On Welfare-Centric Fair Reinforcement Learning

Cyrus Cousins Elita Lobo Kavosh Asadi Michael Littman

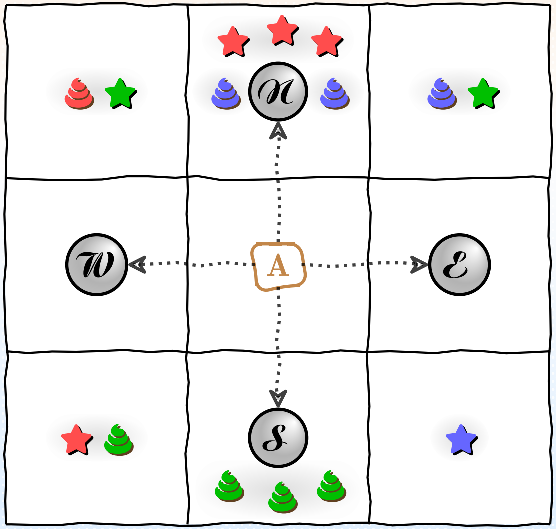
Reinforcement Learning Conference

University of Massachusetts Amherst

www.cyruscousins.online/projects/rlfairness/

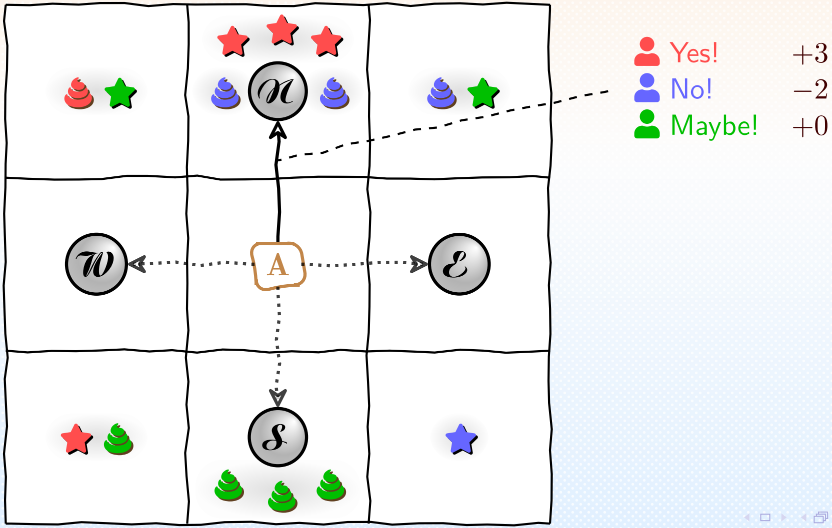
What is Group-Fair Reinforcement Learning?

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



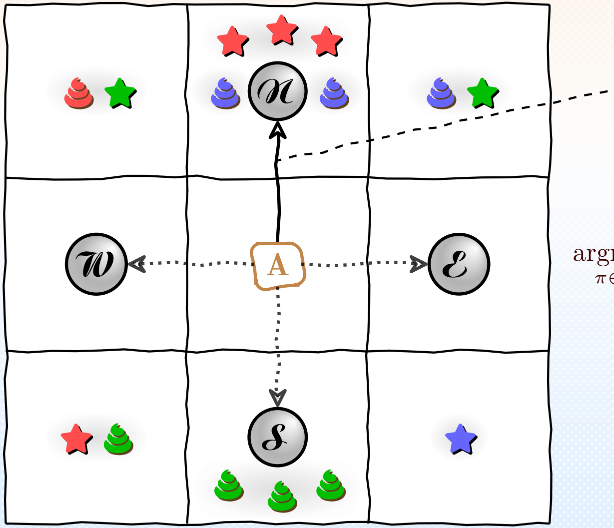
What is Group-Fair Reinforcement Learning?

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



What is Group-Fair Reinforcement Learning?

- Agent $\mathbf{A}$ in world 🌍 receives *vector-valued* reward $\mathbf{R}(s, a) \in \mathbb{R}^g$ from g beneficiaries
  - *Beneficiaries* represent *impacted parties*: Individuals, entities, groups, etc.
  - Reward encodes *their response* to $\mathbf{A}$-🌍 interactions
- Optimize not the value of *what* $\mathbf{A}$ *wants*, but the *welfare* of beneficiary value functions



👤	Yes!	$+3$
👤	No!	-2
👤	Maybe!	$+0$

Objective:

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

Reject Egocentricsm

Egocentric Viewpoint

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

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

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 a Welfare Function?

