A Distribution Mapping Approach to Counterfactually Fair Reinforcement Learning

Jianhan Zhang[1], Jitao Wang[1], John D. Piette[1],
Donglin Zeng[1], Chengchun Shi[2], Zhenke Wu[1]

[1]University of Michigan [2]London School of Economics and Political Science

NeurIPS 2026

Motivating Example: the PowerED Study (Piette et al., 2023)

- **Goal:** reduce opioid analgesic misuse while conserving counselors' time.
- **Method:** 12-week RL-based intervention.
- **Decision:** each week, assign one of three treatments, from automated calls to live counselor calls.
- **Observed states:** self-reported pain and pain interference.

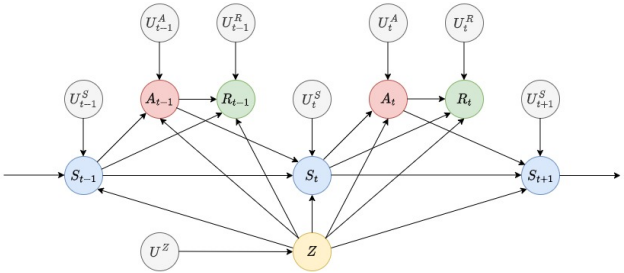
How unfairness can arise

Differences in how patients report pain may cause an RL policy to allocate counselor time differently across sensitive groups, even without using the sensitive attribute directly.

Our Contribution

1. **Method:** CFSMDM estimates counterfactual states and rewards sequentially to learn approximately counterfactually fair policies.
2. **Weaker assumptions:** CFSMDM imposes strictly weaker assumptions than the additive noise assumption used by existing alternatives.
3. **Theory:** bounds on suboptimality gap and level of counterfactual unfairness.
4. **Evidence:** synthetic, semi-synthetic, and real-data evaluations.

Contextual Markov Decision Process (CMDP)



- Z: sensitive attribute
- S_t: state; A_t: action; R_t: reward
- U_t: exogenous noise
- Arrows are represented by **structural equations**.

A counterfactual world changes Z while holding the person's historical exogenous noise and past actions fixed.

Counterfactual Fairness (CF) in a CMDP

Definition (Wang et al., 2025)

For the same observed individual and historical trajectory, changing the sensitive attribute from z to z', while holding constant all historical actions and exogenous variables, should not change the probability distribution of the action:

$$\mathbb{P}^{\pi_t}(A_t^{Z \leftarrow z'}(\bar{U}_t(h_t)) = a) = \mathbb{P}^{\pi_t}(A_t^{Z \leftarrow z}(\bar{U}_t(h_t)) = a).$$

Illustration: fair actions

Sensitive Attribute	Week 1	Week 2	Week 3	Week 4	Week 5
Z=0	**Treatment**	**Treatment**	**No Treatment**	**Treatment**	**No Treatment**
Z=1	Treatment	Treatment	No Treatment	Treatment	No Treatment

Illustration: unfair actions

Sensitive Attribute	Week 1	Week 2	Week 3	Week 4	Week 5
Z=0	**Treatment**	**Treatment**	**No Treatment**	**Treatment**	**No Treatment**
Z=1	Treatment	No Treatment	No Treatment	Treatment	No Treatment

Note: the examples assume deterministic policies, while the formal criterion compares action distributions.

Assumptions: Strict Monotonicity in the Exogenous Noise

Assumption 1 — State Strict Monotonicity

Each component of the initial state and each subsequent state is a **strictly increasing function of its own scalar exogenous noise**, holding the sensitive attribute, previous state and previous action fixed.

$$S_0^{(i)} = f_0^{S^{(i)}}(Z, U_0^{S^{(i)}}),$$

$$S_t^{(i)} = f^{S^{(i)}}(Z, S_{t-1}, A_{t-1}, U_t^{S^{(i)}}), \quad t \geq 1.$$

Assumption 2 — Reward Strict Monotonicity

The reward is **strictly increasing in its scalar exogenous noise** conditional on (Z, S_t, A_t).

$$R_t = f^R(Z, S_t, A_t, U_t^R).$$

Why this matters: it enables counterfactual estimation without assuming additive noise.

The CFSMDM Algorithm

Building on Wang et al. (2025):

$$\widetilde{S}_t = (S_t^{z^{(1)}}, \ldots, S_t^{z^{(K)}}) \quad \widetilde{R}_t = \sum_{k=1}^{K} P(Z = z^{(k)}) R_t^{z^{(k)}}.$$

With *true* counterfactuals, policies learned on these augmented states and rewards are counterfactually fair.

CFSMDM in three steps:

1. **Estimate** counterfactual states $S_t^{z^{(k)}}$ and rewards $R_t^{z^{(k)}}$ by sequential marginal distribution mapping (to be discussed later).
2. **Construct** estimated augmented states and rewards $(\widehat{\widetilde{S}}_t, \widehat{\widetilde{R}}_t)$, with $\widehat{\widetilde{S}}_t = (\widehat{S}_t^{z^{(1)}}, \ldots, \widehat{S}_t^{z^{(K)}})$ and $\widehat{\widetilde{R}}_t = \sum_{k=1}^{K} \widehat{P}(Z = z^{(k)}) \widehat{R}_t^{z^{(k)}}$.
3. **Learn** a policy on the processed trajectories using offline RL (e.g., FQI).

