
Modeling Incomparability: A New Rationality Paradigm for Preference-Based Reinforcement Learning

Simone Drago¹ Marco Mussi¹ Leonardo Bianconi¹ Alberto Maria Metelli¹

Abstract

In this work, we study the *reinforcement learning* (RL) problem from pairwise trajectory comparisons provided by a human expert. We generalize preference-based RL by formalizing a novel setting in which the expert can also label trajectory pairs as *incomparable*, i.e., when neither trajectory “dominates” the other. We introduce the learning problem and the desiderata that its solution should satisfy. Then, we propose a novel Bradley–Terry-inspired rationality model that effectively captures incomparabilities and infers a multi-dimensional reward function, and we study its properties. We provide a sample complexity analysis for learning the model parameters when a dataset is available. Finally, we evaluate our model’s ability to reconstruct a reward function that aligns with the expert’s comparisons in simulated environments and to recover the Pareto frontier of policies, along with a robustness analysis across varying levels of expert rationality.

1. Introduction

In *reinforcement learning* (RL, Sutton & Barto, 2018), a learning agent interacts with an environment observing a *state* and, based on it, selecting an *action*, to optimize a certain objective function over a given time horizon. Crucially, the learning process is guided by a *reward* function, i.e., a numerical signal provided as feedback to the agent after each action (Sutton, 2004). With such a formulation, the agent’s goal is to maximize the expected cumulative reward. The reward is often referred to as “the most succinct description of a task” (Ng & Russell, 2000). However, defining a reward function is a complex endeavor. The system engineer, who designs it, must select numerical values to induce a behavior that “solves” the task, while avoiding unwanted, possibly

dangerous, phenomena (e.g., reward hacking, Amodei et al., 2016). This process, namely *reward engineering* (Dewey, 2014), is a trial-and-error iterative process that relies on domain knowledge and requires tweaking the reward function, as the induced behavior is highly sensitive to misspecified rewards (Pan et al., 2022).

Preference-based RL (PbRL, Wirth et al., 2017) has emerged as a powerful paradigm for overcoming the inherent reward design challenges of RL. It avoids the reward engineering process entirely, relying instead on pairwise comparisons provided by a *human expert* and learning the policy that best aligns with them. PbRL has achieved remarkable results in domains ranging from robotics (Lee et al., 2021b) to *large language models* (Ouyang et al., 2022). One common approach, namely *reinforcement learning from human feedback* (RLHF, Christiano et al., 2017) assumes the existence of an underlying, unknown reward (Friedman & Savage, 1952), and comprises two steps: (i) reward model estimation from the expert’s preferences and (ii) policy optimization via traditional RL methods. PbRL’s achievements are largely built on the use of *rationality models*, which link the underlying (unknown) reward function to observed preferences and guide the stochastic preference generation process. The *Bradley-Terry* (BT, Bradley & Terry, 1952) model is most commonly employed in the literature (see, e.g., Christiano et al., 2017; Rafailov et al., 2023; Munos et al., 2024), and it is fed by utility functions that quantify the “goodness” of the alternatives, often represented by the cumulative reward. Under the BT model, only “clear” preferences are possible, i.e., when presented with two alternatives, the human expert always provides a preference label, usually generated stochastically. The probability of observing the preference of one alternative over the other is assumed to be proportional to (a monotonic transformation of) their difference in utility. Alternative approaches exist, e.g., assuming a different probabilistic model (e.g., Thurstone, 1927) and permitting *indifference* as a feedback (Rao & Kupper, 1967), i.e., allowing for the additional feedback of the two alternatives being equivalent.

Because of the presence of clear preferences and indifference, one crucial assumption of PbRL is that a *scalar* utility function is sufficient to solve the problem. However, in

¹Politecnico di Milano, Milan, Italy.

Correspondence to: Simone Drago <simone.drago@polimi.it>.

2nd Workshop on Epistemic Intelligence in Machine Learning (EIML@ICML 2026), Seoul, South Korea. Copyright 2026 by the author(s).

	Single-Objective	Multi-Objective
Rewards	RL	MORL
Human Feedback	PbRL	CbRL

Table 1. Positioning of CbRL in the literature.

real-world scenarios, an expert may be unable to provide either of these forms of feedback. A practical solution is to attribute this to bad data or to the expert’s irrationality, and discard the sample, as in (Christiano et al., 2017).

We argue, instead, that abstaining from providing a clear preference or stating indifference is an expression of the human expert’s rationality, or, put differently, a feature of human nature.

Indeed, it is intuitive to understand that there exist scenarios in which a human may be unable to compare two alternatives, such as when *multiple objectives are in conflict*, e.g., speed vs. safety in autonomous driving, where it may happen that neither alternative *dominates* the other, i.e., is better in all aspects. In such cases, the inability to choose is not a failure of the expert but the expression of a different model of rationality. We refer to this feedback as *incomparability*. When utility functions are involved, this indicates the existence of an underlying *multi-dimensional* reward function that represents the various objectives. In the literature, the framework that allows dealing with multiple objectives is that of *multi-objective reinforcement learning* (MORL, Hayes et al., 2022). In MORL, a single, universally optimal behavior, so-called *utopian*, is typically unattainable and, thus, the goal is to recover a set of non-dominated behaviors, the so-called *Pareto frontier* (Roijers et al., 2013).

