On Welfare-Centric Fair Reinforcement Learning




Cyrus Cousins   Elita Lobo   Kavosh Asadi   Michael Littman

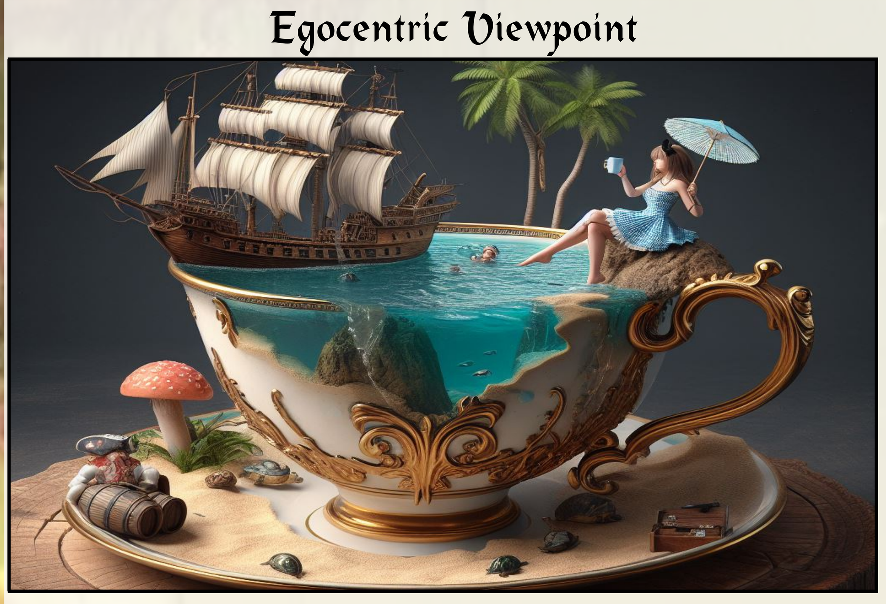
Reinforcement Learning Conference

University *of* Massachusetts Amherst

www.cyruscousins.online/projects/rlfairness/

Reject Egocentricsm

Egocentric Viewpoint



- A acts in 🌏, and 🌏 responds
- Scalar reward $\mathrm{R}(s, a)$ is *intrinsic* to A
- Rational agents selfishly optimize value

$$\operatorname*{argmax}_{\pi \in \Pi} \mathop{\mathbb{E}}_{\pi, s}\left[\sum_{t=0}^{\infty} \gamma^t \mathrm{R}(s_t, \pi(s_t)) \,\middle|\, s_0\right]$$

