Mediator-Based Reward Design in Online Contextual Bandit

Lutong Zou[*]
Peking University
Beijing, China 100871
bbf@stu.pku.edu.cn

Ziping Xu
Harvard University
Cambridge, MA 02138
zipingxu@fas.harvard.edu

Daiqi Gao
Harvard University
Cambridge, MA 02138
dgao@fas.harvard.edu

Susan Murphy[†]
Harvard University
Cambridge, MA 02138
samurphy@fas.harvard.edu

Abstract

In reinforcement learning (RL), different reward functions may lead to the same optimal policy, while some reward functions can be substantially easier to learn. This paper proposes a framework that constructs surrogate rewards based on mediators between actions and rewards, informed by expert-provided causal directed acyclic graphs (DAGs). These DAGs encode domain knowledge from scientists. Under the surrogacy assumption, which assumes that the mediator fully captures all causal links from the action to the reward, we prove that our surrogate reward is unbiased and has reduced variance compared to the original reward. Specifically, we introduce an online reward-design agent that adaptively learns a surrogate reward within an unknown environment. Combined with standard online learning oracles, we show that the regret guarantees can be improved. Furthermore, our framework highlights improvement even without the surrogacy assumption, when total step T is small compared to the error violating the surrogacy assumption. We complement the theoretical analysis with simulation studies, demonstrating significant performance improvement.

Keywords: Reward design; Causal DAGs; Reinforcement learning; Bandit

Acknowledgements

This research was funded by NIH grants P50DA054039, P41EB028242, UH3DE028723, 2R01HL125440-06A1, and 5P30AG073107-03 GY3 Pilots. Susan Murphy holds concurrent appointments at Harvard University and as an Amazon Scholar. This paper describes work performed at Harvard University and is not associated with Amazon.

[*] Work done during an internship at Harvard.

[†] Corresponding author.

1 Introduction

Reinforcement learning (RL), in which an agent sequentially interacts with an unknown environment and learns to maximize cumulative rewards, is widely used for solving decision-making problems. The reward function, often a mapping from the environment's state to a scalar, is a crucial part of RL, as it determines the agent's learning signal. Although many RL problems have a pre-specified reward function, prior literature [1, 2] has shown that a well-designed reward function can lead to faster learning while maintaining the same optimal policy.

This paper addresses a common challenge in reward design: how can we design a reward function to reduce the variance in learning when the original reward to be maximized is **noisy**? This problem is motivated by real-world applications like mobile health [3, 4] and advertising [5], where the reward is defined by highly noisy human behaviors. Figure 1 illustrates an example in smoking cessation [6], where a mobile app delivers digital interventions to prompt users' stress management behavior. These interventions reduce the stress level, thereby decreasing the odds of subsequent smoking behavior. In this scenario, the causal effect between prompts and stress levels is often less noisy than that between stress levels and smoking behavior.

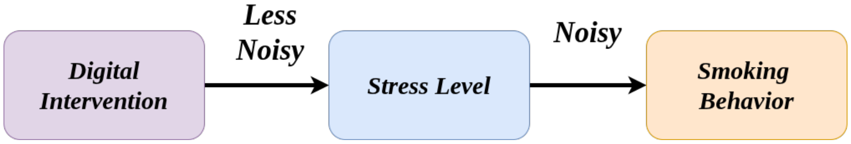


Figure 1: A noisy reward example in mobile health, where a digital intervention reduces smoking behavior by prompting stress management behavior, whose effect is mediated by stress level.

In domains like behavior psychology, the domain knowledge is often available in the form of causal directed acyclic graphs (DAGs) representing causal relations between variables. These causal relations are usually the domain experts' knowledge of the major causal effects between variables in the system [7], and they may be included in the RL agent design to improve the learning efficiency [8, 9], especially in scientific fields [10, 11]. This paper aims to design surrogate rewards based on the **mediators**, the variables that mediate causal paths between actions and rewards. For example, the *stress level* is the mediator between digital interventions and smoking behavior in Figure 1.