- Given g beneficiaries
  - Utility (value) vector $\boldsymbol{v} \in \mathbb{R}^{g}_{0+}$

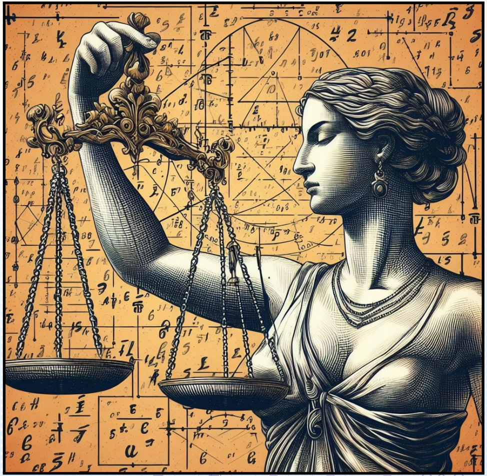
$$\boldsymbol{v} = \left\langle \bigstar\bigstar\bigstar, \bigstar\bigstar, \star \right\rangle$$

What is a Welfare Function?

- Given g beneficiaries
  - Utility (value) vector $\boldsymbol{v} \in \mathbb{R}^{g}_{0+}$

$$\boldsymbol{v} = \left\langle ★★★, ★⯪, ☆ \right\rangle$$

  - $\mathrm{W}(\boldsymbol{v}) : \mathbb{R}^{g}_{0+} \to \mathbb{R}_{0+}$ *aggregates* utility across beneficiaries
    - *Welfare functions* encode *social values*

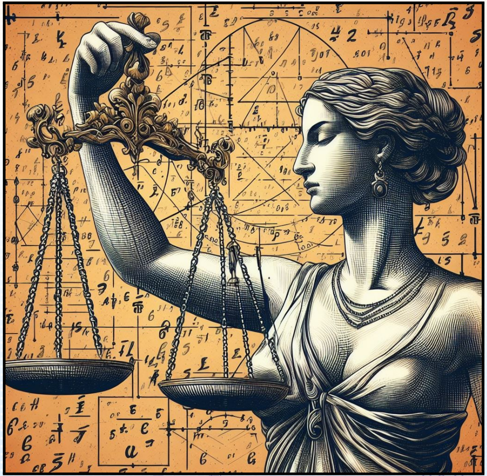


What is a Welfare Function?

- Given g beneficiaries
  - Utility (value) vector $\boldsymbol{v} \in \mathbb{R}^g_{0+}$

$$\boldsymbol{v} = \left\langle ★★★, ★⯪, ☆ \right\rangle$$

  - $\mathrm{W}(\boldsymbol{v}) : \mathbb{R}^g_{0+} \to \mathbb{R}_{0+}$ *aggregates* utility across beneficiaries
    - *Welfare functions* encode *social values*
- Common welfare functions
  - 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: $\mathrm{W}_p(\boldsymbol{v}) \doteq \sqrt[p]{\frac{1}{g}\sum_{i=1}^{g} \boldsymbol{v}_i^p}$

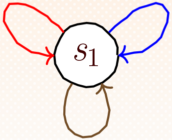


Even Bandits are Tricky!

“Compromise” 3-Armed Bandit

$\mathbf{R}(s_1, a_1) = \langle 1, 0 \rangle$

$\mathbf{R}(s_1, a_2) = \langle 0, 1 \rangle$



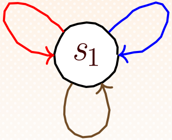
$\mathbf{R}(s_1, a_3) = \langle \frac{2}{3}, \frac{2}{3} \rangle$

Even Bandits are Tricky!

“Compromise” 3-Armed Bandit

$\mathbf{R}(s_1, a_1) = \langle 1, 0 \rangle$

$\mathbf{R}(s_1, a_2) = \langle 0, 1 \rangle$



$\mathbf{R}(s_1, a_3) = \langle \frac{2}{3}, \frac{2}{3} \rangle$

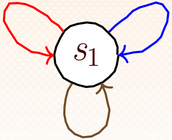
$$\pi^1 = \langle \mathbf{1}, 0, 0 \rangle$$
$$\pi^2 = \langle 0, \mathbf{1}, 0 \rangle$$
$$\pi^\star = \langle 0, 0, \mathbf{1} \rangle$$

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

Even Bandits are Tricky!