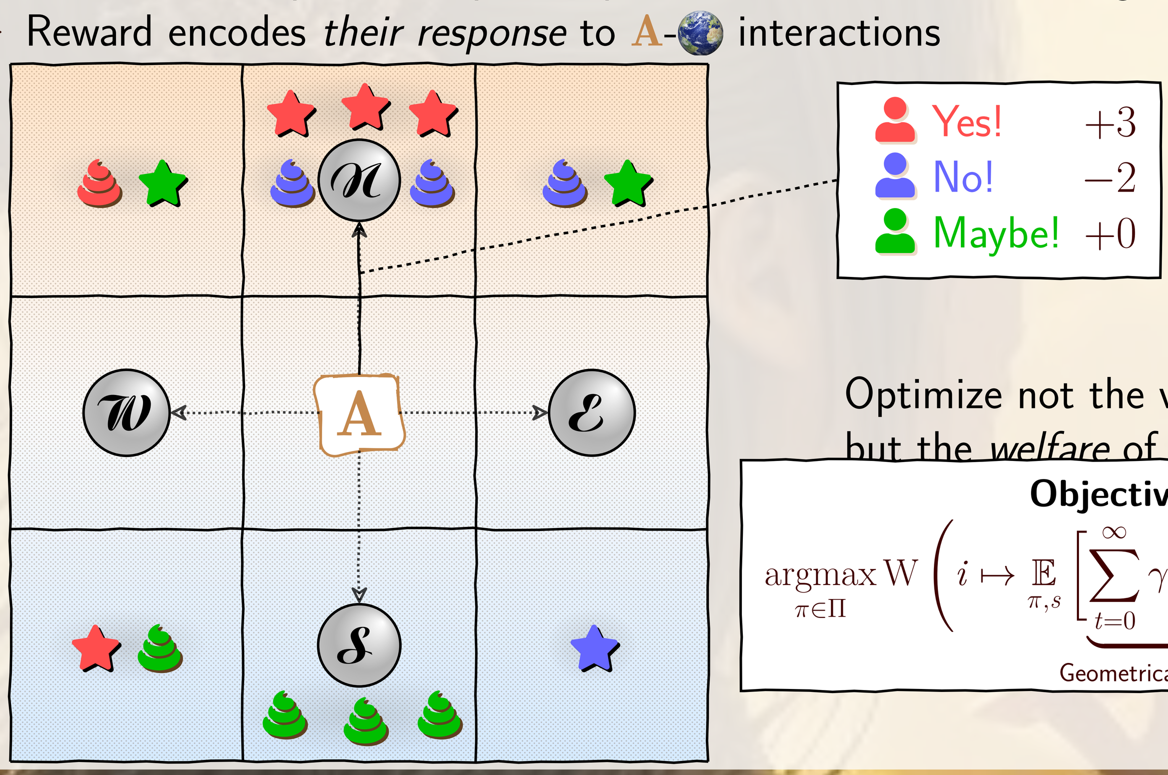
Altruistic Viewpoint

- A's actions in 🌏 impact *beneficiaries*
- Vector reward $\mathbf{R}(s, a)$ quantifies impact
- Altruistic agents optimize *societal welfare*

$$\operatorname*{argmax}_{\pi \in \Pi} \mathrm{W}\left(i \mapsto \mathop{\mathbb{E}}_{\pi, s}\left[\sum_{t=0}^{\infty} \gamma^t \mathbf{R}_i(s_t, \pi(s_t)) \,\middle|\, s_0\right]\right)$$

What is Group-Fair Reinforcement Learning?

- Agent A in world 🌏 receives *vector-valued* reward $\mathbf{R}(s, a) \in \mathbb{R}^g$ for g beneficiaries
  - *Beneficiaries* represent *impacted parties*: Individuals, entities, groups, etc.
  - Reward encodes *their response* to A-🌏 interactions



Optimize not the value of *what* A *wants*, but the *welfare of value functions*

Objective:

$$\operatorname*{argmax}_{\pi \in \Pi} \mathrm{W}\left(i \mapsto \mathop{\mathbb{E}}_{\pi, s}\left[\underbrace{\sum_{t=0}^{\infty} \gamma^t \mathbf{R}_i(s_t, \pi(s_t))}_{\text{Geometrically discounted reward}} \,\middle|\, s_0\right]\right)$$

What is a Welfare Function?

- Utility (value) vector $\boldsymbol{v} \in \mathbb{R}^g_{0+}$:
- $\mathrm{W}(\boldsymbol{v}) : \mathbb{R}^g_{0+} \to \mathbb{R}_{0+}$ *aggregates* utility across beneficiaries

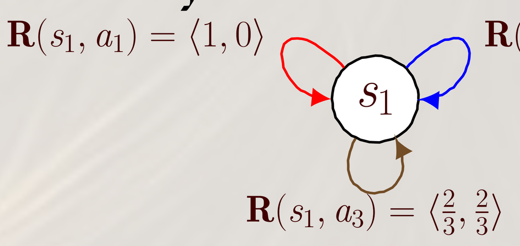
Utilitarian: $\mathrm{W}_1(\boldsymbol{v}) \doteq \frac{1}{g}\sum_{i=1}^{g} \boldsymbol{v}_i$

Egalitarian: $\mathrm{W}_{-\infty}(\boldsymbol{v}) \doteq \min_{i \in 1, \dots, g} \boldsymbol{v}_i$

p **Power-Mean:** $\mathrm{W}_p(\boldsymbol{v}) \doteq \sqrt[p]{\frac{1}{g}\sum_{i=1}^{g} \boldsymbol{v}_i^p}$

Even Bandits are Tricky!

