Understanding the Impact of Reinforcement Learning Personalization on Subgroups of Students in Math Tutoring

Allen Nie[0000−0001−6483−2111], Ann-Katrin Reuel*[0000−0002−7913−9296], and Emma Brunskill[0000−0002−3971−7127]

Stanford University, Stanford, CA, USA
{anie, anka, ebrun}@cs.stanford.edu

Abstract. Reinforcement learning has the promise to help reduce the cost of creating effective educational software through automatically adapting the experience to each individual. Most reinforcement learning algorithms aim to learn an automated pedagogical strategy that optimizes performance on average across the population and outputs a decision policy that may rely on complex representations, like deep neural networks, that are largely opaque. Yet, in most educational contexts, we would like a deeper understanding of educational interventions, such as if the machine-learned pedagogical strategy differs in its benefits to different students, how it differentiates instruction across individuals or situations, and if the personalized strategy learned has benefits over alternative personalizations or automated strategies. Here we explore such analyses for a reinforcement learning decision policy for educational software teaching students about the concept of volume. While some related work covers part of these analyses, we suggest that conducting all three such analyses can help enhance our understanding of the impact of a reinforcement learning decision policy in education and help inform stakeholders' decisions around the use of a particular learned decision policy.

Keywords: Reinforcement Learning · Conditional Average Treatment Effect · Offline Policy Evaluation.

1 Introduction

Reinforcement learning algorithms can learn adaptive decision policies that map from a context to an intervention in order to optimize an expected outcome. Such algorithms hold great promise for optimizing educational software to best support the learning experience of individual students. However, most reinforcement learning algorithms optimize for outcomes on average, often employ complex, hard-to-interpret models like deep neural networks, and frequently lack formal guarantees of convergence to a globally optimal solution. Therefore, an important area of inquiry is to better understand the impact of a particular reinforcement learned pedagogical policy on different students.

* Equal contribution

While some studies have examined the outcomes and implications of reinforcement learning policies in education, the amount of research conducted in this area remains limited. Prior work on college students using a logic tutoring system suggested that some students may be relatively insensitive to different automated pedagogical policies, but some other students benefited significantly from the RL policy personalization [3]. Concurrent to our paper, Abdelshiheed et al.[1] found that an adaptive deep RL policy yielded substantial gains for students who initially were unlikely to try new metacognitive strategies for a logic tutor, but seemed to have little effect for students who already employed such strategies. In the context of a machine learning method for video recommendations for algebra learning, Leite et al.[4] used causal decision trees to identify if subgroups of students significantly varied in their treatment effects. To our knowledge, a more holistic set of analyses to understand the personalization done by a learned reinforcement learning policy, and its impact on student subgroups, has not been proposed.

To gain insight into the impact and effects of a personalization policy (highlighted in Figure 1), we suggest three useful analyses:

(**R1**) **Subgroup Identification**: Identifying subgroups of students according to their differential treatment effect under the RL adaptive policy vs a standard control.
(**R2**) **Analysis of Personalization**: Employing insights from the model interpretation to analyze the difference in RL automated tutoring strategies for different subgroups of students.
(**R3**) **Impact of the Specific RL Personalization**: Constructing alternative policies and using offline policy evaluation methods to estimate the impact of the specific RL policy personalization on subgroups of students.

We present a case study that uses these analyses to advance our understanding. To do so, we use data from a study in which RL was used to personalize an educational tool to help elementary school students learn about the concept of volume.

2 Study Description

In the study on student learning of volume concepts, some students used a narrative-based, artificial intelligent educational software tool. While thse students work through a series of volume-related problems, the software can provide adaptive support in response to student questions. The reinforcement learning policy selects among four pedagogical strategies: providing direct hints, encouragement, Socratic questioning or prompting the student to reframe, or a simple acknowledgment.

In this study, 270 participants in grades 3–5 were recruited from across the United States. Children were randomly assigned to one of two conditions: a standard interface that provided students with a volume practice task without hints or a narrative and a condition with a storyline and a reinforcement-learning

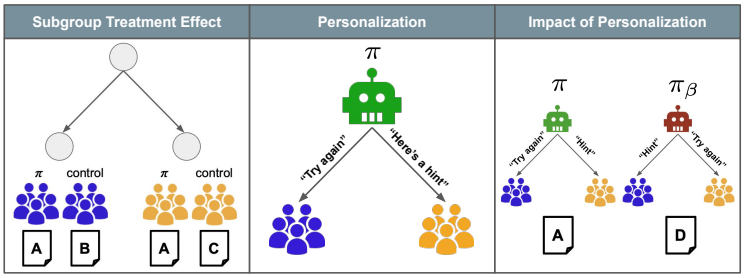

Fig. 1. Three useful Analyses of RL Personalization: (left) Understanding differential treatment effects of the learned RL policy, vs a standard benchmark approach. (middle) Features that describe differences in the learned decision policy. Blue and yellow icons represent subgroups of students with different covariates (for example, low and high pre-test scores). (right) Exploring if different personalization policies yield significantly more or less effective teaching.

augmented agent. 67 children used the control system, and 203 children used the system with RL agent-mediated guidance. Gender and grade were balanced between the two conditions.