Related work. The term **reward engineering** is used interchangeably with **reward design** in [12], where both terms refer to the creation and modification of reward functions to align the learned policy with the goal of the task. However, in other literature [13], **reward engineering** refers to creating a reward function that aligns with the human intent. Our method differs from the latter setting as we directly observe rewards. **Reward shaping** involves modifying the reward function to guide exploration or reduce variance in the learning process, without changing the optimal policy [1]. Existing methods such as potential-based reward shaping [14] and dense reward functions that allocate a sparse reward into multiple steps [15, 2] guide exploration but rarely address high-noise scenarios, which is the central issue tackled in this paper. Our work uniquely focuses on mediator-based surrogate rewards to reduce variance from immediate noisy rewards.

The mediator-based surrogate index was introduced in [16] to estimate the average treatment effect under the surrogacy assumption, with the primary outcome unobserved in the experimental sample and the treatment variable unobserved in the observational sample. In policy learning, [10] imputed missing long-term outcomes using the mediator-based surrogate index and maximized the imputed outcomes via off-policy optimization in a batched bandit setting. Our work differs from the above by adaptively learning the surrogate rewards online. **A key feature of our work is its modularity**—we develop an online reward design agent that can interface with any existing online bandit oracle, which is an agent making decisions in bandit environments; thus separating the reward design from the decision-making process. A technical challenge of this general framework is the non-stationarity in the surrogate reward due to the reward-learning process. We propose to use an adversarial bandit oracle to address this challenge.

Contributions. We propose a framework for mediator-based reward design in contextual bandits. Under the surrogacy assumption—that the reward is independent of the action given the mediator and state—we prove that our surrogate reward is unbiased with respect to the true expected reward and has a strictly lower variance. Further, we propose an online reward design agent that adaptively learns the target mediator-based reward function and integrates it with any existing online bandit oracle for online decision-making. We show theoretically that, when employing an adversarial bandit oracle to manage reward non-stationarity, our method achieves tighter regret bounds compared to standard linear UCB algorithms, with or without the surrogacy assumption. (3) We validate the performance improvement with simulation experiments.

2 Problem Setup

Notation. We use I_d to denote the identity matrix of dimension $d \times d$, and $\| \cdot \|$ to denote the L_2-norm. A random vector $\mathbf{X} \sim SG_d(\sigma)$ is a d-dimensional sub-Gaussian variable such that, for any $\mathbf{u} \in \mathbb{R}^d$ with $\|\mathbf{u}\| = 1$, $\mathbb{E}[\exp(\mathbf{u}^\top \mathbf{X} t)] \leq \exp(\sigma^2 t^2/2), \forall t \in \mathbb{R}$.

Contextual bandit with mediators. We consider a contextual bandit problem where the agent at each step $t = 1, \ldots, T$ observes context $\boldsymbol{S}_t \in \mathcal{S} \subset \mathbb{R}^{d_S}$, based on which it chooses an action $A_t \in \mathcal{A}$. The environment reveals the mediator $\boldsymbol{M}_t \in \mathbb{R}^{d_M}$ and the reward $R_t \in \mathbb{R}$ jointly from a distribution $P(\cdot \mid \boldsymbol{S}_t, A_t)$, where each $P(\cdot \mid \boldsymbol{s}, a)$ is a probability distribution over $\mathbb{R}^{d_M+1}$. The agent aims to optimize the cumulative reward $\sum_{t=1}^T R_t$. We define the mean reward function $R(\boldsymbol{s}, a) = \mathbb{E}[R_t \mid A_t = a, \boldsymbol{S}_t = \boldsymbol{s}]$.