"Compromise" 3-Armed Bandit

$\mathbf{R}(s_1, a_1) = \langle 1, 0 \rangle$ $\mathbf{R}(s_1, a_2) = \langle 0, 1 \rangle$



$\mathbf{R}(s_1, a_3) = \langle \frac{2}{3}, \frac{2}{3} \rangle$

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

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

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

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

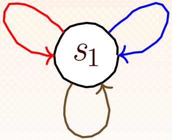
$$\pi(s) = \langle \mathbf{1}, 0, 0 \rangle$$
$$\mathbf{V}^\pi(s) = \langle \tfrac{1}{1-\gamma}, 0 \rangle$$

$$\pi(s) = \langle 0, \mathbf{1}, 0 \rangle$$
$$\mathbf{V}^\pi(s) = \langle 0, \tfrac{1}{1-\gamma} \rangle$$

Even Bandits are Tricky!

"Compromise" 3-Armed Bandit

$\mathbf{R}(s_1, a_1) = \langle 1, 0 \rangle$ $\qquad$ $\mathbf{R}(s_1, a_2) = \langle 0, 1 \rangle$



$\mathbf{R}(s_1, a_3) = \langle \frac{2}{3}, \frac{2}{3} \rangle$

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

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

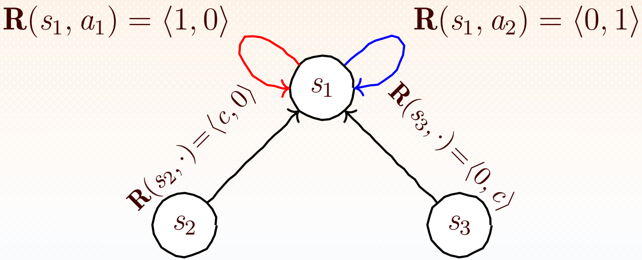
If $\gamma \geq \frac{1}{2}$: Egalitarian policy iteration oscillates indefinitely

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

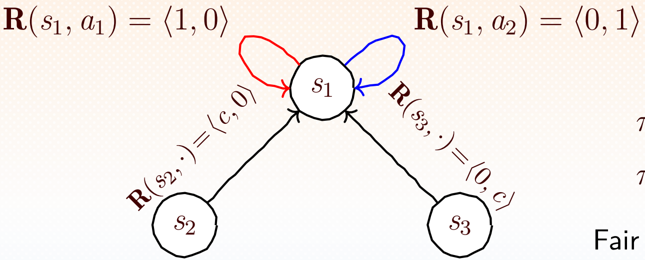
$\pi(s) = \langle \mathbf{1}, 0, 0 \rangle$	$\pi^\star(s) = \langle 0, 0, \mathbf{1} \rangle$	$\pi(s) = \langle 0, \mathbf{1}, 0 \rangle$
$\mathbf{V}^\pi(s) = \langle \frac{1}{1-\gamma}, 0 \rangle$	$\mathbf{V}^{\pi^\star}(s) = \langle \frac{2/3}{1-\gamma}, \frac{2/3}{1-\gamma} \rangle$	$\mathbf{V}^\pi(s) = \langle 0, \frac{1}{1-\gamma} \rangle$

Overcoming Initial Disparity

"Asymmetric Start Bandit" MDP



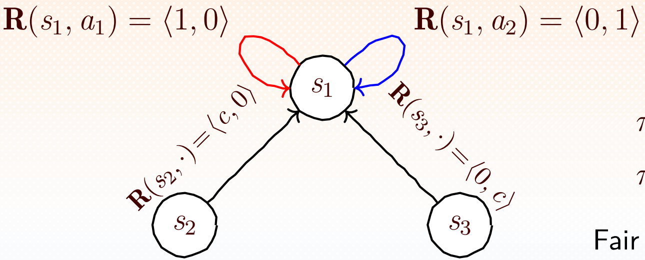
"Asymmetric Start Bandit" MDP



$$\pi^1(s_1) = \langle \mathbf{1}, 0 \rangle$$
$$\pi^2(s_1) = \langle 0, \mathbf{1} \rangle$$
$$\pi^\star(s_1 \text{ from } s_2) = \left\langle \tfrac{1}{2} - \tfrac{1-\gamma}{2\gamma}c, \tfrac{1}{2} + \tfrac{1-\gamma}{2\gamma}c \right\rangle$$
$$\pi^\star(s_1 \text{ from } s_3) = \left\langle \tfrac{1}{2} + \tfrac{1-\gamma}{2\gamma}c, \tfrac{1}{2} - \tfrac{1-\gamma}{2\gamma}c \right\rangle$$