The previous argument reveals that PbRL and MORL, taken individually, each address only half of the problem: MORL handles multiple conflicting objectives but requires dense reward signals, while PbRL handles human feedback but assumes a scalar utility that cannot represent incomparability. To the best of our knowledge, a framework capable of recovering a Pareto frontier of behaviors from *comparison*¹ feedback alone is still lacking. In this work, to fill this gap, we introduce *comparison-based reinforcement learning* (CbRL). As shown in Table 1, CbRL establishes a connection between the two, serving as the human-feedback counterpart of MORL and the multi-objective counterpart of PbRL. By modeling incomparability in all regards as a feedback signal, CbRL augments MORL with pairwise comparisons and extends PbRL beyond scalar utility functions. In doing so, it must confront the challenges of PbRL and MORL jointly, i.e., learning from stochastically-generated preferences and dealing with a multi-dimensional utility, respectively.

¹We use the term “comparison” to include clear preferences, indifference, and incomparability.

Related Works. Our work aims to extend PbRL to the multi-objective scenario. Traditional PbRL assumes that human preferences can be defined by a single scalar reward function, employing the BT model or similar alternatives (e.g., Thurstone, 1927) to link cumulative rewards to observed feedback. However, in recent years, this assumption has come under scrutiny from the perspective of social choice theory, showing that scalar rewards are often insufficient to capture the heterogeneity of preferences across a diverse set of users (Chakraborty et al., 2024). Whereas the concept of indifference between alternatives has a long history in the literature (Rao & Kupper, 1967; Davidson, 1970), the phenomenon of *incomparability*, i.e., an expert refusing to choose between alternatives due to an inherent conflict between contrasting objectives, has largely been overlooked, either attributing incomparable samples to bad data and discarding them (e.g., Christiano et al., 2017), or bypassing the problem by eliciting a multi-objective preference conditioned to a known weighting of the objectives (Mu et al., 2025). In the literature, incomparability was first introduced to model the *rational choice* of the expert to not take a definite position w.r.t. a comparison (Sen, 1970; Roy & Vincke, 1984). Our CbRL framework, motivated by the idea of incomparabilities being a consequence of a multi-objective problem, and thus inducing a Pareto frontier of policies (Abdelkareem et al., 2022), enables recovering a Pareto frontier of policies by explicitly modeling incomparability through a novel multi-objective rationality model, while limiting the cognitive burden on the human expert.

Original Contributions. The contributions of this paper are summarized as follows:

- In Section 2, we formalize the novel MDPC framework, incorporating incomparability as a rational feedback, and the desiderata of an MDPC-compliant rationality model. We show that no rationality model can satisfy the desiderata and yield a convex negative log-likelihood (Proposition 2.1).
- In Section 3, we devise a novel BT-inspired rationality model for CbRL, we analyze its properties showing that it complies with our desiderata (Lemma 3.1), and we study the sample complexity for learning its parameters (Theorems 3.2 and 3.3).

In Appendix A, we report a more in-depth discussion of related works. In Appendix B, we report the derivation of the function we use to represent the incomparability score. In Appendix C, we report omitted proofs. In Appendix D, we report the numerical validation of our model, together with methodology and implementation details, discussing the limitations of PbRL approaches in tackling CbRL, providing a robustness analysis to experts’ irrationality, and discussing how to learn the Pareto frontier of behaviors starting from the reconstructed reward function.

Notation. Given $a, b \in \mathbb{N}$ with $a < b$, we define $\llbracket a \rrbracket := \{1, 2, \dots, a\}$ and $\llbracket a, b \rrbracket := \{a, a+1, \dots, b\}$. Given a finite set $\mathcal{X}$, we denote as $\Delta(\mathcal{X})$ the probability simplex over $\mathcal{X}$, and with $|\mathcal{X}|$ its cardinality. We denote as $\mathbf{1}_n, \mathbf{0}_n \in \mathbb{R}^n$ the column vectors of n ones and zeros, respectively. Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ be real-valued vectors, we denote as $\mathbf{x} \odot \mathbf{y}$ the Hadamard product, as $t\mathbf{x}$ the rescaling of $\mathbf{x}$ by $t \in \mathbb{R} \cup \{\pm\infty\}$, and as $\text{std}(\mathbf{x}) = \|\mathbf{x} - \bar{x}\mathbf{1}_n\|_2/\sqrt{n}$ its standard deviation, where $\bar{x}$ is the empirical mean. Given a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, we denote the column vector obtained by stacking the columns of a matrix as $\text{vec}(\mathbf{A}) \in \mathbb{R}^{mn}$. Let $P, Q \in \Delta(\mathcal{X})$ such that $Q(x) = 0$ implies $P(x) = 0$ for every $x \in \mathcal{X}$, we define their Kullback-Leibler (KL) divergence as $D_{\text{KL}}(P \| Q) := \sum_{x \in \mathcal{X}} P(x) \log(P(x)/Q(x))$. We use the $\tilde{O}(\cdot)$ notation to omit constant terms and logarithmic factors that do not depend on δ .