Mediator-based surrogate reward. Figure 2 demonstrates a typical DAG for contextual bandit problems. When the dashed line is absent, $\boldsymbol{M}_t$ blocks all causal paths from A_t to R_t. In the causal inference literature, this is called surrogacy (Assumption 1). In our theoretical and numerical analysis, we will **allow weak violation of the surrogacy assumption** to accommodate different real world scenarios. Although it is implicitly assumed in general, we emphasize that **causal sufficiency** is assumed in our DAG, which means that for a set of variables V, every common cause of any pair of variables in V is also in V [17].

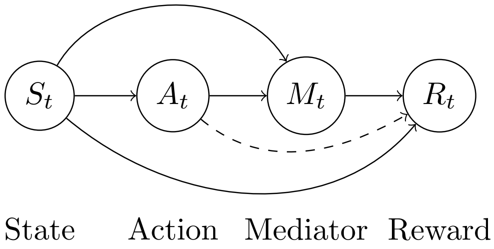


Figure 2: A DAG for contextual linear bandit with mediators. The surrogacy assumption is violated **when the dashed line between R and A exists**.

Assumption 1 (Surrogacy, [18]). *The reward is independent of the action given the mediator and the state, i.e.* $R_t \perp A_t \mid \boldsymbol{M}_t, \boldsymbol{S}_t$.

We aim to design a surrogate reward $\tilde{R}_t = f(\boldsymbol{M}_t, \boldsymbol{S}_t)$ for some function $f : \mathbb{R}^{d_M} \times \mathcal{S} \mapsto \mathbb{R}$. We show in Proposition 1 that under surrogacy, $\tilde{R}_t = \mathbb{E}[R_t \mid \boldsymbol{M}_t, \boldsymbol{S}_t]$ is an unbiased surrogate reward having lower variance than the true reward.

Proposition 1 (Unbiasedness under surrogacy). *Define* $\tilde{R}^*(\boldsymbol{m}, \boldsymbol{s}) = \mathbb{E}[R_t \mid \boldsymbol{M}_t = \boldsymbol{m}, \boldsymbol{S}_t = \boldsymbol{s}]$, *the mean function of* $\tilde{R}_t$ *given* $\boldsymbol{M}_t = \boldsymbol{m}$ *and* $\boldsymbol{S}_t = \boldsymbol{s}$. *If Assumption 1 holds, we have*

$$\mathbb{E}[R_t \mid A_t = a, \boldsymbol{S}_t = \boldsymbol{s}] = \mathbb{E}[\tilde{R}^*(\boldsymbol{M}_t, \boldsymbol{S}_t) \mid A_t = a, \boldsymbol{S}_t = \boldsymbol{s}], \tag{1}$$

$$\operatorname{Var}(R_t \mid A_t = a, \boldsymbol{S}_t = \boldsymbol{s}) \geq \operatorname{Var}(\tilde{R}^*(\boldsymbol{M}_t, \boldsymbol{S}_t) \mid A_t = a, \boldsymbol{S}_t = \boldsymbol{s}) \tag{2}$$

for any $a \in \mathcal{A}$ *and* $\boldsymbol{s} \in \mathcal{S}$. *The inequality strictly holds if and only if there exists some subset* $\mathcal{M} \subseteq$ *supp*$(\boldsymbol{M}_t | A_t = a, \boldsymbol{S}_t = \boldsymbol{s})$ *with* $P(\mathcal{M}) > 0$, *such that* $\operatorname{Var}(R_t | \boldsymbol{M}_t = \boldsymbol{m}, A_t = a, \boldsymbol{S}_t = \boldsymbol{s}) > 0$ *for all* $\boldsymbol{m} \in \mathcal{M}$.