"Compromise" 3-Armed Bandit



$\pi^1 = \langle \mathbf{1}, 0, 0 \rangle$

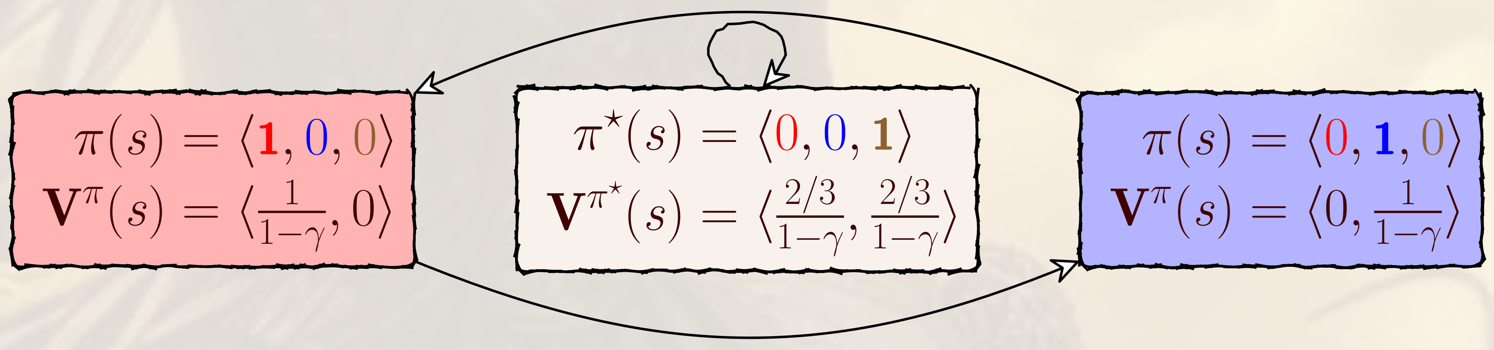
$\pi^2 = \langle 0, \mathbf{1}, 0 \rangle$

$\pi^\star = \langle 0, 0, \mathbf{1} \rangle$

Beneficiary policies π^1 and π^2 and *fair policy* $\pi^\star$ are *disjoint!*

If $\gamma \geq \frac{1}{2}$: Egalitarian policy iteration oscillates indefinitely

$$\pi^{(t+1)} \leftarrow \operatorname*{argmax}_{\pi \in \Pi_{\mathcal{M}}} \mathrm{W}_{-\infty}\left(i \mapsto \mathop{\mathbb{E}}_{\pi, s_1}\left[\mathbf{R}_i(s_0, \pi(s_0)) + \gamma \mathbf{V}_i^{\pi^{(t)}}(s_1)\right]\right)$$



On Planning

- **Policy Iteration**
  - ✗ Nonconvergent; can oscillate indefinitely
- **Value Iteration**
  - With what Bellman operator? Many obstacles here:
    - ✗ Beneficiaries each have their own *value function* $\mathbf{V}_{1:g}$, but not their own *policy* π
    - ✗ No greedy-optimal substructure (start-state dependence)
- Planning with *geometrically-discounted state-action occupancy frequencies*

$$\boldsymbol{d}^\star = \operatorname*{argmax}_{\boldsymbol{d} \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}_{0+}} \mathrm{W}\left(\sum_{s \in \mathcal{S}, a \in \mathcal{A}} \boldsymbol{d}_{s,a} \mathbf{R}_1(s, a), \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \boldsymbol{d}_{s,a} \mathbf{R}_2(s, a), \dots, \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \boldsymbol{d}_{s,a} \mathbf{R}_g(s, a)\right)$$

$$\text{such that } \forall s \in \mathcal{S} : \sum_{a \in \mathcal{A}} \boldsymbol{d}_{s,a} = \boldsymbol{p}_s + \gamma \sum_{s' \in \mathcal{S}, a' \in \mathcal{A}} \mathbf{P}_s(s', a') \boldsymbol{d}_{s',a'} \ ,$$