Fair policy $\pi^\star$ is *start-state dependent!*

"Asymmetric Start Bandit" MDP



$$\pi^1(s_1) = \langle \mathbf{1}, 0\rangle$$
$$\pi^2(s_1) = \langle 0, \mathbf{1}\rangle$$
$$\pi^\star(s_1 \text{ from } s_2) = \left\langle \tfrac{1}{2} - \tfrac{1-\gamma}{2\gamma}c, \tfrac{1}{2} + \tfrac{1-\gamma}{2\gamma}c \right\rangle$$
$$\pi^\star(s_1 \text{ from } s_3) = \left\langle \tfrac{1}{2} + \tfrac{1-\gamma}{2\gamma}c, \tfrac{1}{2} - \tfrac{1-\gamma}{2\gamma}c \right\rangle$$

Fair policy $\pi^\star$ is *start-state dependent!*

Lemma (Optimality of Stationary Policies)

For any start state $s_0 \in \mathcal{S}$, there exists some $\mathrm{W}(\cdot)$-optimal policy

$$\pi^\star \in \operatorname*{argmax}_{\pi \in \Pi_{\mathcal{M}}} \mathrm{W}\left(\mathbf{V}_1^\pi(s_0), \ldots, \mathbf{V}_g^\pi(s_0)\right)$$

that is a stationary (Markovian) stochastic policy

On Planning

- **Policy Iteration**
  - ✗ Nonconvergent; can oscillate indefinitely

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)

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*

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

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

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

✓ Approximately optimize $\pi^\star$ with *convex programming*

Regret and Mistakes

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

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

Regret and Mistakes

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

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

- 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)
  - ? **Episodic:** End episode, draw s_{t+1} from start-state distribution

Regret and Mistakes

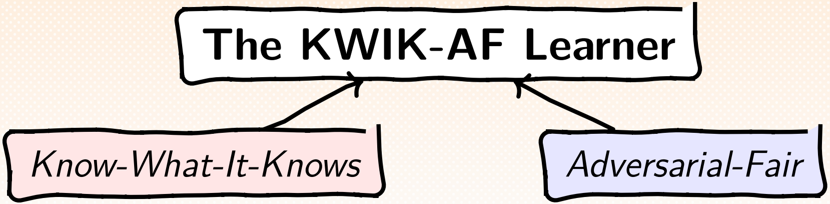
- Optimal policy is *stochastic*, can't assess individual actions

  Assess regret of welfare of *agent policies* $\hat{\pi}_1, \ldots, \hat{\pi}_T$

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

- 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)
  - ? **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



