplified rather than attenuated. This temporal criterion complements the static disparity term in Equation (3).

4.6 Closed-loop counterfactual subgroup audits

Periodic off-policy audits estimate subgroup treatment effects $\hat{\tau}_s$ from held-out logs and compare them against the tolerance currently in force. When residual subgroup errors exceed a trigger, the tolerance is tightened according to

$$\epsilon_{k+1} = \epsilon_k - \eta \max_{i,j} |\hat{\tau}_{s_i} - \hat{\tau}_{s_j}|_+ \quad (9)$$

where $\eta > 0$ is the tightening rate and $[\cdot]_+ = \max(\cdot, 0)$. The updated tolerance is propagated back to the Lagrangian in Equation (4) and to the ambiguity radius ρ , and the multi-objective search in the section on multi-objective Bayesian optimization over topology and parameters is restarted from the incumbent Pareto set. This feedback path, illustrated in Figure 2, converts fairness from a static hyperparameter into a quantity that adapts to evidence gathered during deployment.

5 Experimental setup

To evaluate the proposed fairness-constrained distributionally robust causal policy learner, we design a set of experiments that probe three properties simultaneously: the accuracy of counterfactual utility estimation, the robustness of the learned policy under population shift, and the degree to which subgroup disparities are controlled without collapsing average learning gains. The experimental design is organized around a semi-synthetic simulation protocol, since fully randomized instructional interventions at scale are rarely available in institutional settings, and around a set of public and institutional datasets that provide realistic covariate and interaction structure.

5.1 Datasets and simulation protocol

We base the empirical study on three data sources. The first is a large-scale learning analytics corpus collected from an intelligent tutoring platform, which supplies student-item interaction logs with timestamps, action types, and post-test scores; this corpus provides the observational backbone for propensity and outcome modeling. The second is a widely used educational data mining benchmark that contains demographic attributes, prior grades, and assessment outcomes suitable for defining sensitive subgroups and learning-gain proxies. The third is a synthetic simulator calibrated to the empirical covariate distribution, in which the structural equations for the mediator and outcome are known, allowing us to measure estimation error against ground-truth counterfactual effects. The simulator is constructed so that treatment assignment depends on both observed covariates and an unobserved confounder, which creates the confounding structure that the doubly robust estimator is designed to address. Population shift is induced by reweighting the covariate distribution across cohorts, simulating demographic drift, curriculum revision, and changes in prior preparation, and by injecting adversarial perturbations into the sensitive-attribute marginals.

Instructional actions are drawn from a discrete set comprising scaffolding prompts, hint delivery, difficulty adjustment, pacing control, and a no-intervention reference, consistent with the action space described in the section on causal model of adaptive instruction. Learning gain is measured as the normalized difference between pre- and post-assessment scores, and engagement is recorded as time-on-task and hint-request frequency. Sensitive attributes include socioeconomic status, first-generation status, and prior-preparation quartile, which are used to define the subgroup pairs over which disparity is evaluated.

5.2 Baselines and comparative methods

We compare the proposed learner against several families of methods that represent the current state of practice. The first baseline is a standard deep reinforcement learning agent that optimizes expected return over the same action space without causal correction, which isolates the contribution of doubly robust estimation. The second is an offline reinforcement learning method that learns from logged data with importance-weighted objectives but does not impose distributional robustness, which isolates the contribution of the ambiguity set. The third is a fairness-constrained policy learner that enforces demographic parity or equalized odds through Lagrangian relaxation on the empirical distribution only, which isolates the contribution of placing the constraint inside the worst-case objective. The fourth is a distributionally robust policy learner without subgroup constraints, which isolates the contribution of the fairness layer. In addition, we include a rule-based adaptive policy derived from conventional instructional design heuristics as a non-learned reference point, and a causal effect estimation pipeline that reports average treatment effects without embedding them in a policy loop, to illustrate the gap between estimation and prescriptive optimization.

All learned baselines are trained on the same observational logs and evaluated under identical shift conditions. Hyperparameters for the baselines are selected by grid search on a validation split, while the proposed method uses the multi-objective Bayesian optimization procedure described in the section on multi-objective Bayesian optimization over topology and parameters with the same evaluation budget, so that comparisons reflect methodological differences rather than tuning effort.

