



POLITECNICO
MILANO 1863



Reinforcement Learning Beyond Scalar Rewards

Alberto Maria Metelli

AIMLSystems 2026
Lecco, October 9, 2026

Outline

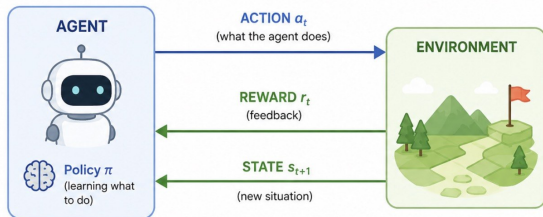
- ➊ **Introduction** – scalar rewards, MDPs, bandits
- ➋ **Preferences** – ordinal structure, utility and link functions
- ➌ **Comparisons** – indifference, incomparability, MOBT
- ➍ **Case study** – GridWorld and identification
- ➎ **RLHF** – scalar vs multi-objective feedback models

Introduction

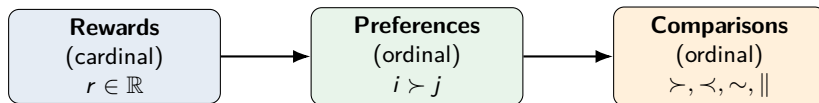
From cardinal reward to preference feedback

Message of the talk

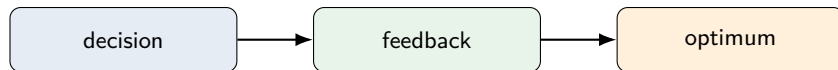
- Reinforcement learning (Sutton and Barto, 2018) assumes that feedback is a **number**: a reward
- Human feedback is often **relative**: “A is better than B”, “these are tied”, or “I cannot compare them”
- Research question: ***How can we extend reinforcement learning to non-numerical feedback***
- We start with **MDPs** and use **bandits** to isolate the role of feedback (Lattimore and Szepesvári, 2020)



From rewards to human feedback



Taxonomy for Human Feedback



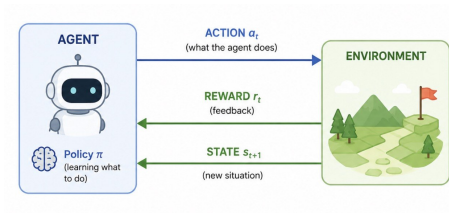
- **Decision:** one action, a pair of arms, two trajectories, a policy, ...
- **Feedback:** a real-valued reward, an ordinal preference, a four-way comparison, a critique, a veto, ...
- **Optimum:** highest expected reward, Condorcet winner, maximal/Pareto set, feasible policies, ...

The basic reinforcement-learning model

- A discounted MDP is defined by $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$ (Sutton and Barto, 2018)
- At time t , the agent observes state S_t , chooses action A_t , then the environment generates

$$S_{t+1} \sim P(\cdot \mid S_t, A_t), \quad r_t = r(S_t, A_t)$$

- The **policy** $\pi(a \mid s)$ maps states to action distributions
- Rewards provide the *numerical* signal of desirability



The goal is to maximize cumulative reward

- A trajectory is $\tau = (s_0, a_0, \dots, s_H)$
- Its discounted **return** is (Sutton and Barto, 2018)

$$G(\tau) = \sum_{t=0}^{H-1} \gamma^t r(s_t, a_t), \quad 0 \leq \gamma \leq 1$$

- The **optimal policy** maximizes expected return:

$$\pi^* \in \arg \max_{\pi} J(\pi), \quad J(\pi) = \mathbb{E}_{\tau \sim \pi}[G(\tau)]$$

The Reward Hypothesis

Sutton's Reward Hypothesis (paraphrased)

All goals and purposes can be represented as maximizing the expected value of the cumulative sum of a received scalar signal

- It is the conceptual foundation of standard RL: $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ (Bowling et al., 2023; Sutton and Barto, 2018)
- But the existence and accessibility of such a representation is a **modeling assumption**

Can every human objective be expressed by *one number*? Can the human communicate it?

