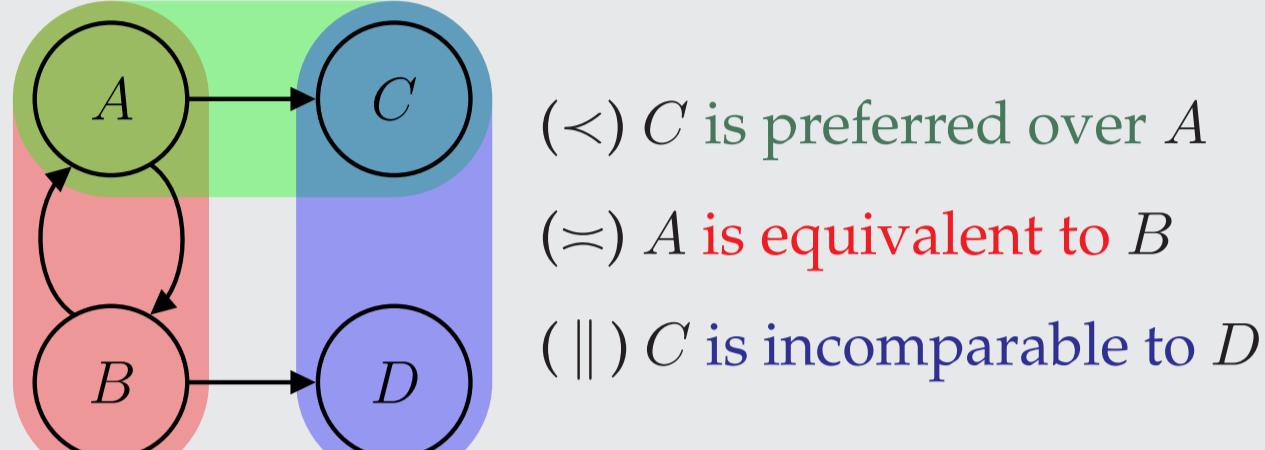



**SETTING**
**PREFERENCE**

$$\text{PrefMDP} = (\underbrace{\mathcal{S}, \mathcal{A}, H, p, \mu}_{\text{MDP}\setminus R}, \leq_{\mathcal{T}})$$

$\leq_{\mathcal{T}} \subseteq \mathcal{T} \times \mathcal{T}$  is a **partial (pre)order** over the trajectory space  $\mathcal{T}$



**Assumption:** Human expresses preferences based on an underlying (unknown) utility

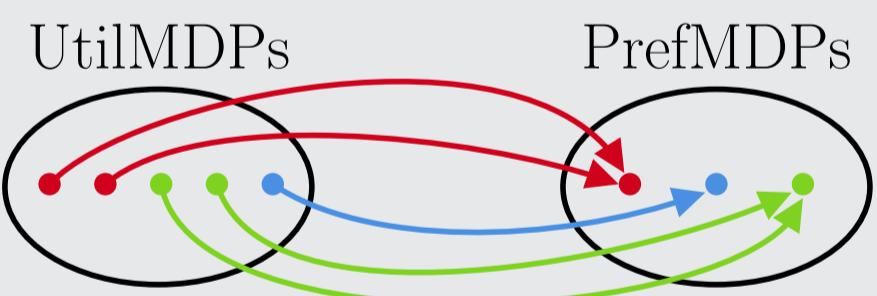
$$\text{UtilMDP} = (\underbrace{\mathcal{S}, \mathcal{A}, H, p, \mu, u}_{\text{MDP}\setminus R})$$

$u : \mathcal{T} \rightarrow \mathbb{R}^m$  is a **multi-dimensional utility**

Expected utility of a policy  $\pi \in \Pi$ :

$$J(\pi; u) := \sum_{\tau \in \mathcal{T}} d_\pi(\tau) u(\tau) = \langle d_\pi, u \rangle,$$

where  $d_\pi$  is the distribution over trajectories induced by  $\pi$



**Assumption:** Human expresses preferences based on an (unknown) Markovian reward

$$\text{MDP} = (\underbrace{\mathcal{S}, \mathcal{A}, H, p, \mu, r}_{\text{MDP}\setminus R})$$