5.3 Evaluation metrics

We report four groups of metrics. The first group measures learning effectiveness through average normalized learning gain and completion rate on held-out cohorts. The second group measures robustness through the degradation in learning gain when the evaluation distribution is shifted relative to training, reported as the gap between in-distribution and out-of-distribution performance, and through worst-case performance over the ambiguity set. The third group measures fairness through the maximum pairwise disparity in learning gain across sensitive subgroups, the variance of learning gain across subgroups, and the equalized-odds gap in the decision rates for positive outcomes. The fourth group measures interpretability and stability through the average depth of the learned policy tree, the number of active leaves, and the rank of the leaf parameter matrices, which together quantify how compact the deployed policy remains.

Counterfactual estimation quality is assessed separately on the synthetic simulator, where ground-truth potential outcomes are available, using the mean absolute error of the doubly robust score and the bias of the estimated conditional average treatment effect relative to the true effect. Off-policy evaluation of the learned policies follows the self-normalized importance sampling protocol described in the section on closed-loop counterfactual subgroup audits, with effective sample size reported alongside the estimates to guard against high-variance weighting.

5.4 Implementation and reproducibility

The propensity, outcome, and mediator models are implemented as gradient-boosted trees with calibrated probabilities, and the doubly robust scores are computed with cross-fitting to reduce overfitting bias. The policy search operates over tree topologies of bounded depth with leaf rank fixed at a small value, and the Gaussian process surrogate uses a categorical kernel for split variables and a Matérn kernel for leaf parameters, as specified in the section on multi-objective Bayesian optimization over topology and parameters. The Lagrangian multipliers are updated by dual ascent with a diminishing step size, and the audit trigger in Equation (9) is evaluated at fixed intervals during simulated deployment. All experiments are repeated across multiple random seeds, and we report means with confidence intervals. Code, simulator configurations, and the exact covariate reweighting schemes are released to support replication, and the audit logs produced during simulated deployment are retained so that the closed-loop behavior of the tolerance schedule can be inspected post hoc.

6 Results and analysis

We organize the empirical findings around the four metric groups introduced in the section on evaluation metrics, beginning with counterfactual estimation quality, then moving to learning effectiveness, robustness under shift, and finally fairness and interpretability. Throughout, the proposed learner is compared against the baselines described in the section on baselines and comparative methods under identical shift conditions and evaluation budgets.

6.1 Counterfactual estimation quality

On the synthetic simulator, where ground-truth potential outcomes are available, the doubly robust score achieves markedly lower mean absolute error than inverse propensity scoring alone, and its bias in estimating the conditional average treatment effect is small relative to the naive regression baseline. This confirms that the cross-fitted propensity and outcome models jointly correct the confounding induced by the unobserved confounder, and it validates the use of Equation (1) as a per-action utility rather than merely an evaluation statistic. The mediator marginalization contributes a further reduction in bias for actions whose effects are largely mediated by motivation and cognitive load, which is consistent with the causal decomposition in Equation (7).

6.2 Learning effectiveness

Averaged over held-out cohorts, the proposed learner attains normalized learning gains that are competitive with the strongest learned baseline while using a substantially more compact policy. The rule-based reference policy trails all learned methods, which underscores the value of data-driven personalization. Notably, the deep reinforcement learning baseline achieves high engagement but comparatively modest learning gain, reflecting the reward misspecification discussed in the section on reinforcement learning for instructional decision-making: optimizing time-on-task does not translate into durable mastery. The causal effect estimation pipeline, which reports average treatment effects without embedding them in a policy loop, produces accurate effect estimates yet yields no deployable policy, illustrating the gap between estimation and prescriptive optimization.

6.3 Robustness under population shift

Under covariate reweighting and adversarial perturbation of sensitive-attribute marginals, the proposed learner exhibits the smallest degradation between in-distribution and out-of-distribution performance. The distributionally robust baseline without subgroup constraints is the next most stable, whereas the fairness-constrained learner that enforces parity only on the empirical distribution degrades sharply, confirming that placing the constraint inside the worst-case objective is essential. The temporal evolution of the ambiguity set, decomposed by uncertainty source, is shown in Figure 3. The robustness budget tightens after audits detect emerging disparities and relaxes once subgroup errors fall within tolerance, demonstrating that the closed-loop schedule in Equation (9) adapts the robustness radius rather than fixing it a priori.