The simplest example: bandit feedback

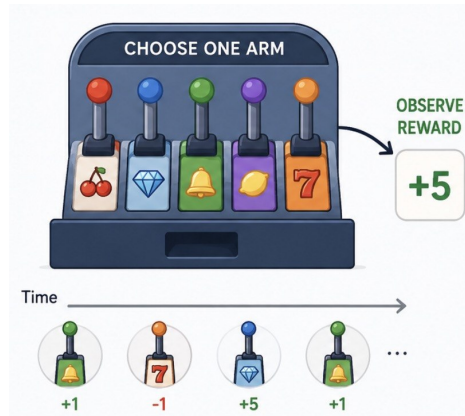
- Choose an arm $I_t \in [K]$, observe only its *numeric* reward $X_{I_t,t}$ (Lattimore and Szepesvári, 2020)
- Each arm i has an unknown expected reward

$$u_i = \mathbb{E}[X_i]$$

- **Definition of the optimum:**

$$i^* \in \arg \max_{i \in [K]} u_i$$

- The agent learns *which number is largest* from noisy numeric observations



Preferences

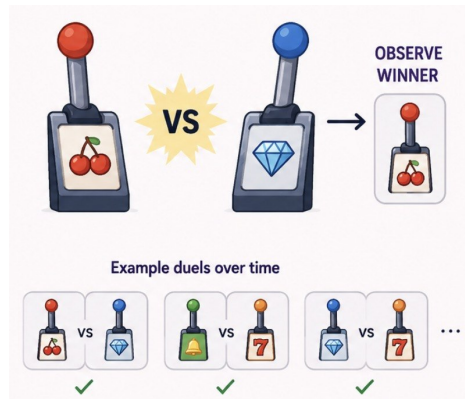
Relations, utility representations, and stochastic choice

Dueling bandits

- Choose a **pair** of arms $I_t, J_t \in [K]$
- Environment reveals a preference (Yue et al., 2009; Busa-Fekete and Hüllermeier, 2014)

$$Y_{I_t, J_t, t} \in \{\succ, \prec\}$$

- The feedback is binary
 - ▶ $\succ$ means that I_t is preferred over J_t ($I_t \succ J_t$)
 - ▶ $\prec$ means that J_t is preferred over I_t ($J_t \succ I_t$)
- Often represented as the outcome of Bernoulli sampling: 1 for $\succ$ and 0 for $\prec$
- No scalar reward is directly observed
- Without a scalar reward, what should count as an **optimum**?



Arbitrary preferences

- Let P_{ij} denote the probability that alternative i wins a duel against j (Yue et al., 2009; Busa-Fekete and Hüllermeier, 2014):

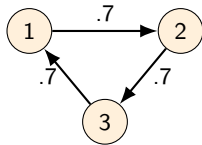
$$\mathbf{P} \in [0, 1]^{K \times K}, \quad P_{ij} = \mathbb{P}(i \succ j), \quad P_{ij} + P_{ji} = 1$$

- The induced majority relation is

$$i \succ_{\mathbf{P}} j \iff P_{ij} > \frac{1}{2}$$

- This relation need not be transitive:

$$P_{12} = 0.7, \quad P_{23} = 0.7, \quad P_{31} = 0.7$$



Arbitrary preferences: optimality

- A **Condorcet winner** beats every other alternative (Yue et al., 2009; Busa-Fekete and Hüllermeier, 2014):

$$P_{i^*j} > \frac{1}{2} \quad \forall j \neq i^*$$

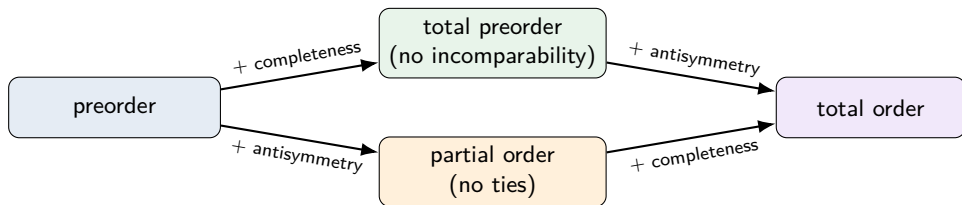
- Alternative solution concepts:

$$s_i^{\text{Borda}} = \frac{1}{K-1} \sum_{j \neq i} P_{ij},$$

$$s_i^{\text{Copeland}} = \sum_{j \neq i} \mathbb{1}\{P_{ij} > 1/2\}$$

Preference relations

- A relation $\succeq$ on $\mathcal{X}$ is a **preorder** if it is reflexive and transitive (Ok, 2002)
- A preorder is **complete** if $x \succeq y$ or $y \succeq x$ for all x, y
- A preorder is **antisymmetric** if $x \succeq y$ and $y \succeq x$ imply $x = y$