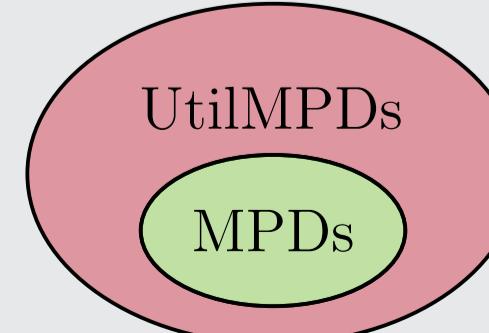
$r := (r_h)_{h \in \llbracket H \rrbracket}$  is a **stage-dependent multi-dimensional reward function**

Trajectory return:

$$u_r(\tau) := \sum_{h=1}^H r_h(s_h, a_h)$$

Expected policy return:

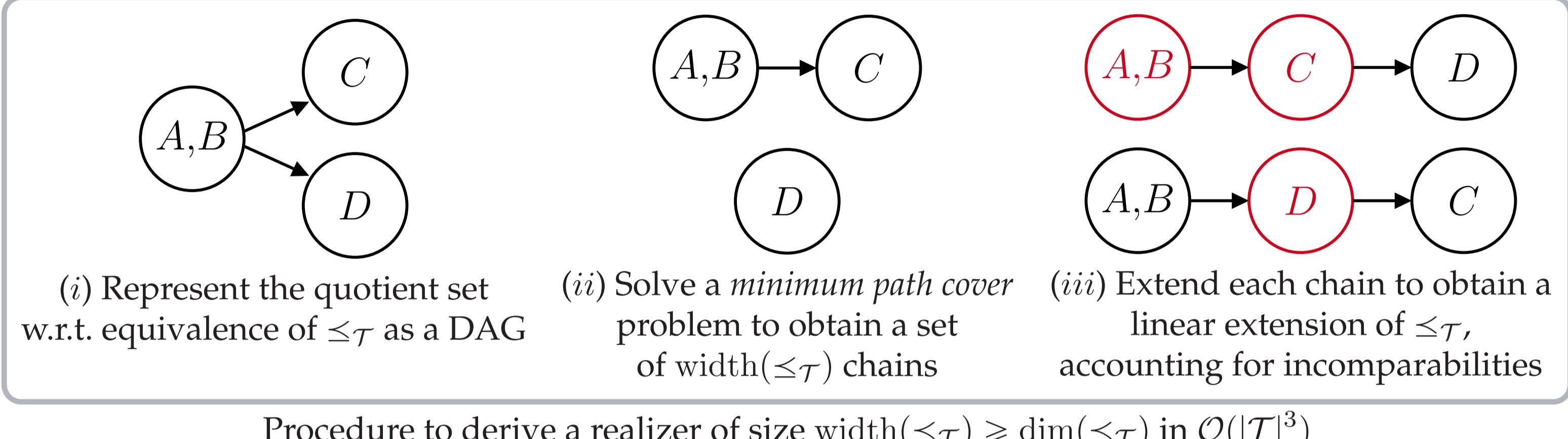
$$J(\pi; r) := J(\pi; u_r)$$


**UTILITY**
**Reward**
**COMPATIBLE UTILITIES**
**COMPATIBLE UTILITY**

$u$  is **compatible** with  $\leq_{\mathcal{T}}$  if  $\forall \tau, \tau' \in \mathcal{T}: \tau \leq_{\mathcal{T}} \tau' \Rightarrow u(\tau) \leq u(\tau')$  (element-wise)

Realizer Dimension	Exists?	Computational Complexity
$< \dim(\leq_{\mathcal{T}})$	✗	—
$= \dim(\leq_{\mathcal{T}})$	✓	NP-hard
$> \dim(\leq_{\mathcal{T}})$	✓	$\text{Poly}( \mathcal{T} )$

where  $\dim(\cdot)$  is the order dimension


**POLICY DOMINANCE**
**POLICY DOMINANCE**

Policy  $\pi \leq_{\mathcal{T}}$ -strictly dominates policy  $\pi'$  ( $\pi' <_{\Pi} \pi$ ) if it yields a strictly better expected utility

$$J(\pi; u) - J(\pi'; u) > 0 \quad (\text{element-wise})$$

for every compatible utility function  $u$

 **$\leq_{\mathcal{T}}$ -PARETO OPTIMALITY**

Set of  $\leq_{\mathcal{T}}$ -Pareto optimal policies:

$$\Pi^*(\leq_{\mathcal{T}}) := \{\pi \in \Pi : \nexists \pi' \in \Pi \text{ s.t. } \pi <_{\Pi} \pi'\}$$