3 Analysis and Results

3.1 Student Subgroup Identification

We first examine if there are significant differences among subgroups in terms of the impact of reinforcement learning vs. a control condition. Similar to Leite et al. [4], we employ a subgroup treatment effect analysis using a two-stage cluster-robust causal forest [10] to estimate the individual treatment effect and identify subgroups. A causal forest is an ensemble of causal trees that have been grown on a random subsample of the data during training to predict individual treatment effect[2]. Causal forests are more robust to nuances of the data-splitting procedure, but are a bit more limited for identifying consistent subgroups, which may vary substantially across trees in the forest. Leite et al. [4] addressed this by using a best linear analysis. However we are particularly interested in the non-linear but interpretable benefits of decision trees. Therefore we proposed an alternate heuristic method to identify the student subgroups and estimate the subgroup-level treatment effect, which preserves a tree representation flexibility. For our data, the student features we include are "gender" (0/1), "math anxiety" (9-45), and "pre-test score"(0-8).

We first subsample 43% of our data to build the causal forest. Then we use the R-loss[5], which calculates the expected difference between the estimated and the true treatment effect, to select the best tree out of the ensemble. We follow this tree's decision rules and allocate students in each subgroup. Then we use the holdout 50% of our data to calculate the conditional average treatment effects

(CATE) for each subgroup. This constitutes an honest estimation and mimics the procedures in honest forests. See Wager et al.[10] for a formal discussion.

We built a causal forest with 500 trees and a minimum node size of 7. Due to the small size of the dataset, we further set the sample fraction that is used to grow an individual tree to 0.8. The 'grf' R package was used to fit the causal forest [9]. We use a difference-in-means estimator to calculate the CATE for each subgroup and construct the 95% confidence interval.

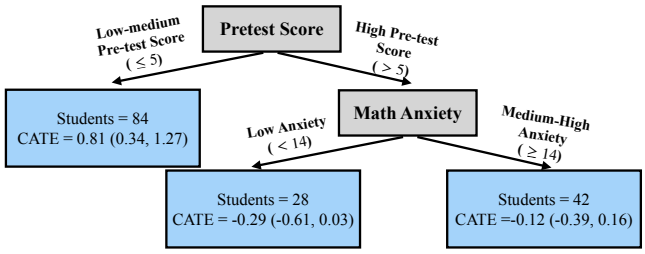

Fig. 2. The best tree selected from the causal forest. It shows three identified subgroups with the respective CATE for each group and the 95% confidence interval.

Three subgroups have been identified, with splits occurring based on the features 'math anxiety' and 'pre-test score.' We show them in Figure 2. Finally, the holdout students ($n = 154$) are divided into subgroups determined by the best causal tree before the respective CATE is calculated. Students with a low pre-test score had the highest average group average treatment effect in our tree: these students particularly benefited from the RL adaptive policy condition compared to a simple control condition. The impact on other students overlaps with zero, suggesting a slight negative or null result. This is consistent with past work [3] that founds benefits in a logic tutor with RL most benefited students who initially appeared to be struggling.

3.2 Analysis of Personalization

The learned RL policy uses a neural network to produce a probability distribution over the four pedagogical strategies given an input learning context. In previous work on this dataset, Ruan et al.[8] found that integrated gradients, a method in explainable machine learning, suggested math anxiety and student pre-test scores are the most influential student features to the policy's decision-making. Here, we compute the average probability of taking each action for the student subgroups, created by taking the top and bottom 25%-quantile of both features. We note that the two groups that exhibit the largest difference in terms of probability between all actions are the students with high pre-test scores and high math anxiety and students with low pre-test scores and low math anxiety. However, a chi-squared analysis did not show a difference between the groups.

Table 1. We show the average probability of policy taking each action in each student subgroup. Top/bottom means the top or bottom 25% quantile of the pre-test score. High/low means the top or bottom 25% quantile of the math anxiety.

RL Policy Action	Top × High	Bottom × Low	Top × Low	Bottom × High
Pr(Direct Hint)	57.6%	35.7%	50.7%	43.2%
Pr(Acknowledgment)	3.35%	18.3%	7.36%	8.14%
Pr(Encouragement)	20.6%	25.9%	27.9%	16.8%
Pr(Guided Prompt)	18.4%	20.1%	14.1%	31.8%

3.3 Impact of Personalization

In the previous section, we identified two subgroups of students that induce the most difference in the action probabilities of the policy. We choose the two subgroups to be the Top × High and Bottom × Low groups defined in Table 1. We now explore the following counterfactual: if the policy had not assigned the learned personalized interventions to these two groups of students, would we have seen a difference in the learning outcomes of these two groups?