From preferences to utilities

Scalar representation theorem – finite case (Debreu, 1954)

For a finite set $\mathcal{X}$, a relation $\succeq$ admits a scalar utility representation if and only if it is a total preorder

$$\exists u : \mathcal{X} \rightarrow \mathbb{R} : \quad x \succeq y \iff u(x) \geq u(y)$$

- Utility **represents the preference order**
- The representation is **ordinal**: many numerical functions induce the same relation

From preferences to utilities

- The two numerical representations below induce *exactly the same preferences*

	A	B	C	$B - A$	$C - B$
u	0	1	2	1	1
v	0	1	100	1	99

Ordinal invariance

If f is strictly increasing, then

$$u(x) \geq u(y) \iff f(u(x)) \geq f(u(y))$$

The preference order is identified, but cardinal gaps are not

Utilitie for MDPs

- A representation theorem might give a utility for *entire behaviors* (Drago et al., 2026b; 2025):

$$u : \mathcal{T} \rightarrow \mathbb{R}$$

- A Markovian additive reward demands a stronger factorization:

$$u(\tau) \stackrel{?}{=} \sum_{t=0}^{H-1} \gamma^t r(s_t, a_t)$$

- If not, a Markovian reward can still be an **approximation**

Random Utility Model

- Assume that each arm has a latent unobserved scalar **utility** $u_i \in \mathbb{R}$ (Bradley and Terry, 1952; Thurstone, 1927; Luce, 1959)
- Consider a noise distribution Ψ
- Pairwise feedback is generated as follows
 - ▶ Sample ϵ_i and ϵ_j independently from Ψ with CDF ψ
 - ▶ Define the noisy utilities $U_i = u_i + \epsilon_i$ and $U_j = u_j + \epsilon_j$
 - ▶ Generate feedback as

$$Y_{i,j} = \mathbb{1}\{U_i > U_j\}$$

- ▶ The feedback $Y_{i,j} \sim \text{Ber}(P_{ij})$ is Bernoulli distributed with parameter

$$P_{ij} = \mathbb{P}(Y_{i,j} = 1) = \psi(u_i - u_j)$$

Random Utility Model

- Thus, we can relate win probabilities to utilities

$$u_i > u_j \iff P_{ij} = \psi(u_i - u_j) > 1/2$$

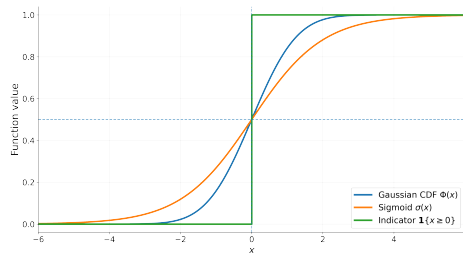
- Clear definition of the optimum, consistent with standard rewards

$$i^* \in \arg \max_{i \in [K]} u_i$$

- The induced relation $\succ_{\mathbf{P}}$ is a total preorder with no incomparabilities

Link functions

- One may also directly define a link function $\psi : \mathbb{R} \rightarrow [0, 1]$ (Bradley and Terry, 1952; Thurstone, 1927; Drago et al., 2025)
 - ▶ **monotonicity**: $x < y \implies \psi(x) \leq \psi(y)$
 - ▶ **symmetry**: $\psi(x) + \psi(-x) = 1$
- This way, ψ is the CDF of a symmetric zero-mean distribution



Model	Link	Invertible	Random-utility model
Bradley–Terry	$\psi_{\text{BT}}(x) = \frac{1}{1+e^{-x}}$	Yes	Gumbel noise
Thurstone / probit	$\psi_{\text{Th}}(x) = \Phi(x/s)$	Yes	Gaussian noise
Hard threshold	$\psi(x) = \mathbb{1}\{x > 0\}$	No	deterministic

What can be identified?