**HOW TO EVALUATE?**

$\pi' \leq_{\Pi} \pi$  can be verified by evaluating if:

$$\forall n \in \llbracket |\mathcal{T}| \rrbracket : \sum_{i=1}^n (d_\pi(i) - d_{\pi'}(i)) \geq 0$$

holds for every linear extension of  $\leq_{\mathcal{T}}$ , where index  $i$  represents the  $i$ -th trajectory sorted w.r.t. the linear extension

**OPEN QUESTION**

Is there an efficient evaluation method?

Trivial solution requires the evaluation of  $\mathcal{O}(|\mathcal{T}|!)$  linear extensions of  $\leq_{\mathcal{T}}$

We conjecture this to be computationally hard

**APPROXIMATION VIA MARKOVIAN REWARDS**
**WHY THE NEED FOR APPROXIMATION?**
**Preferences**

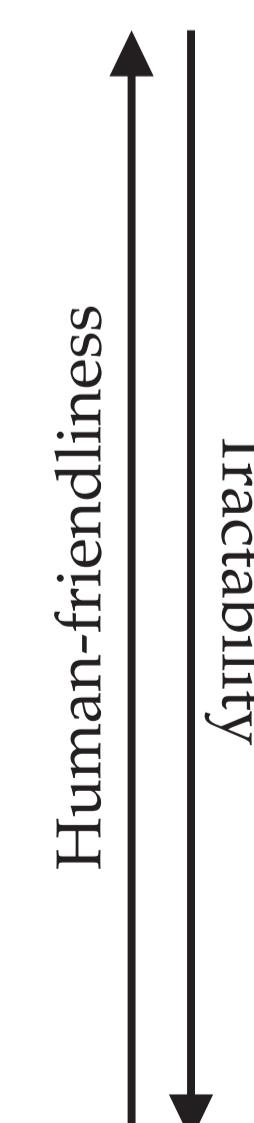
- + Most general type of feedback
- Intractable without introducing a structure

**Utilities**

- + Assign numerical signals to each trajectory
- Complexity of learning is exponential in  $H$

**Rewards**

- + Enable efficient learning
- Less representational power


**(CONVEX) QUADRATIC PROGRAM**

**Goal:** Find  $u$  and  $r$  that best represent  $\leq_{\mathcal{T}}$

**Input:** A realizer  $\{\leq_{\mathcal{T}, i}\}_{i \in \llbracket m \rrbracket}$  of  $\leq_{\mathcal{T}}$  of size  $m$

**Idea:** Jointly choose  $u$  compatible with the realizer and  $r$  as Markovian approximation of  $u$

**Define:**  $B \in \{0, 1\}^{|\mathcal{T}| \times |\mathcal{S}| \times |\mathcal{A}|^H}$  (binary encoding of  $\mathcal{T}$ ) and  $A := \mathbf{I}_{|\mathcal{T}|} - B(B^\top B)^{-1}B^\top$  (OLS)

$$\eta^* := \min \|Au\|_F^2$$

$$\text{s.t. } u_j(i+1) \leq u_j(i) - \varepsilon \quad \forall i \in \llbracket |\mathcal{T}| - 1 \rrbracket, j \in \llbracket m \rrbracket$$

If  $\eta^* = 0 \rightarrow u_r = u \rightarrow$  Preferences derive from a Markovian reward function

If  $\eta^* > 0 \rightarrow$  Preferences cannot be expressed via Markovian rewards

→ Using  $r$ , we learn how to solve a simpler surrogate problem

→ Such an approximation introduces a suboptimality in terms of the performance of the optimal policy bounded by  $2\sqrt{mn\eta^*}$