Controllability in preference-conditioned multi-objective reinforcement learning

Pau de las Heras Molins, Beyazit Yalcinkaya,
Lasse Peters, David Fridovich-Keil, Georgios Bakirtzis

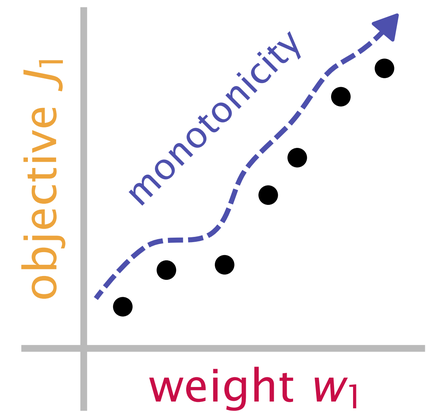
Multi-objective reinforcement learning (MORL) controls agents via preference

Preference

is commonly encoded using linear weight vectors
- weight vector $\mathbf{w} \in \mathbb{R}^D$ — bounded to the simplex

represents the desired trade-off between the D objectives
- e.g. $\mathbf{w} = [0.8, 0.2]$ — the first objective is much more important

conditions policies to influence agent behavior
- policy $\pi(s)$ becomes $\pi(s; \mathbf{w})$, inducing vector returns $\mathbf{v}^{\pi}_{\mathbf{w}}$



Changes in preference should change the agent's behavior in the intended way (i.e., returns; e.g., monotonically with weights) $\Longrightarrow$

We show that standard metric cannot reliably capture this property — controllabilty

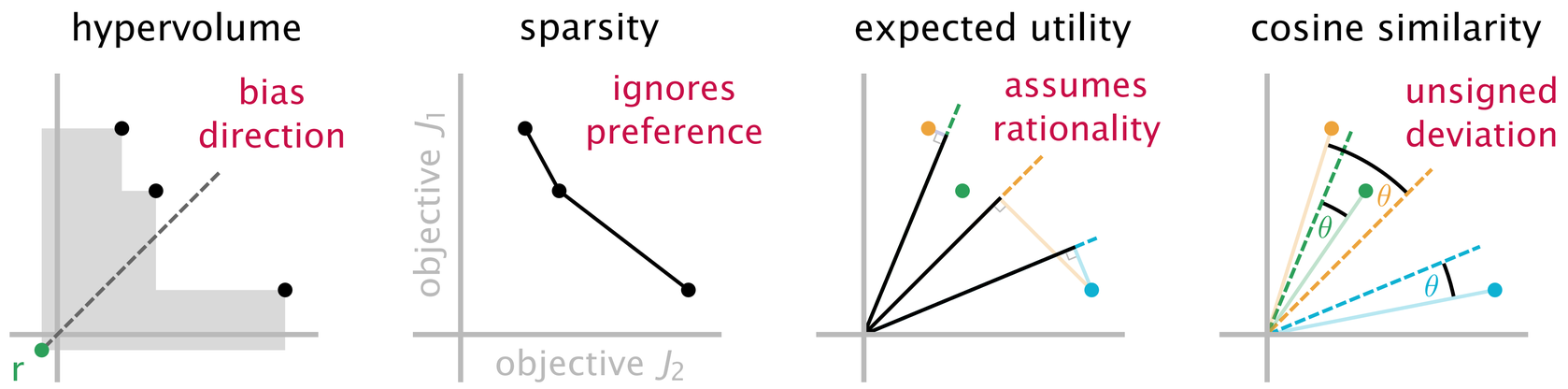
The mapping $\mathbf{w} \mapsto \mathbf{v}^{\pi}_{\mathbf{w}}$ should be

predictable
- it exhibits structured dependence on $\mathbf{w}$ (e.g., monotonicity)

distinguishable
- different preferences induce measurably different returns

Standard MORL metrics are insufficient to evaluate preference-adaptable agents $\Longrightarrow$

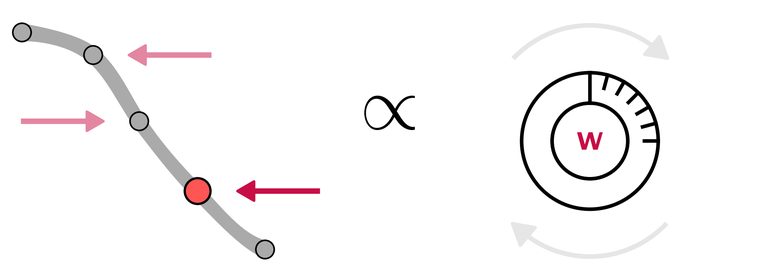
They ignore preference, make strong assumptions or provide limited insight on controllability



Our experiments show that an agent can score well in MORL metrics despite ignoring preferences

Preference provides a symbolic interface between user intent and agent behavior $\Longrightarrow$

But only if we can reliably assess our ability to control agents



We propose a complementary controllability metric for linear preference $\Longrightarrow$

Rank correlation quantifies monotonicity of per-objective returns with weight