Take $\pi^\star(s, a) \propto \boldsymbol{d}^\star_{s,a}$ for all $s \in \mathcal{S}$, $a \in \mathcal{A}$

Regret and Mistakes

- Optimal policy is *stochastic*, can't assess individual actions
  - Assess regret of welfare of *agent policies* $\hat{\pi}_1, \dots, \hat{\pi}_T$

$$\mathrm{Regret}(T) = \sum_{t=1}^{T}\left(\mathrm{W}\left(\mathbf{V}^{\pi_t^\star}(s_t)\right) - \mathrm{W}\left(\mathbf{V}^{\hat{\pi}_t}(s_t)\right)\right)$$

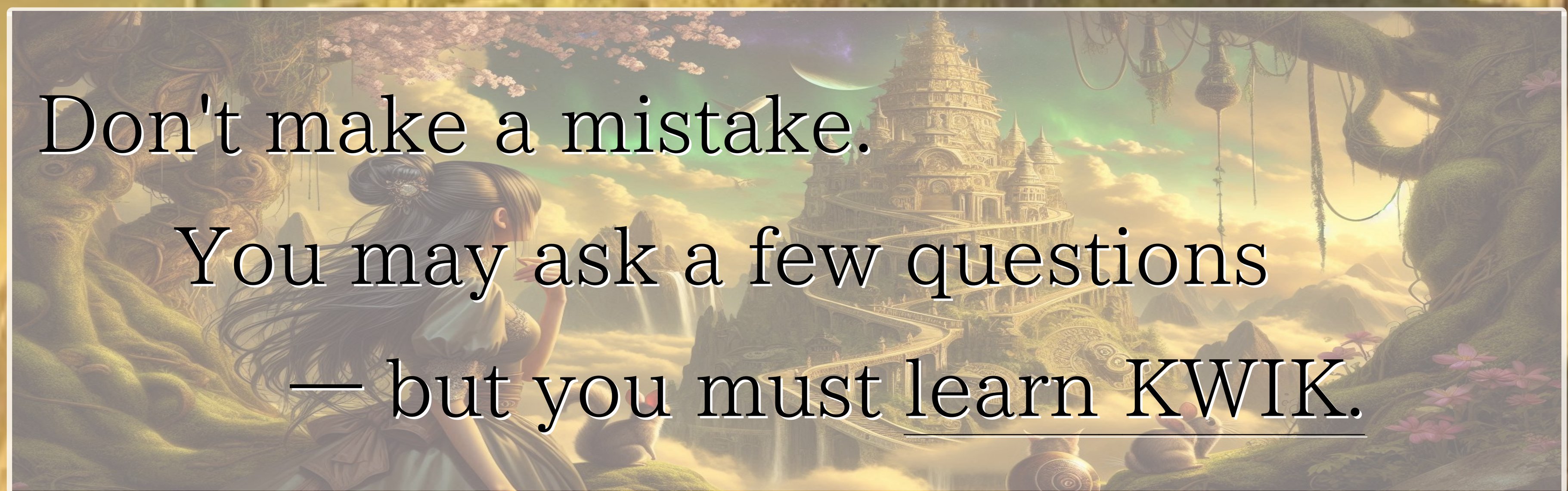

- When should we evaluate the agent?
  - ✗ Incoherent to take $s_{t+1} \sim \hat{\pi}_t(s_t)$
    - *Geometric discounting* suggests *geometric episode length*
    - Unfair to execute each $\hat{\pi}_t(s_t)$ (start-state dependence)
  - ? **Continuous:** Follow $\hat{\pi}_t$ for $\mathrm{Geometric}(1-\gamma)$ steps, resume
  - ? **Episodic:** End episode, draw s_{t+1} from start-state distribution
- A policy $\hat{\pi}$ is a *mistake* at s if $\mathrm{W}\left(\mathbf{V}^{\pi_s^\star}(s)\right) - \mathrm{W}\left(\mathbf{V}^{\hat{\pi}}(s)\right) > \varepsilon$
  - ✗ Exploration actions are probably mistakes
  - ? Can exploitation confidently avoid mistakes?