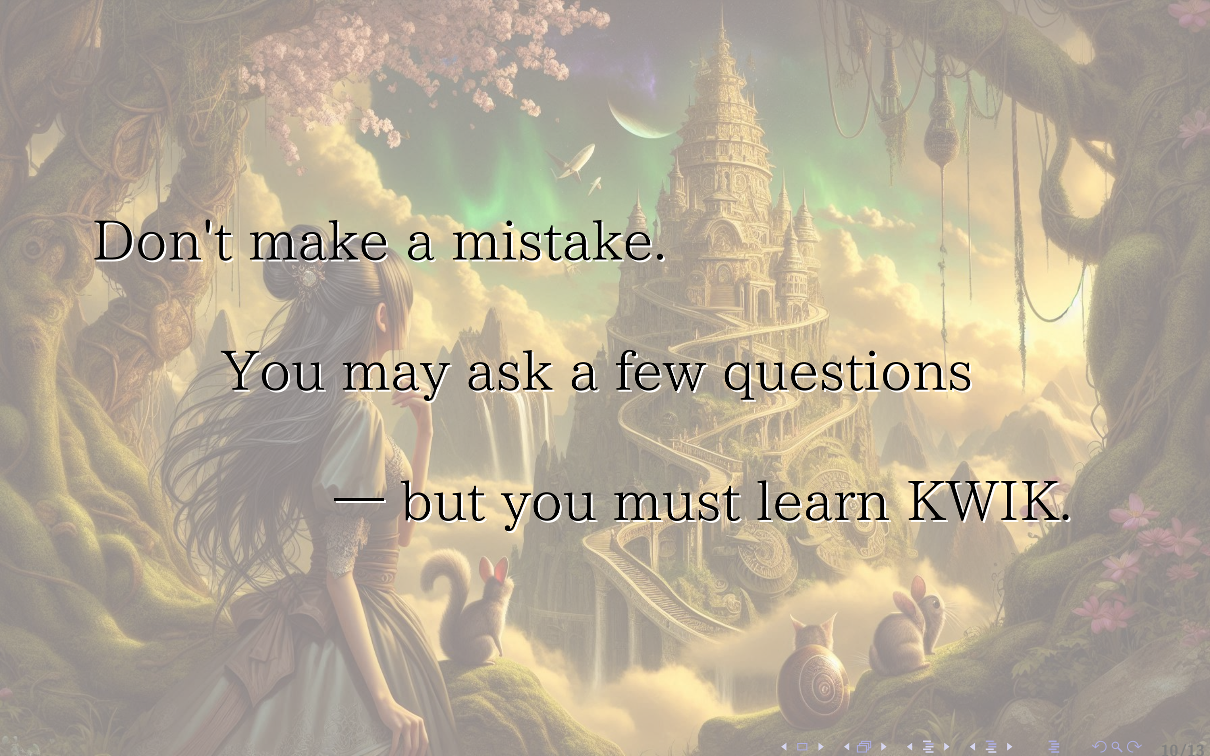
- **KWIK Learner:** At each step, in state s, $\mathbf{A}$ can either
  1. Output an ε-optimal *exploitation policy* π_{xpt}
     - ✗ With probability at least $1-\delta$, *for all time*
     - ✗ *No mistakes*: $\mathrm{W}\left(\mathbf{V}^{\pi_s^\star}(s)\right) - \mathrm{W}\left(\mathbf{V}^{\pi_{\text{xpt}}}(s)\right) > \varepsilon$
  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

Learning Model: KWIK-AF

The KWIK-AF Learner

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

- **KWIK Learner:** At each step, in state s, A can either
  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
- **Adversarial-Fair:** A must be *flexible* and *robust*
  - Optimize *adversarially selected* welfare function $\mathrm{W}_t(\cdot)$ at each step
  - When A outputs a policy π_{xpt}:
    - ⚠ Move A to *adversarial* s', provide *no feedback!*



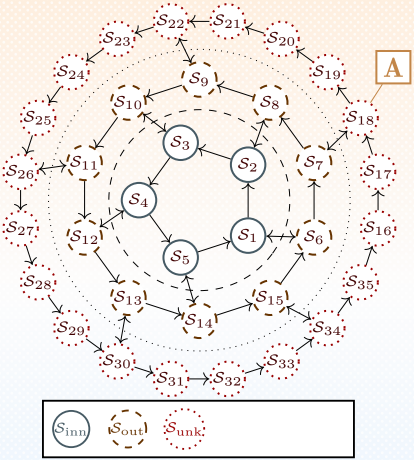
Don't make a mistake.

You may ask a few questions

— but you must learn KWIK.

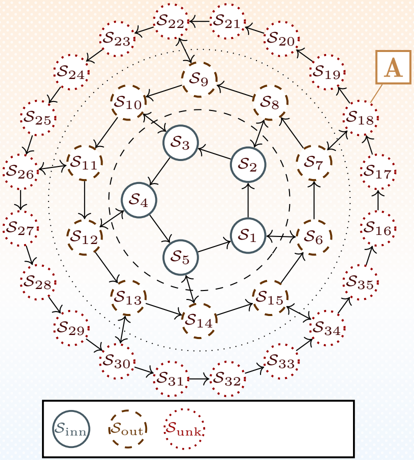
E^4: The Equitable Explicit Explore Exploit Algorithm

- Partition state space into three sets: $\mathcal{S}_{\text{unk}}$, $\mathcal{S}_{\text{out}}$, $\mathcal{S}_{\text{inn}}$
  - **Unknown** $\mathcal{S}_{\text{unk}}$: Insufficient samples ($\leq m_{\text{knw}}$) to estimate reward $\mathbf{R}(s, a)$ and transition $\mathbf{P}(s)$