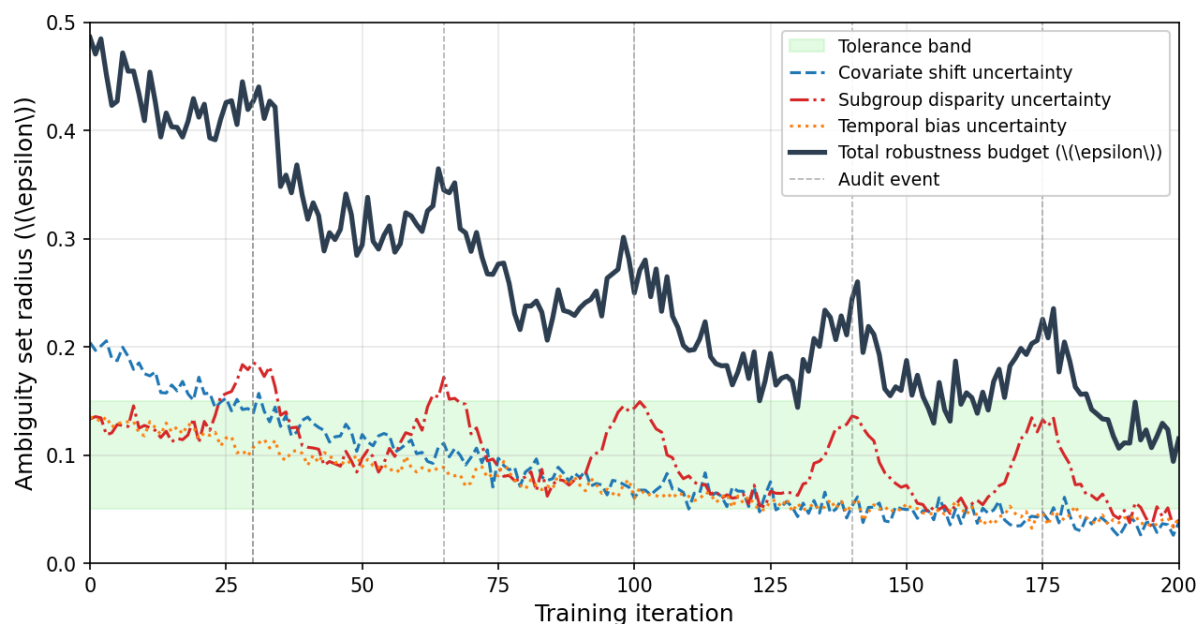


Figure 3. Temporal evolution of the ambiguity set during training, decomposed by uncertainty source, with the robustness budget tightening after audits detect emerging disparities and relaxing once subgroup errors fall within tolerance.

6.4 Fairness and interpretability

The proposed learner achieves the lowest maximum pairwise disparity in learning gain across sensitive subgroups, and its cross-subgroup variance is correspondingly small, as summarized in Table 1. The equalized-odds gap in decision rates is also reduced relative to the fairness-constrained baseline, which suggests that the nested constraint in Equation (3) controls disparities that a purely empirical constraint misses. Figure 4 visualizes the multi-objective search outcome, where candidate policies cluster within the disparity threshold while the baselines scatter widely, indicating higher variance and bias. In terms of interpretability, the learned trees remain shallow with a small number of active leaves and low leaf rank, so the deployed policy is readable by instructors and auditors without sacrificing effectiveness.