Learning Model: KWIK-AF

The KWIK-AF Learner

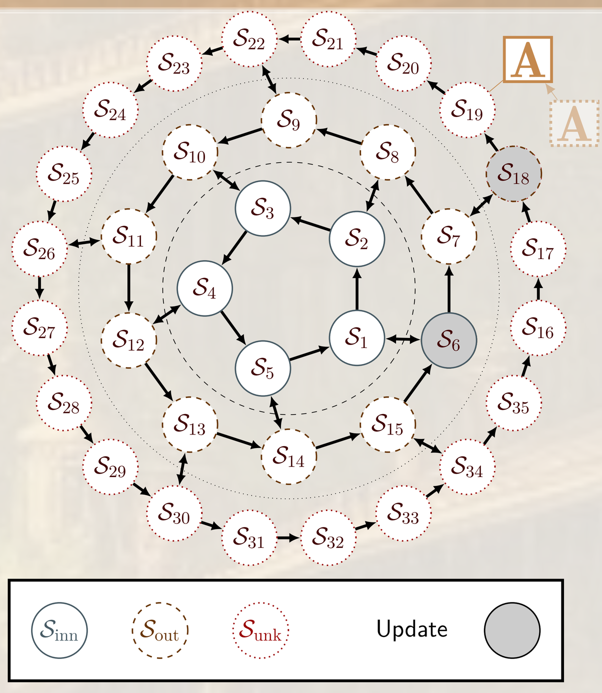
Know-What-It-Knows   *Adversarial-Fair*

- **KWIK Learner:** At each step, agent has two choices:
  1. Output an ε-optimal *exploitation policy* π_{xpt}
     - ✗ With probability at least $1-\delta$, *for all time*
     - ✗ *No mistakes*: $\mathrm{W}\left(\mathbf{V}^{\pi_s^\star}(s)\right) - \mathrm{W}\left(\mathbf{V}^{\pi_{\mathrm{xpt}}}(s)\right) > \varepsilon$
  2. Output an *exploration action* a
     - ✓ Receive $(s, a, \boldsymbol{r}, s')$ tuple, return control to agent in s'
     - ✗ Limited budget: Only $\mathrm{m}\left(|\mathcal{S}|, |\mathcal{A}|, \gamma, \mathrm{R}_{\max}, g, \varepsilon, \delta\right)$ exploration actions, ever
- **Adversarial-Fair:** Algorithm must be *flexible* and *robust*
  - Optimize for *adversarially selected* welfare function $\mathrm{W}(\cdot)$ at each step
  - When A outputs a policy π_{xpt}:
    - Move A to *adversarial* s', provide *no feedback!*



E^4: The Equitable Explicit Explore Exploit Algorithm

- Partition state space into three sets: $\mathcal{S}_{\mathrm{unk}}$, $\mathcal{S}_{\mathrm{out}}$, $\mathcal{S}_{\mathrm{inn}}$
  - **Unknown** $\mathcal{S}_{\mathrm{unk}}$: Insufficient samples (fewer than m_{knw}) to estimate reward $\mathbf{R}(s, a)$ and transition $\mathbf{P}(s)$
  - **Outer-Known** $\mathcal{S}_{\mathrm{out}}$: Some escape policy π_{esc} can reach $\mathcal{S}_{\mathrm{unk}}$ in T steps with probability at least E
  - **Inner-Known** $\mathcal{S}_{\mathrm{inn}}$: No policy can reach $\mathcal{S}_{\mathrm{unk}}$ in T steps with probability at least E
  - Learning moves states from $\mathcal{S}_{\mathrm{unk}} \to \mathcal{S}_{\mathrm{out}} \to \mathcal{S}_{\mathrm{inn}}$