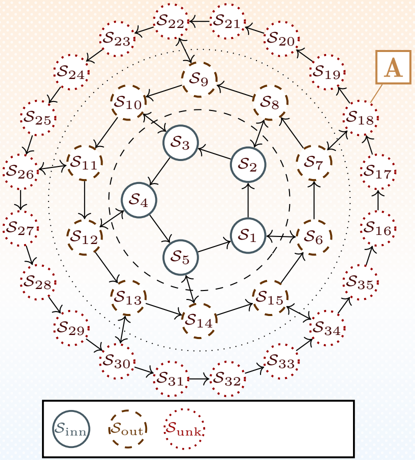
E^4: The Equitable Explicit Explore Exploit Algorithm

- Partition state space into three sets: $\mathcal{S}_{\text{unk}}$, $\mathcal{S}_{\text{out}}$, $\mathcal{S}_{\text{inn}}$
  - **Unknown** $\mathcal{S}_{\text{unk}}$: Insufficient samples ($\leq m_{\text{knw}}$) to estimate reward $\mathbf{R}(s, a)$ and transition $\mathbf{P}(s)$
  - **Outer-Known** $\mathcal{S}_{\text{out}}$: Some *escape policy* π_{esc} can reach $\mathcal{S}_{\text{unk}}$ in T steps with probability at least E



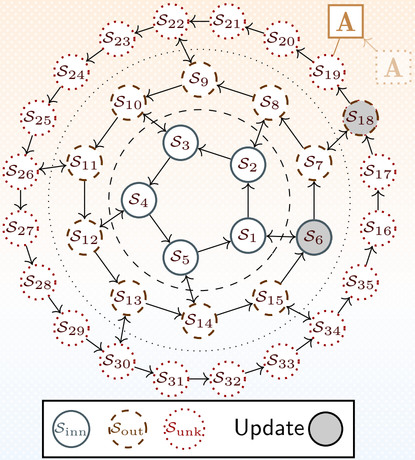
E^4: The Equitable Explicit Explore Exploit Algorithm

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



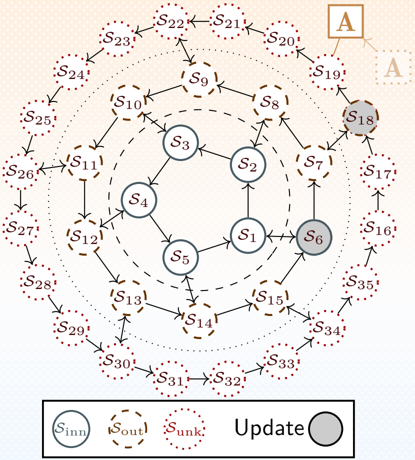
E^4: The Equitable Explicit Explore Exploit Algorithm

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



E^4: The Equitable Explicit Explore Exploit Algorithm

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



∼ The E^4 Algorithm ∼

1. If in $\mathcal{S}_{\text{unk}}$: Explore, observe $(s, a, \boldsymbol{r}, s')$, update empirical MDP $\hat{\mathcal{M}}$, update $\mathcal{S}_{\text{unk}}, \mathcal{S}_{\text{out}}, \mathcal{S}_{\text{inn}}$
2. If in $\mathcal{S}_{\text{out}}$: Begin escape attempt (follow π_{esc} for T steps)

$$\pi_{\text{esc}} \leftarrow \operatorname*{argmax}_{\pi \in \Pi_T} \sum_{s \in \mathcal{S}} \mathop{\mathbb{P}}_{s_{t+1} \sim \hat{\mathbf{P}}(s_t, \pi(s_t, t))} \left(\bigvee_{i=0}^{T} s_i \in \mathcal{S}_{\text{unk}} \,\middle|\, s_0 = s \right)$$

3. Otherwise in $\mathcal{S}_{\text{inn}}$: Output exploit policy $\pi_{\text{xpt}} \leftarrow \operatorname*{argmax}_{\pi \in \Pi_{\mathcal{M}}} \mathrm{W}\left(\hat{\mathbf{V}}^{\pi}(s)\right)$

E^4 Theory

Lemma (Explore-Exploit)

At any point in the execution of E^4, $\mathbf{A}$ can act effectively:

1. *Can exploit from S_{inn}*
2. *Can explore directly from S_{unk}*
3. *Can explore indirectly from S_{out} (escape succeeds with some probability)*

E^4 Theory

Lemma (Explore-Exploit)

At any point in the execution of E^4, $\mathbf{A}$ 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)*