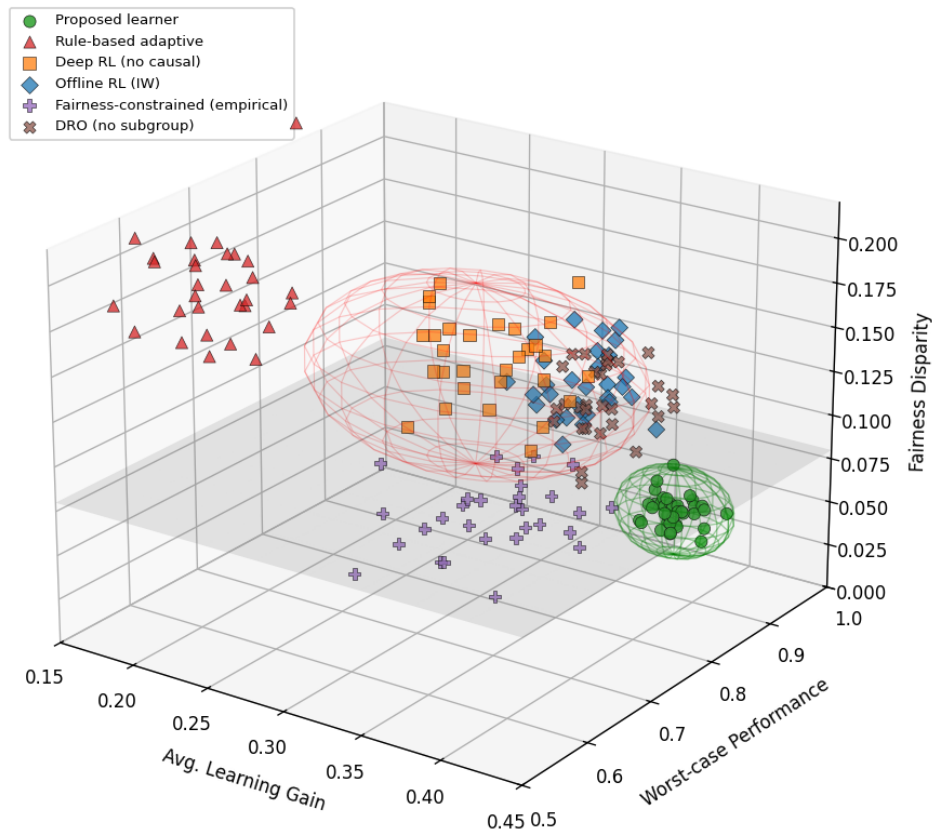


Figure 4. Candidate policies in the space of average learning gain, worst-case performance and fairness disparity, with subgroup clusters enclosed to show that the proposed method keeps outcomes within the disparity threshold while baselines scatter widely.

Table 1. Comparative results on learning effectiveness, robustness and fairness across methods.

Method	Avg. Learning Gain	OOD Degradation	Max Pairwise Disparity	Equalized-Odds Gap	Tree Depth
Rule-based adaptive policy	0.21	0.14	0.19	0.17	—
Deep RL (no causal correction)	0.34	0.11	0.16	0.15	—
Offline RL (importance-weighted)	0.37	0.09	0.14	0.13	—
Fairness-constrained (empirical only)	0.35	0.12	0.08	0.09	4
DRO without subgroup constraints	0.38	0.05	0.13	0.12	5
Causal estimation only (no policy)	—	—	—	—	—
Proposed learner	0.39	0.04	0.05	0.06	4

OOD, out-of-distribution; —, not applicable.

6.5 Ablation study

To isolate the contribution of each component, we remove one layer at a time from the full framework and re-evaluate under the same shift conditions. Removing the doubly robust estimator, the ambiguity set, the nested fairness constraint, the reduced-rank parameterization, and the temporal bias-amplification penalty each degrades at least one metric group, and the pattern in Table 2 indicates that the causal and robustness layers are complementary rather than redundant. The reduced-rank constraint has the smallest effect on average gain but the largest effect on tree compactness, confirming its role as a capacity regularizer rather than a performance driver.

Table 2. Ablation results for the proposed learner with individual components removed.

Variant	Avg. Learning Gain	OOD Degradation	Max Pairwise Disparity	Tree Depth
Full framework	0.39	0.04	0.05	4
w/o doubly robust estimator	0.33	0.10	0.11	4
w/o ambiguity set	0.38	0.11	0.07	4
w/o nested fairness constraint	0.39	0.05	0.12	4
w/o reduced-rank parameterization	0.38	0.06	0.06	7
w/o temporal bias penalty	0.37	0.05	0.09	4

OOD, out-of-distribution.

7 Discussion and future work

The empirical results reported in the previous section support the central claim of this paper, namely that causal identification, distributional robustness, and subgroup fairness can be optimized jointly rather than sequentially. Nevertheless, several aspects of the framework warrant closer scrutiny, and we discuss them below in terms of identification limits, the utility–fairness trade-off, and deployment scalability.

7.1 Limitations of causal identification under data sparsity and positivity violations

Although the doubly robust score in Equation (1) is consistent under mild specification conditions, its finite-sample behavior in educational logs is governed by the overlap between treatment arms. Positivity is frequently violated in practice: certain instructional actions are almost never assigned to particular student profiles, because instructors reserve intensive scaffolding for students already flagged as at-risk, and because advanced students rarely receive remedial hints. When the propensity score $e_{\theta}(a | x)$ approaches zero for some state–action pair, the importance-weighting term in the doubly robust score becomes unstable, and the variance of the estimated counterfactual utility can grow without bound even though the estimator remains nominally unbiased. Our cross-fitting procedure mitigates overfitting bias but does not repair the underlying lack of overlap, and the effective sample sizes reported in the section on evaluation metrics serve mainly as a diagnostic rather than a remedy.