A linear working model. Throughout the paper, we make an additional linear assumption about the reward and mediator generating processes:

$$R_t = \boldsymbol{\theta}_{A_t}^\top \boldsymbol{S}_t + \boldsymbol{\theta}_S^\top \boldsymbol{S}_t + \boldsymbol{\theta}_M^\top \boldsymbol{M}_t + \epsilon_t, \text{ and } \boldsymbol{M}_t = \boldsymbol{\Gamma}_{A_t} \boldsymbol{S}_t + \boldsymbol{\omega}_t, \tag{3}$$

where $\boldsymbol{\theta}_a \in \mathbb{R}^{d_S}$, $\boldsymbol{\Gamma}_a \in \mathbb{R}^{d_M \times d_S}$ for each $a \in \mathcal{A}$, $\boldsymbol{\theta}_S \in \mathbb{R}^{d_S}, \boldsymbol{\theta}_M \in \mathbb{R}^{d_M}$, and $\epsilon_t \sim SG_1(\sigma_\epsilon)$, $\boldsymbol{\omega}_t \sim SG_{d_M}(\sigma_\omega)$. Note that we have $\boldsymbol{\theta}_{A_t}$ here to **allow violation of the surrogacy assumption**, which holds when $\boldsymbol{\theta}_a \equiv \mathbf{0}$ for all $a \in \mathcal{A}$.

Policy and regret. We define a policy π as a mapping from context to action, i.e., $\pi : \mathcal{S} \mapsto \mathcal{A}$. The performance of an online algorithm is measured by regret, which is defined as the difference between the expected cumulative rewards generated by the optimal policy and the expected cumulative rewards generated by the algorithm. Assuming $\pi^*(\boldsymbol{s}) := \arg\max_a R(\boldsymbol{s}, a)$ is the optimal policy, we define the regret as

$$\operatorname{Regret}_T := \sum_{t=1}^T \left(R(\boldsymbol{S}_t, \pi^*(\boldsymbol{S}_t)) - R(\boldsymbol{S}_t, A_t) \right). \tag{4}$$

3 Online Reward Design with Bandit Oracle

Our proposed solution has two critical components, an **online reward design agent** that adaptively learns a reward mapping based on $\boldsymbol{S}_t$ and $\boldsymbol{M}_t$, and an online bandit oracle. The **online bandit oracle** is a bandit algorithm (e.g. [19]) that at each decision time t takes a tuple $(\boldsymbol{S}_t, A_t, \tilde{R}_t)$ as the input and outputs the next policy π_{t+1}, a mapping from the state space $\mathcal{S}$ to a distribution over the action space $\mathcal{A}$. The **online reward design agent** is learning the function

Algorithm 1 Mediator-based Online Reward Learning

```
Input: online bandit oracle $\mathcal{O}$; dataset for reward design $\mathcal{D}_0 = \{\}$; ridge regularization parameter $\lambda$.
Initialize $\pi_1$ by the uniform random policy.
for $t = 1, \ldots, T$ do
    Observe current state $\boldsymbol{S}_t$
    Sample $A_t \sim \pi_t(\boldsymbol{S}_t)$, and environment generates $(R_t, \boldsymbol{M}_t) \sim P(\cdot \mid \boldsymbol{S}_t, A_t)$
    Add observation for reward design $\mathcal{D}_t = \mathcal{D}_{t-1} \cup \{(\boldsymbol{S}_t, \boldsymbol{M}_t, R_t)\}$
    Run ridge regression in (5) on $\mathcal{D}_t$ to get estimators $\hat{\boldsymbol{\theta}}_{S,t}$ and $\hat{\boldsymbol{\theta}}_{M,t}$
    Construct reward $\tilde{R}_t = \hat{\boldsymbol{\theta}}_{S,t}^\top \boldsymbol{S}_t + \hat{\boldsymbol{\theta}}_{M,t}^\top \boldsymbol{M}_t$
    Query bandit oracle $\pi_{t+1} = \mathcal{O}(\boldsymbol{S}_t, A_t, \tilde{R}_t)$
end for
```