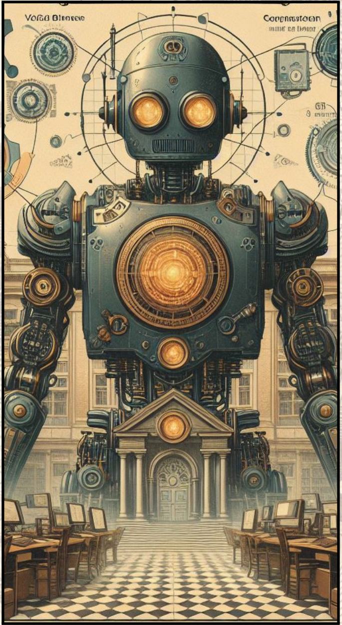
Theorem (E^4 is a KWIK-AF Learner)

E^4 is a KWIK-AF learner w.r.t. the class of all λ-$\|\cdot\|_\infty$ Lipschitz-continuous welfare functions, with exploration budget

$$
\begin{aligned}
\mathrm{m}\left(|\mathcal{S}|, |\mathcal{A}|, \gamma, \mathrm{R}_{\max}, g, \varepsilon, \delta\right) &\in \mathbf{O}\left(|\mathcal{S}|^2 |\mathcal{A}| \left(\frac{\lambda \mathrm{R}_{\max}}{\varepsilon(1-\gamma)} \log_{\frac{1}{\gamma}}\left(\frac{\lambda \mathrm{R}_{\max}}{\varepsilon(1-\gamma)}\right)\right)^3 \log \frac{|\mathcal{S}||\mathcal{A}| g}{\delta}\right) \\
&\subseteq \mathrm{Poly}\left(|\mathcal{S}|, |\mathcal{A}|, \frac{1}{1-\gamma}, \mathrm{R}_{\max}, \log g, \frac{1}{\varepsilon}, \log \frac{1}{\delta}, \lambda\right)
\end{aligned}
$$

In Summary

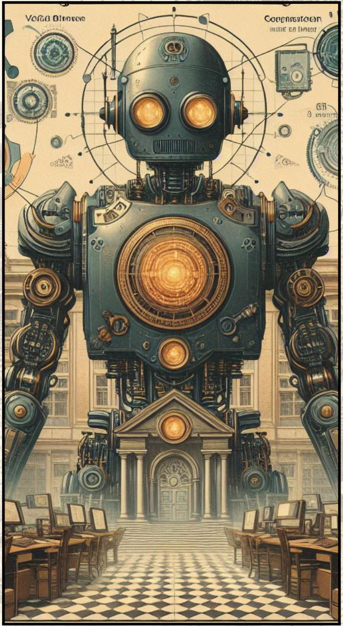
- From Egocentric to Altruistic Agents
  - Agent $\mathbf{A}$ acts in , impacting *beneficiaries*
  - Vector-valued (per-beneficiary) reward $\mathbf{R}(s, a)$
  - Social planner's problem:
    - Optimize *welfare* of *value functions* $\operatorname*{argmax}_{\pi \in \Pi} \mathrm{W}\big(\mathbf{V}^{\pi}(s)\big)$



In Summary

- From Egocentric to Altruistic Agents
  - Agent $\mathbf{A}$ acts in , impacting *beneficiaries*
  - Vector-valued (per-beneficiary) reward $\mathbf{R}(s, a)$
  - Social planner's problem:
    - Optimize *welfare* of *value functions* $\operatorname*{argmax}_{\pi \in \Pi} \mathrm{W}\left(\mathbf{V}^{\pi}(s)\right)$

- KWIK-AF: A Model of Fair RL
  - Adversarial flexibility
  - Societal welfare objectives
  - No mistakes from bounded exploration



In Summary

- From Egocentric to Altruistic Agents
  - Agent $\mathbf{A}$ acts in , impacting *beneficiaries*
  - Vector-valued (per-beneficiary) reward $\mathbf{R}(s, a)$
  - Social planner's problem:
    - Optimize *welfare* of *value functions* $\operatorname*{argmax}_{\pi \in \Pi} \mathrm{W}\big(\mathbf{V}^{\pi}(s)\big)$
- KWIK-AF: A Model of Fair RL
  - Adversarial flexibility
  - Societal welfare objectives
  - No mistakes from bounded exploration
- Efficient Learning and Planning
  - KWIK-AF learn with E^4
  - Plan with convex programming on state-action measure
  - Polynomial exploration budget, time complexity