- ❶ If ψ is strictly increasing with $\psi(0) = 1/2$ (Bradley and Terry, 1952; Negahban et al., 2017):

$$P_{ij} > 1/2 \iff u_i > u_j \quad \Rightarrow \quad \text{order}$$

- ❷ If ψ is **known and invertible**:

$$u_i - u_j = \psi^{-1}(P_{ij}) \quad \Rightarrow \quad \text{utility gaps}$$

Comparisons

Indifference, incomparability, and vector utilities

Binary feedback forces a decision

- Standard pairwise preferences have two outcomes (Davidson, 1970; Eliaz and Ok, 2006; Drago et al., 2026a):

$$\mathcal{O}_{\text{binary}} = \{\succ, \prec\}$$

- But a more expressive comparison interface supports *four* outcomes:

$$\mathcal{O} = \{\succ, \prec, \asymp, \parallel\}$$

Symbol	Meaning	Example
$x \succ y$	preferred	"A is better"
$x \prec y$	inverse preference	"B is better"
$x \asymp y$	indifference	"They are equally good"
$x \parallel y$	incomparability	"I cannot decide"

Indifference and incomparability are different

Indifference ($x \asymp y$)

$$u(x) = u(y)$$

- Alternatives are **equivalent** under the considered criterion (Eliaz and Ok, 2006; Ok, 2002)
- Can be represented by a scalar utility
- Example: two summaries with the same quality

Incomparability ($x \parallel y$)

$$\neg(x \succeq y) \wedge \neg(y \succeq x)$$

- The relation does not rank the alternatives
- Can express conflicting, non-comparable objectives
- Example: faster but less safe versus safer but slower

Not choosing can be a rational and informative response

Scalar utility cannot represent incomparability

- Any scalar mapping $u : \mathcal{X} \rightarrow \mathbb{R}$ induces a **total** preorder (Ok, 2002; Roijers et al., 2013):

$$\forall x, y \quad u(x) \geq u(y) \vee u(y) \geq u(x)$$

- Hence no scalar utility can represent a relation with $x \parallel y$
- A natural alternative is a **vector utility**:

$$\mathbf{u}(x) = (u_1(x), \dots, u_d(x)) \in \mathbb{R}^d$$

- It yields a **Pareto relation**:

$$x \succeq_P y \iff u_k(x) \geq u_k(y) \quad \forall k \in [d]$$

Incomparability can be a manifestation of a multi-objective problem

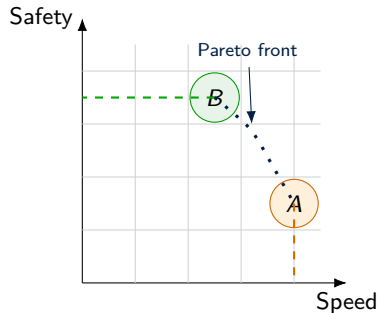
A two-dimensional example

- Policy A : faster but less safe (Roijers et al., 2013)
- Policy B : slower but safer

$$\mathbf{u}(A) = (8, 3), \quad \mathbf{u}(B) = (5, 7)$$

- Neither dominates:

$$A \parallel B$$



Finite partial orders admit vector representations

Vector utility representation (finite domain) (Ok, 2002)

For a finite preorder, after identifying equivalent alternatives, there exists a finite d and a vector utility $\mathbf{u} : \mathcal{X} \rightarrow \mathbb{R}^d$ satisfying

$$x \succeq y \iff \mathbf{u}(x) \geq \mathbf{u}(y) \text{ componentwise}$$

- Thus incomparability corresponds to *disagreement across objectives* (Ok, 2002; Drago et al., 2025)
- The smallest possible d relates to the **order dimension** of the quotient order
- Computing the smallest representation dimension is NP-hard (Yannakakis, 1982)

Comparison feedback: four observations

- Let $o \in \mathcal{O} = \{\succ, \prec, \asymp, \parallel\}$ (Drago et al., 2026a)
- The learner observes a random comparison for a pair of trajectories:

$$Q_{\tau, \tau'}^o = \mathbb{P}(\tau \circ \tau'), \quad \sum_{o \in \mathcal{O}} Q_{\tau, \tau'}^o = 1.$$

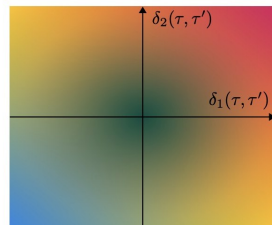
- Under a vector-utility model, comparison probabilities depend on the **vector gap**

$$\delta(\tau, \tau') = \mathbf{u}(\tau) - \mathbf{u}(\tau')$$

- We seek $f(o \mid \delta)$ that distinguishes equivalence (near-zero gap) from conflict (opposing signs)