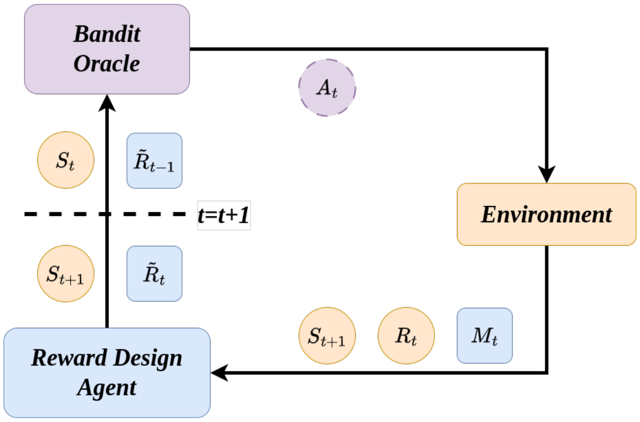


Figure 3: At each decision time t, the oracle observes the state $\boldsymbol{S}_t$, chooses an action $A_t \in \mathcal{A}$, and receives the surrogate reward $\tilde{R}_t$. Variables in **orange (circle)** are observed variables in the standard bandit setting, and those in **blue (rounded square)** are the additional variables introduced in this framework for surrogate reward design. Action is in **purple (dashed circle)** to show that it's dependent on the oracle.

$\tilde{R}^*(\boldsymbol{m}, \boldsymbol{s}) = \mathbb{E}[R_t | \boldsymbol{M}_t = \boldsymbol{m}, \boldsymbol{S}_t = \boldsymbol{s}] = \boldsymbol{s}^\top \boldsymbol{\theta}_S + \boldsymbol{m}^\top \boldsymbol{\theta}_M$, where $\boldsymbol{\theta}_S$ and $\boldsymbol{\theta}_M$ are the coefficients of the linear model in (3). The interaction between the online reward design agent and the online bandit oracle is characterized in Figure 3.

We use online ridge regression to learn the coefficients of the linear model. At each decision time t, the estimators are

$$\begin{bmatrix} \hat{\boldsymbol{\theta}}_{S,t} \\ \hat{\boldsymbol{\theta}}_{M,t} \end{bmatrix} = \left(\sum_{\tau=1}^{t-1} \boldsymbol{X}_\tau \boldsymbol{X}_\tau^\top + \lambda \mathbf{I}_{d_S + d_M} \right)^{-1} \sum_{\tau=1}^{t-1} \boldsymbol{X}_\tau R_\tau, \tag{5}$$

where λ is a tuning parameter and $\boldsymbol{X}_t = (\boldsymbol{S}_t^\top, \boldsymbol{M}_t^\top)^\top$ denotes the regression covariates. The overall process is described in Algorithm 1.

4 Results

We first introduce an additional regularity assumption concerning the boundedness of the coefficients in (3).

Assumption 2 (Bounded coefficients). *We assume that there exist constants $\mathcal{E}, C > 0$ s.t. (1) $\forall t \in \{1, 2, \cdots, T\}, \|S_t\| \leq 1$ almost surely; (2) $\max_{a \in \mathcal{A}} \|\Gamma_a\| \leq C, \max_{a \in \mathcal{A}} \|\theta_a\| \leq \mathcal{E}C$; (3) $\|\theta_M\| \leq C$.*

*Note that **surrogate error** $\mathcal{E}$ controls the ratio of the effect of A_t on R_t not captured by M_t to that captured by M_t. A greater $\mathcal{E}$ implies more severe violation of the surrogacy assumption. $\mathcal{E} = 0$ implies surrogacy (Assumption 1).*

Regret Bound. We now present the high-probability regret bound for Algorithm 1 when the adversarial online bandit oracle is chosen as *RealLinExp3* [19]. The standard proof is extended to incorporate the noise of the reward.