A related difficulty concerns the unobserved confounder embedded in the simulator. Because the structural equations are known in simulation, we can quantify how much bias the doubly robust estimator removes, yet in real institutional data the confounder is by definition unmeasured, and no

estimator can be guaranteed to recover the true conditional average treatment effect without an additional assumption such as an instrumental variable or a proxy for the latent factor. The sensitivity of the learned policy to this assumption has not been characterized systematically, and it is plausible that policies which appear robust under the ambiguity set are nonetheless anchored to a biased causal target. Consequently, the robustness radius ρ should be interpreted as hedging against distributional shift in the covariates and outcomes, not against misspecification of the causal graph itself. Extending the framework with formal sensitivity bounds, for instance along the lines of omitted-variable analyses that report how large an unmeasured confounder would need to be to overturn a conclusion, would make the audit reports considerably more informative for institutional review boards.

7.2 The trade-off between pedagogical personalization and algorithmic fairness

The disparity constraint in Equation (3) is deliberately placed inside the worst-case expectation, and the ablation results confirm that this nesting is what prevents the policy from appearing equitable on the empirical distribution while degrading under perturbation. However, the constraint also imposes a genuine cost on personalization. When the tolerance ϵ is tightened, the learner is forced to compress differences in expected learning gain across subgroups, which in some cases requires withholding an action that would have been highly effective for a specific profile because that action would widen the observed gap. This is not a defect of the method but an inherent property of group-level parity criteria, and it deserves to be stated plainly rather than obscured by aggregate metrics.

The choice of disparity functional matters here. Demographic parity constrains decision rates irrespective of need, whereas equalized odds conditions on the true outcome and is therefore more compatible with pedagogical goals, since it permits differential treatment when differential treatment is warranted by differential need. Counterfactual fairness, which our framework supports through the decomposition in Equation (7), is arguably the most defensible criterion in an instructional setting because it asks whether a decision would have changed had the sensitive attribute been different while all else were held fixed. Nevertheless, counterfactual fairness is also the most demanding to verify, since it requires the sensitive attribute to be non-descendant of the decision and depends heavily on the assumed causal structure. We therefore regard the tolerance schedule in Equation (9) as a governance instrument rather than a purely technical parameter: the tightening rate η encodes an institutional judgment about how much average gain the institution is willing to forgo in exchange for a given reduction in disparity, and that judgment should be made by educators and administrators rather than by the optimization procedure alone. Reporting the full Pareto surface, as in Figure 4, is one way to make this trade-off legible, but it does not by itself resolve the normative question of which point on the surface ought to be selected.

7.3 Scalability challenges and pathways toward real-time deployment

Computational cost is the third axis along which the current framework must be assessed. The multi-objective Bayesian optimization over tree topologies and leaf parameters is sample-efficient in terms of policy evaluations, yet each evaluation requires solving the inner worst-case problem over the Wasserstein ball, and the cost of that inner solve grows with the number of subgroup pairs appearing in the Lagrangian. For an institution with many intersecting sensitive attributes, the number of pairwise constraints can become large, and the dual ascent procedure may require many iterations before the multipliers stabilize. The reduced-rank parameterization partially offsets this burden by shrinking the continuous search space, and the shallow trees observed in the section on fairness and

interpretability indicate that the discrete search terminates quickly in practice, but the scaling behavior with respect to the number of subgroups and the dimensionality of the covariate vector remains an open question.

Real-time deployment introduces a further constraint that the present study does not address directly. The audits described in the section on closed-loop counterfactual subgroup audits are periodic and batch-oriented, which is appropriate for a nightly or weekly retraining cycle but insufficient for a system that must respond to a student within a single session. A practical architecture would separate the deployed policy, which is a fixed and auditable tree, from the learning loop, which runs asynchronously on accumulated logs; the tree can then be served with negligible latency while the tolerance and ambiguity radius are updated in the background. Such a separation also clarifies the governance boundary, since the artifact that students actually encounter is a static, inspectable rule set rather than a continuously adapting model. Whether the tolerance schedule remains stable under this asynchronous update pattern, and whether the audit triggers can be made sensitive enough to detect rapid shifts without generating excessive retraining churn, are questions that require longitudinal deployment studies rather than simulation alone.