Four outcomes from vector utility gaps

- Preference: $\delta \in \mathbb{R}_+^d \setminus \{0\}$ (Drago et al., 2026a)
- Inverse preference: $-\delta \in \mathbb{R}_+^d \setminus \{0\}$
- Indifference: $\delta \approx \mathbf{0}$
- Incomparability: δ has **mixed signs**



- | | |
|--|---|
| ■ $>$ (Direct Preference) | ■ $\approx$ (Indifference) |
| ■ $<$ (Inverse Preference) | ■ $\parallel$ (Incomparability) |

A multi-objective Bradley–Terry model: intuition

- Extend BT to **four** outputs via a softmax (Drago et al., 2026a):

$$f(o \mid \delta) = \frac{e^{h_o(\delta)}}{\sum_{o' \in \mathcal{O}} e^{h_{o'}(\delta)}}$$

- The scores can encode *direction*, *equality* and *conflict*:

$$\begin{aligned} h_{\succ}(\delta) &= \tfrac{1}{2} \mathbf{1}^\top \delta, & h_{\prec}(\delta) &= -\tfrac{1}{2} \mathbf{1}^\top \delta, \\ h_{\asymp}(\delta) &= \alpha, & h_{\parallel}(\delta) &= \beta + \gamma \sqrt{d} \operatorname{std}(\delta) \end{aligned}$$

- The **average direction** determines preference; the **dispersion** across objectives drives incomparability

MOBT: what can comparisons identify?

Objectives	Identifiable information	Remaining ambiguity
$d = 1$	Scalar utility differences	Shared additive offset
$d = 2$	Two objectives, up to relabeling	Permutation and additive offsets
$d \geq 3$	Aggregate utility and magnitude of conflict	Continuum of rotations/reflections mixing individual objectives

The **distribution of comparisons** can be identified without identifying the **individual reward coordinates** (Drago et al., 2026b)

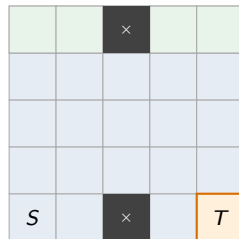
Case study

Multi-objective comparisons in a GridWorld

Case study: the paper's 5×5 GridWorld

- **Start:** lower-left cell; deterministic moves in four directions or *stay*
- **Target:** bottom-right; episodes continue after reaching it
- **Three objectives:**
 - 1 stay at target (bottom-right),
 - 2 reach upper rows,
 - 3 avoid obstacle-adjacent cells
- Two-objective variant: remove obstacles and objective 3

top-row objective



GridWorld: rewards, trajectories, and comparison data

Cell-based reward $r(s)$

- ❶ **Target:** +1 in the bottom-right cell, -0.1 elsewhere
- ❷ **Upper rows:** -0.1 times the row index from the top
- ❸ **Obstacle safety:** -0.5 adjacent to an obstacle, 0 elsewhere

Synthetic preference experiment

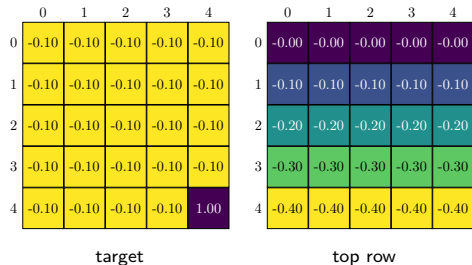
- 2,000 trajectories of length $H = 20$;
- 5,000 sampled pairs;
- Expert outcomes drawn from **MOBT**;
- One-hot cell features and

$$\mathbf{u}(\tau) = M \phi(\tau), \quad \phi(\tau) \in \mathbb{R}^{25}$$

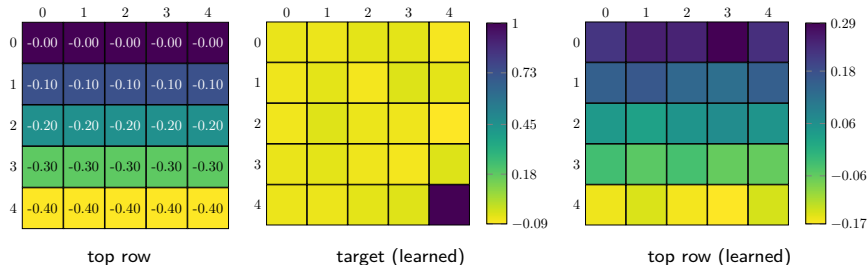
- In $d = 2$, obstacles and objective 3 are removed; the target is not terminal (Drago et al., 2026a)