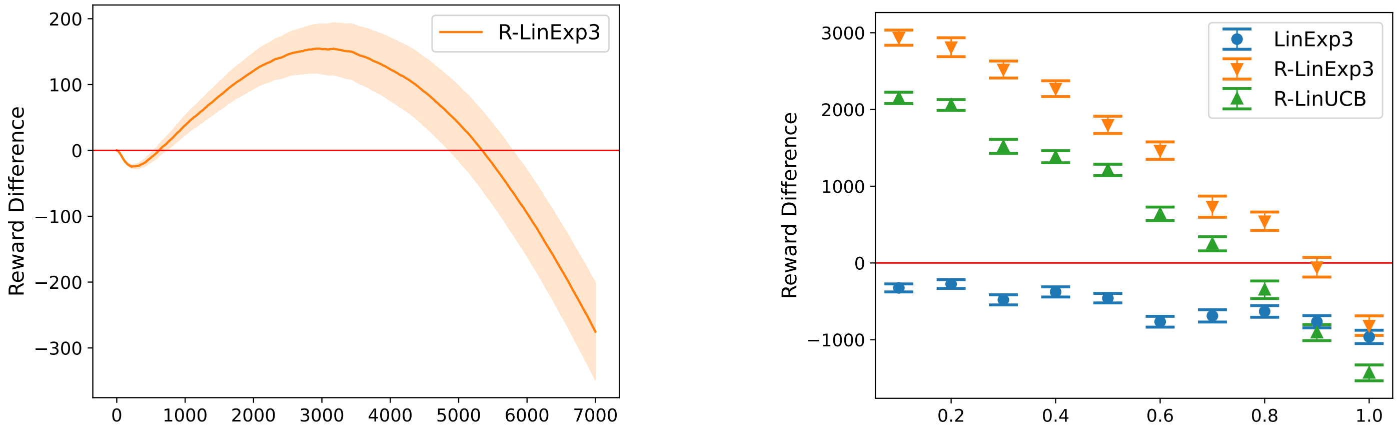
Theorem 4.1 (Regret bound for Algorithm 1)**.** *When Assumption 2 holds, the regret of Algorithm 1 is bounded with a high probability by*

$$\tilde{\mathcal{O}} \left(\sigma_\omega \sqrt{d_S |\mathcal{A}| T} + \sigma_\epsilon \sqrt{(d_S + d_M) T} + \mathcal{E} T \right). \tag{6}$$

Now we compare our regret bound with the original bound, $\tilde{\mathcal{O}}$ signs omitted. The minimax regret bound for stochastic contextual linear bandits is $\sigma_\omega \sqrt{d_S |\mathcal{A}| T} + \sigma_\epsilon \sqrt{d_S |\mathcal{A}| T}$. To interpret the improvement, we note that our regret bound in (6) has a reduction in a multiplicative factor of $\sqrt{|\mathcal{A}|}$ at the price of an additional $\sqrt{d_M}$ term. This improvement is due to the fact that the ridge regression estimator pools information across all arms to learn the effect of the mediator on the reward, while naive contextual linear bandits with a discrete action set estimates the coefficient of each arm separately. The price of making this improvement is that we now need to run an additional regression based on the mediator, hence the additional $\sqrt{d_M}$ term. The linear term $\mathcal{E}T$ comes from the bias in the surrogate reward due to the violation of Assumption 1. Accordingly, our method has better regret bound when

$$\sqrt{T} \leq \frac{\sigma_\epsilon}{\mathcal{E}} (\sqrt{d_S |\mathcal{A}|} - \sqrt{d_S + d_M}). \tag{7}$$

The improvement is significant when: (1) $\sigma_\omega \ll \sigma_\epsilon$: the mediator generating process is much less noisy than the primary reward generating process; (2) $d_S \gg d_M$: the context is of a much higher dimension than the mediator; (3) large $|\mathcal{A}|$: we have a relatively large action set; and (4) small $\mathcal{E}$: the action has a small uncaptured direct effect size on the reward.



(a) Round T

(b) Surrogate error $\mathcal{E}$

Figure 4: Performance of R-LinExp3, R-LinUCB and LinExp3 compared to that of LinUCB. The y-axis is reward difference, which is the difference between the cumulative reward of the algorithm and that of the baseline (LinUCB here). The error bars are the standard errors over 700 independent runs.