~ The E^4 Algorithm ~

1. If in $\mathcal{S}_{\mathrm{unk}}$: Explore, observe $(s, a, \boldsymbol{r}, s')$, update empirical MDP $\hat{\mathcal{M}}$, update $\mathcal{S}_{\mathrm{unk}}, \mathcal{S}_{\mathrm{out}}, \mathcal{S}_{\mathrm{inn}}$
2. If escape in progress: Follow π_{esc} and decrement timer
3. If in $\mathcal{S}_{\mathrm{out}}$: Begin T-step escape attempt in $\pi_{\mathrm{esc}} \leftarrow \operatorname*{argmax}_{\pi \in \Pi_T} \sum_{s \in \mathcal{S}} \mathop{\mathbb{P}}_{s_{t+1} \sim \mathbf{P}(s_t, \pi(s_t))}\left(\bigvee_{i=0}^{T} s_i \in \mathcal{S}_{\mathrm{unk}} \,\middle|\, s_0 = s\right)$
4. Otherwise in $\mathcal{S}_{\mathrm{inn}}$: Output exploit policy $\pi_{\mathrm{xpt}} \leftarrow \operatorname*{argmax}_{\pi \in \Pi_{\mathcal{M}}} \mathrm{W}\left(\hat{\mathbf{V}}^\pi(s)\right)$

E^4 Theory

- Can set T, E, m_{knw} to ε-δ KWIK-AF learn
- At any point in the execution of E^4, A can act effectively:
  1. Can exploit from $\mathcal{S}_{\mathrm{inn}}$
  2. Can explore directly from $\mathcal{S}_{\mathrm{unk}}$
  3. Can explore indirectly from $\mathcal{S}_{\mathrm{out}}$
     - Escape succeeds with some probability
- E^4 KWIK-AF learns w.r.t. the class of all $\lambda\text{-}\|\cdot\|_\infty$ Lipschitz-continuous welfare functions

Exploration Budget: $\mathrm{m}\left(|\mathcal{S}|, |\mathcal{A}|, \gamma, \mathrm{R}_{\max}, g, \varepsilon, \delta\right) \in \mathbf{O}\left(|\mathcal{S}|^2|\mathcal{A}|\left(\frac{\lambda \mathrm{R}_{\max}}{\varepsilon(1-\gamma)}\log_{\frac{1}{\gamma}}\left(\frac{\lambda \mathrm{R}_{\max}}{\varepsilon(1-\gamma)}\right)\right)^3 \log\frac{|\mathcal{S}||\mathcal{A}|g}{\delta}\right)$

$\subseteq \mathrm{Poly}\left(|\mathcal{S}|, |\mathcal{A}|, \frac{1}{1-\gamma}, \mathrm{R}_{\max}, \log g, \frac{1}{\varepsilon}, \log\frac{1}{\delta}, \lambda\right)$

In Summary

- From Egocentric to Altruistic Agents
  - Agent A acts in 🌏, impacting *beneficiaries*
  - Vector-valued (per-beneficiary) reward $\mathbf{R}(s, a)$
  - Social planner's problem:
    - Optimize *welfare* of *value functions* $\operatorname*{argmax}_{\pi \in \Pi} \mathrm{W}(\mathbf{V}^\pi(s))$
  - Incorporate *fairness* into *sequential learning* problems
- KWIK-AF: A Model of Fair RL
  - Adversarial flexibility
  - Societal welfare objectives
  - Tolerate *no mistakes*, allow *bounded exploration*
  - Challenging model of learning, subsumes PAC-MDP
- Efficient Learning and Planning
  - Learn with E^4: $\mathrm{Poly}(\dots)$ exploration budget
  - Plan with *convex programming* on *state-action measure*
  - Fair RL and classic RL are comparably difficult