GridWorld ($d = 2$): true and recovered objectives

Ground-truth rewards



Estimated rewards



- The recovered maps preserve the two objectives (Drago et al., 2026a)

GridWorld ($d = 3$): the true rewards

	0	1	2	3	4
0	-0.10	-0.10		-0.10	-0.10
1	-0.10	-0.10	-0.10	-0.10	-0.10
2	-0.10	-0.10	-0.10	-0.10	-0.10
3	-0.10	-0.10	-0.10	-0.10	-0.10
4	-0.10	-0.10		-0.10	1.00

Reach target

	0	1	2	3	4
0	-0.00	-0.00		-0.00	-0.00
1	-0.10	-0.10	-0.10	-0.10	-0.10
2	-0.20	-0.20	-0.20	-0.20	-0.20
3	-0.30	-0.30	-0.30	-0.30	-0.30
4	-0.40	-0.40		-0.40	-0.40

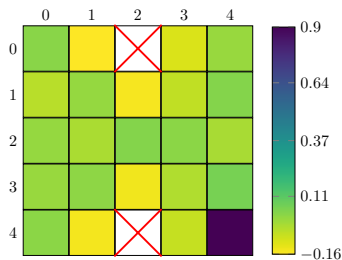
Reach top row

	0	1	2	3	4
0	0.00	-0.50		-0.50	0.00
1	0.00	0.00	-0.50	0.00	0.00
2	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	-0.50	0.00	0.00
4	0.00	-0.50		-0.50	0.00

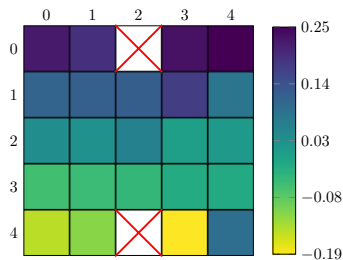
Avoid obstacles

- Third objective penalizes adjacency to obstacles; objectives can conflict across trajectories
- Each visited cell generates a reward vector $\mathbf{r}(s) \in \mathbb{R}^3$ (Drago et al., 2026a)

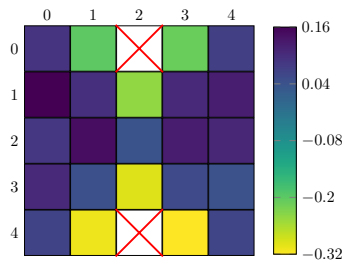
GridWorld ($d = 3$): the recovered rewards



Learned coordinate 1



Learned coordinate 2

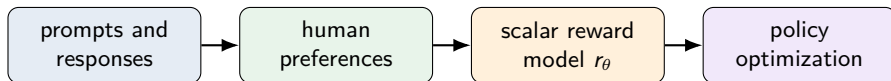


Learned coordinate 3

- We see **cross-objective leakage**: obstacle penalties also appear in other learned coordinates
- Lack of identifiability (Drago et al., 2026a)

RLHF

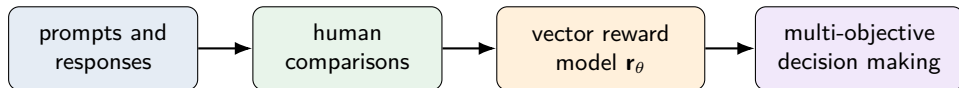
Scalar and multi-objective feedback models



- Standard reward modeling assumes a **scalar** score for each response (Christiano et al., 2017; Ouyang et al., 2022):

$$P_\theta(y^+ \succ y^- \mid x) = \sigma\left(r_\theta(x, y^+) - r_\theta(x, y^-)\right)$$