Simulation. Here we emphasize the effect of different **number of rounds** T and **surrogate error** $\mathcal{E}$. See Figure 4. The two base algorithms are LinExp3 [19], an adversarial bandit oracle, and LinUCB [5], a stochastic bandit oracle. For each oracle, we test both how the naive version learns on the original reward, and the version with the reward design agent learns on the surrogate reward. We refer to the latter versions as R-LinExp3 and R-LinUCB, respectively. Figure 4a depicts how the round T affects the performance. We can see that at the start, our algorithm's performance is inferior to that of LinUCB because the reward design agent needs some data on which to "warm up". Before warming up, the surrogate reward cannot accurately indicate the contribution of the chosen action. With a warm start, though, our algorithm outperforms LinUCB until T reaches a relatively high level, where the influence of bias preponderates over that of variance. This behavior is consistent with the outcome in (7). Figure 4b shows that an increase in surrogate error would reduce the performance of the reward design agent, which is consistent with the outcome in (6).

References

[1] A. D. Laud. *Theory and application of reward shaping in reinforcement learning*. University of Illinois at Urbana-Champaign, 2004.

[2] J. Eschmann. "Reward function design in reinforcement learning". In: *Reinforcement Learning Algorithms: Analysis and Applications* (2021), pp. 25–33.

[3] A. L. Trella et al. "Reward design for an online reinforcement learning algorithm supporting oral self-care". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 2023, pp. 15724–15730.

[4] S. Ghosh et al. "Did we personalize? assessing personalization by an online reinforcement learning algorithm using resampling". In: *Machine Learning* (2024), pp. 1–37.

[5] L. Li et al. "A contextual-bandit approach to personalized news article recommendation". In: *Proceedings of the 19th international conference on World wide web*. 2010, pp. 661–670.

[6] S. L. Battalio et al. "Sense2Stop: a micro-randomized trial using wearable sensors to optimize a just-in-time-adaptive stress management intervention for smoking relapse prevention". In: *Contemporary Clinical Trials* 109 (2021), p. 106534.

[7] A. Gopnik et al. "A theory of causal learning in children: causal maps and Bayes nets." In: *Psychological review* 111.1 (2004), p. 3.

[8] Z. Deng et al. "Causal Reinforcement Learning: A Survey". In: *Transactions on Machine Learning Research* (2023).

[9] E. Bareinboim, J. Zhang, and S. Lee. *An Introduction to Causal Reinforcement Learning*. Tech. rep. R-65. Causal Artificial Intelligence Lab, Columbia University, Dec. 2024.

[10] J. Yang et al. "Targeting for long-term outcomes". In: *Management Science* 70.6 (2024), pp. 3841–3855.

[11] D. Gao et al. "Harnessing Causality in Reinforcement Learning with Bagged Decision Times". In: *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*. PMLR, 2025, pp. 658–666.

[12] T. Hirtz et al. "Unsupervised reward engineering for reinforcement learning controlled manufacturing". In: *Journal of Intelligent Manufacturing* (2024), pp. 1–14.

[13] A. Gupta et al. "Unpacking reward shaping: Understanding the benefits of reward engineering on sample complexity". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 15281–15295.

[14] A. Ng, D. Harada, and S. J. Russell. "Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping". In: *International Conference on Machine Learning*. 1999.

[15] J. Hare. *Dealing with Sparse Rewards in Reinforcement Learning*. 2019. arXiv: `1910.09281 [cs.LG]`.

[16] S. Athey et al. *The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely*. Tech. rep. National Bureau of Economic Research, 2019.

[17] P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search*. MIT press, 2001.

[18] R. L. Prentice. "Surrogate endpoints in clinical trials: definition and operational criteria". In: *Statistics in medicine* 8.4 (1989), pp. 431–440.

[19] G. Neu and J. Olkhovskaya. "Efficient and robust algorithms for adversarial linear contextual bandits". In: *Conference on Learning Theory*. PMLR. 2020, pp. 3049–3068.