To explore this, we built an alternative anti-policy π_β, in contrast to our original policy π. We construct π_β with the following procedure: First, we compute the maximum for both the pre-test and the math anxiety scores across all students in group A (Top x High). We then replace the pre-test and math anxiety scores for all students in group B (Bottom x Low) with these two maxima. All other features in group B (negative/positive sentiment in text, failed attempts, etc.) remain unchanged. Similarly, we take the original data in group B and compute the minimum across all students for the pre-test and the math anxiety scores in this group. Subsequently, we replace the corresponding feature values in group A with these two minima. Again, all other features in group A remain the same. This feature swap allows us to obtain a counterfactual policy without re-training by rerunning the original policy on the alternated dataset.

An alternative to adaptive, differentiated policis is a static, non-personalized decision policy. This might be particularly useful if sometimes student features are noisy or mismeasured. Therefore, we also estimated the performance of a non-personalized static policy $\bar{\pi}$, constructed by computing the average probability of each action over all states: $\bar{\pi}(a) = \mathbb{E}_s[\pi(a|s)]$. To estimate the performance of these alternative policies, we use a popular offline policy evaluation technique: weighted importance sampling [7], which allows us to use historical data collected by policy π to compute the expected student improvement (Y) of π_β. Let $w_i = \prod_{j=1}^{L} \frac{\pi_\beta(a|s)}{\pi(a|s)}$ for the i-th student: $\mathbb{E}_{\tau\sim\rho_{\pi_\beta}}[Y] = \mathbb{E}_{\tau\sim\rho_\pi}\left[\frac{w_i}{\sum_{j=0}^{n_1} w_j} Y\right]$ and $\mathrm{Var}_{\pi_\beta}(Y) = \mathbb{E}_{\tau\sim\rho_\pi}\left[\left(\frac{w_i}{\sum_{j=0}^{n_1} w_j}\right)^2 (Y - \mathbb{E}_{\pi_\beta}[Y])^2\right]$. Suppose there is a significant difference between $\mathbb{E}_\pi[Y]$ and $\mathbb{E}_{\pi_\beta}[Y]$. In that case, we can conclude that the policy's choice of personalization impacts these subgroups of students. We can derive the variance of the weighted importance sampling estimator, with details in Owen[6].

Table 2. Mean & variance of expected student performance improvement (WIS)

Sub-group	$\mathbb{E}_{\pi}[Y]$ (Original)	$\mathbb{E}_{\pi_\beta}[Y]$ (Anti)	$\mathbb{E}_{\bar{\pi}}[Y]$ (Static)
Top Pre-test × High Anxiety	0.42 (0.06)	0.26 (0.04)	0.33 (0.06)
Bottom Pre-test × Low Anxiety	3.55 (0.23)	1.71 (1.27)	3.21 (0.54)

We report our findings in Table 2. There is a significant benefit for both student subgroups from using the original policy (see $\mathbb{E}_{\pi}[Y]$) over our estimate of the alternate anti-optimized π_β (see $\mathbb{E}_{\pi_\beta}[Y]$). The potential change in performance vs using the static policy is small, indicating there exists a static policy that is robust to potential mismeasurement of student features. This provides stakeholders options on which policy to implement.

4 Conclusion

Here we proposed analyses to help understand the personalization done by an RL policy, and the impact outcomes. We apply our framework to data from a real-life study on math tutoring, showing RL personalizes in an impactful way.

5 Acknowledgements

This work was supported by a Stanford Hoffman-Yee & a NSF #2112926 grant.

References

1. Abdelshiheed, M., Hostetter, J.W., Barnes, T., Chi, M.: Leveraging deep reinforcement learning for metacognitive interventions across intelligent tutoring systems. In: Artificial Intelligence in Education: AIED 2023 (2023)
2. Athey, S., Tibshirani, J., Wager, S., et al.: Generalized random forests. The Annals of Statistics **47**(2)
3. Ausin, M.S.: Leveraging deep reinforcement learning for pedagogical policy induction in an intelligent tutoring system. In: Proceedings of EDM (2019)
4. Leite, W.L., Kuang, H., Shen, Z., Chakraborty, N., Michailidis, G., D'Mello, S., Xing, W.: Heterogeneity of treatment effects of a video recommendation system for algebra. In: Learning@Scale. pp. 12–23 (2022)
5. Nie, X., Wager, S.: Quasi-oracle estimation of heterogeneous treatment effects. Biometrika **108**(2), 299–319 (2021)
6. Owen, A.B.: Monte carlo theory, methods and examples (2013)
7. Precup, D.: Eligibility traces for off-policy policy evaluation. Computer Science Department Faculty Publication Series p. 80 (2000)
8. Ruan, S., Nie, A., Steenbergen, W., He, J., Zhang, J., Guo, M., Liu, Y., Nguyen, K.D., Wang, C.Y., Ying, R., et al.: Reinforcement learning tutor better supported lower performers in a math task. arXiv preprint arXiv:2304.04933 (2023)
9. Tibshirani, J., Athey, S., Wager, S.: grf: Generalized random forests. r package version 120 (2020)
10. Wager, S., Athey, S.: Estimation and inference of heterogeneous treatment effects using random forests. JASA **113**(523), 1228–1242 (2018)