Multi-objective RLHF

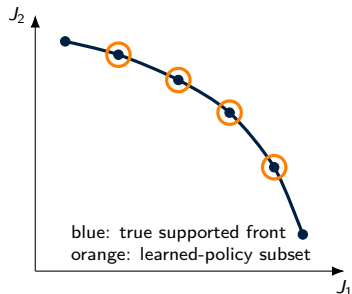


- Instead of a single score, learn a vector of criteria (Roijsers et al., 2013; Drago et al., 2026a; Ouyang et al., 2022)

$$\mathbf{r}_\theta(x, y) = (r_{\text{help}}, r_{\text{safety}}, r_{\text{truth}}, r_{\text{privacy}}) \in \mathbb{R}^d$$

- Feedback may distinguish **preference, inverse preference, indifference, incomparability**
- The decision layer can select among Pareto-efficient policies using application-dependent trade-offs

Reconstructing the Pareto frontier from comparisons



- For weights $\mathbf{w} \in \Delta_{d-1}$, optimize a scalarized return (Drago et al., 2026a; Roijers et al., 2013)

$$\pi_{\mathbf{w}} \in \arg \max_{\pi} \mathbf{w}^{\top} \mathbf{J}(\pi)$$

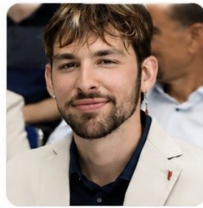
- $d = 2$: learned scalarizations recover *true Pareto-optimal policies*
- $d \geq 3$: exact guarantees are strongest near the uniform weighting

Learning the comparisons does not automatically reconstruct every trade-off

Take-home messages

- ① **Scalar rewards** are a powerful representation assumption
- ② **Preferences** are ordinal relations; recovering cardinal utilities requires a link model and additional assumptions
- ③ **Comparisons** can explicitly distinguish indifference from incomparability and motivate vector-valued rewards
- ④ **Identifiability issues**: equal comparison distributions need not reveal individual objectives or the entire Pareto front

Thank You for Your Attention!



References I

- Michael Bowling, John D. Martin, David Abel, and Will Dabney. Settling the reward hypothesis. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 3003–3020, 2023. URL <https://proceedings.mlr.press/v202/bowling23a.html>.
- Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.
- Róbert Busa-Fekete and Eyke Hüllermeier. A survey of preference-based online learning with bandit algorithms. In *Proceedings of the 25th International Conference on Algorithmic Learning Theory*, pages 18–39, 2014.
- Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017.
- Roger R. Davidson. On extending the Bradley–Terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, 65(329):317–328, 1970.
- Gerard Debreu. Representation of a preference ordering by a numerical function. In R. M. Thrall, C. H. Coombs, and R. L. Davis, editors, *Decision Processes*, pages 159–165. Wiley, 1954.
- Simone Drago, Marco Mussi, and Alberto Maria Metelli. Towards theoretical understanding of sequential decision making with preference feedback. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 14499–14514, 2025.
- Simone Drago, Marco Mussi, Leonardo Bianconi, and Alberto Maria Metelli. On generalizing preference-based reinforcement learning to multi-objective problems, 2026a. Under review for AISTATS 2027.

References II

- Simone Drago, Marco Mussi, and Alberto Maria Metelli. On the computational limits of sequential decision making with trajectory-preference feedback, 2026b. Technical report.
- Kfir Eliaz and Efe A. Ok. Indifference or indecisiveness? choice-theoretic foundations of incomplete preferences. *Games and Economic Behavior*, 56(1):61–86, 2006.
- Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.
- R. Duncan Luce. *Individual Choice Behavior: A Theoretical Analysis*. Wiley, 1959.
- Sahand Negahban, Sewoong Oh, and Devavrat Shah. Rank centrality: Ranking from pairwise comparisons. *Operations Research*, 65(1):266–287, 2017.
- Efe A. Ok. Utility representation of an incomplete preference relation. *Journal of Economic Theory*, 104(2):429–449, 2002. doi: 10.1006/jeth.2001.2814.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022.
- Diederik M. Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48:67–113, 2013.
- Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2 edition, 2018.
- Louis L. Thurstone. A law of comparative judgment. *Psychological Review*, 34(4):273–286, 1927.
- Mihalis Yannakakis. The complexity of the partial order dimension problem. *SIAM Journal on Algebraic and Discrete Methods*, 3(3):351–358, 1982.
- Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The K-armed dueling bandits problem. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.