Counterfactual State and Reward Estimation

Key result: conditional quantile level is preserved across counterfactual worlds

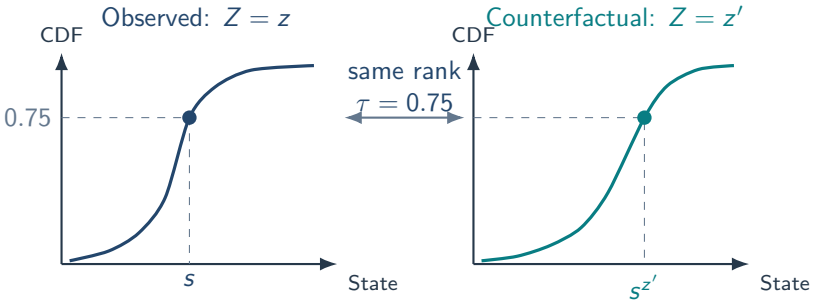
Under the strict monotonicity assumptions, an individual's **latent rank** remains the same when their sensitive attribute is counterfactually changed.

Initial-state example (one state component):

$$s_0^{z'} = F^{-1}_{S_0|Z=z'}\left(F_{S_0|Z=z}(s_0)\right)$$

- **States at later steps:** condition on the preceding *counterfactual* state and the *observed* past action, and then match the conditional ranks.
- **Rewards:** condition on the current *counterfactual* state and the *observed* current action, and then match the conditional ranks.

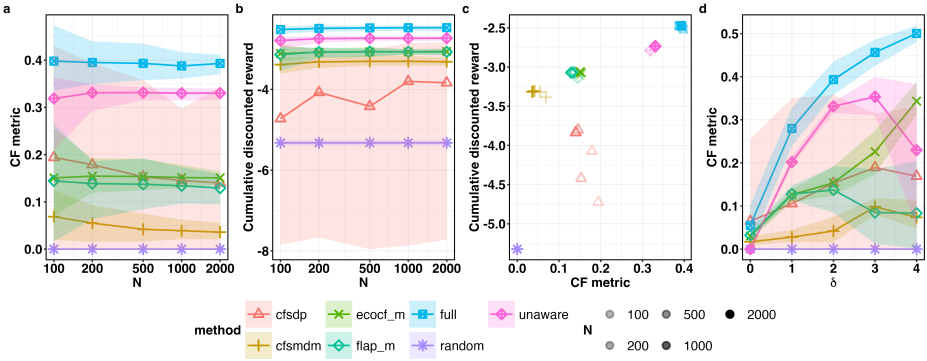
Counterfactual State and Reward Estimation (cont'd)



Same conditional percentile $\longrightarrow$ corresponding counterfactual value

For later steps, repeat conditionally using the previously estimated counterfactual state.

Synthetic Data Experiments (in a non-additive-noise environment)



Takeaway: Despite the noise being non-additive, CFSMDM improves counterfactual fairness relative to other non-random methods while retaining useful policy value. 50 replications; points are empirical means and ribbons show the 2.5th–97.5th empirical quantiles. Lower CF metric is better.

Application to the PowerED Study Data

Mean estimated CF metric (lower is better); 50 replications, four sensitive attributes

Method	Education	Ethnicity	Sex	Age
Full	0.45	0.46	0.66	0.35
Random	0.00	0.00	0.00	0.00
Unaware	0.17	0.17	**0.06**	0.13
CFSDP	0.14	0.18	0.07	**0.10**
ECOCF_M	0.16	0.20	0.14	0.15
FLAP_M	0.17	0.20	0.14	0.17
CFSMDM	**0.11**	**0.10**	0.09	**0.10**

- CFSMDM reduces the estimated CF metric vs Full for all four sensitive attributes.
- CFSMDM achieves the lowest CF metric among non-random methods under most, though not all, of the four sensitive attributes.

Thank you!

github.com/JianhanZhang/CFSMDM_RL

jianhanz@umich.edu

References I

Piette, J. D., Thomas, L., Newman, S., Marinec, N., Krauss, J., Chen, J., Wu, Z., and Bohnert, A. S. B. (2023). An automatically adaptive digital health intervention to decrease opioid-related risk while conserving counselor time: Quantitative analysis of treatment decisions based on artificial intelligence and patient-reported risk measures. *Journal of Medical Internet Research*, 25:e44165.

Wang, J., Shi, C., Piette, J. D., Loftus, J. R., Zeng, D., and Wu, Z. (2025). Counterfactually fair reinforcement learning via sequential data preprocessing. *arXiv preprint arXiv:2501.06366v2*.