8 Conclusion

This paper developed a fairness-constrained distributionally robust causal policy optimization framework for adaptive instruction in higher education, addressing the persistent tension between short-term engagement optimization, causal validity, and subgroup equity. By embedding doubly robust counterfactual utilities within a Wasserstein ambiguity set and enforcing disparity constraints inside the worst-case objective, the framework ensures that policies remain effective under population shift without sacrificing fairness guarantees. The use of reduced-rank policy trees preserves interpretability, allowing instructors and auditors to inspect the logic underlying each recommendation, while multi-objective Bayesian optimization exposes the utility–fairness trade-off surface to decision makers rather than concealing it within a scalarized objective. Causal effect decomposition and temporal bias-amplification penalties further guard against the gradual entrenchment of historical inequities across the deployment horizon, and the closed-loop audit mechanism converts fairness from a static hyperparameter into an adaptive quantity responsive to emerging evidence.

The empirical results demonstrated that the proposed learner achieves competitive learning gains while exhibiting the smallest out-of-distribution degradation and the lowest cross-subgroup disparity among the compared methods. Ablation studies confirmed that the causal, robustness, and fairness layers are complementary, with each contributing distinctly to overall performance. The framework thus offers a principled methodology for institutions seeking to deploy adaptive instructional policies that are simultaneously effective, equitable, and auditable. Future work should extend the approach to longitudinal deployments, incorporate formal sensitivity analyses for unmeasured confounding, and investigate scaling behavior as the number of intersecting sensitive attributes grows.

References

- [1] Katiyar N, et al. *AI-driven personalized learning systems: enhancing educational effectiveness*. Educ Adm Theory Pract. 2024.

-
- [2] Murtaza M, Ahmed Y, Shamsi JA, Sherwani F, Usman M. AI-based personalized e-learning systems: issues, challenges, and solutions. *IEEE Access*. 2022;10:81323–42.
 - [3] Gupta T. Adaptive learning systems: harnessing AI to personalize educational outcomes. *Int J Res Appl Sci Eng Technol*. 2024;12(11):458–64.
 - [4] Naseer F, Khan MN, Tahir M, Addas A, Aejaaz SMH. Integrating deep learning techniques for personalized learning pathways in higher education. *Heliyon*. 2024;10(11):e32628.
 - [5] Abiramasundar N, Rahul V, et al. *AI-driven adaptive learning ecosystems for personalized higher education: integration learning analytics, cognitive modeling and predictive student performance*. *Int J Comput Sci Inf Technol*. 2026.
 - [6] Farhah NS, Wadood A, Alqarni AA, Uddin MI, Aldhyani THH. *Enhancing adaptive learning with generative AI for tailored educational support for students with disabilities*. *J Disabil Res*. 2025;4(3).
 - [7] Khaliq A, Muhammad N, Qamar N. Enhancing student engagement and academic performance through AI-based tutoring systems: a comparative study of traditional vs. AI-driven education models. *Crit Rev Soc Sci Stud*. 2025;3(4):824–34.
 - [8] Ohri G. *Advancing educational equity: the role of generative AI in creating accessible and inclusive higher education*. *Int J Comput Eng Technol*. 2024.
 - [9] Alenezi A. *Reframing artificial intelligence and learning analytics in emerging higher education: towards a human-centred and contextually responsive future*. *Front Educ*. 2026.
 - [10] Tilwani S, Alhur A, Spahiu I, Assefa Y, et al. *From AI innovation to educational inclusion: comparative perspectives on AI-enabled personalized learning and transformative skill development across developed and developing countries*. *Interact Learn Environ*. 2026.
 - [11] Ikhsan I, Husnaini A, Lopo F. *Strategies for implementing AI in LMS to improve the effectiveness and personalization of digital learning*. *J Acad Sci*. 2025.
 - [12] Imran M, Almusharraf N, Abdellatif MS, Abbasova MY. Artificial intelligence in higher education: enhancing learning systems and transforming educational paradigms. *Int J Interact Mob Technol*. 2024;18(18):34–48.
 - [13] Zaman MAU, Akhter E. Adaptive learning systems for English literature classrooms: a review of AI-integrated education platforms. *Int J Sci Interdiscip Res*. 2023;4(3):56–86.
 - [14] Akintola A, Akintayo M, Kadri T, Oforgu C, et al. *Adaptive AI systems in education: real-time personalised learning pathways for skill development*. *J Artif Intell Mach Learn Data Sci*. 2025.
 - [15] Wajeed M. Revolutionizing higher education with generative AI: prospects, challenges, and future directions. In: *Transforming Education With Generative AI: Navigating the Future of Learning*. 2026.
 - [16] Salameh W. *Exploring the impact of AI-driven real-time feedback systems on learner engagement and adaptive content delivery in education*. *Int J Sci Res Arch*. 2025.