2. Framework for Incomparability

In this section, we introduce *incomparability* feedback, propose the novel setting of the *Markov decision processes with comparison* (MDPC), a mathematical framework for learning from comparisons, discuss the desiderata for a rationality model, and formulate the learning goal. Finally, we provide an important impossibility result.

Comparison Feedback. Suppose a human expert is shown a pair of alternatives and is asked to provide a *comparison* between the two. Intuitively, we expect one of the following outcomes:

- ($>$) *Direct preference*, when the expert shows a clear preference for the first alternative;
- ($<$) *Inverse preference*, when the expert shows a clear preference for the second alternative;
- ($\asymp$) *Indifference*, when the expert considers the two alternatives to be equivalent;
- ($\parallel$) *Incomparability*, when none of the above holds.

Incomparability was first introduced in decision theory as the expert’s choice to “not take a definite position” w.r.t. a comparison (Roy & Vincke, 1984). An axiomatic foundation based on first-order logic (Tsoukias & Vincke, 1995) was later proposed, capturing two fundamental reasons for the arising of incomparabilities. The first is when the data is incomplete, the request is ambiguous, or the expert presents some degree of irrationality and cannot choose.² The second, instead, occurs when there is an inherent conflict among alternatives, or when the expert considers a set of criteria that cannot be directly combined. In this work, we focus on the latter. When cast to RL, a convenient characterization is that of multiple, possibly conflicting, objectives, i.e., the setting of MORL.

²This is the case commonly assumed in PbRL, when data is labeled by human experts (e.g., Christiano et al., 2017).

Framework. Before defining the *Markov decision process with comparisons*, we report the framework it generalizes, namely the finite-horizon *Markov Decision Process without reward* (MDP\texttt{R}, Abbeel & Ng, 2004), i.e., a tuple $(\mathcal{S}, \mathcal{A}, H, p, \mu)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $H \in \mathbb{N}$ is the horizon, $p = (p_h)_{h \in \llbracket H \rrbracket}$ is the stage-dependent transition model defined as $p_h : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$ for every $h \in \llbracket H \rrbracket$, and $\mu \in \Delta(\mathcal{S})$ is the initial state distribution. A *trajectory* is a sequence of state-action pairs, as $\tau := (s_i, a_i)_{i \in \llbracket H \rrbracket}$, and the space of possible trajectories as $\mathcal{T}$.

We now formalize the MDPC, which allows us to define the problem of learning from comparisons.

Definition 2.1 (Markov Decision Process with Comparisons). A *Markov decision process with comparisons* (MDPC) is a tuple $(\mathcal{S}, \mathcal{A}, H, p, \mu, \rho)$, where $(\mathcal{S}, \mathcal{A}, H, p, \mu)$ is an MDP\texttt{R} and $\rho : \mathcal{T}^2 \rightarrow \Delta(\mathcal{C})$ is the comparison probability distribution where $\mathcal{C} = \{>, <, \asymp, \parallel\}$ is the outcome set.

Notably, the MDPC can be interpreted as the “multi-objective counterpart” of the *Markov decision process with preferences* (Wirth et al., 2017). Thus, upon selecting two trajectories $\tau, \tau' \in \mathcal{T}$, the agent receives a *comparison* from the set of outcomes $\mathcal{C} = \{>, <, \asymp, \parallel\}$ generated stochastically by an expert according to comparison probability distribution ρ . For example, we represent the probability of a trajectory pair being labeled as indifferent as $\rho(\asymp | (\tau, \tau'))$. Throughout the paper, we will use the infix notation $\rho(\tau \circ \tau')$ to denote $\rho(\circ | (\tau, \tau'))$ for an outcome $\circ \in \mathcal{C}$.

We model the behavior of an agent interacting with an MDPC via a stage-dependent Markovian policy $\pi := (\pi_h)_{h \in \llbracket H \rrbracket}$, where $\pi_h : \mathcal{S} \rightarrow \Delta(\mathcal{A})$ for each $h \in \llbracket H \rrbracket$.

To provide a tractable model, as customary in PbRL, we follow the *utility representation theorem* (von Neumann & Morgenstern, 1947), which postulates that any *rational agent*³ generates preferences in terms of an underlying utility function defined over the set of available options. In CbRL, this translates to the definition of a multi-dimensional utility function $\mathbf{u} : \mathcal{T} \rightarrow \mathbb{R}^d$, mapping trajectories to d -dimensional utility vectors, where $d \in \mathbb{N}$ is the number of objectives, such that the probability of observing a comparison is proportional to a function of the utility difference of the two trajectories (Drago et al., 2025). Formally, this corresponds to enforcing $\rho(\tau \circ \tau') := f(\circ | \delta(\tau, \tau'))$ for every $\circ \in \mathcal{C}$ and $\tau, \tau' \in \mathcal{T}$, with $\delta(\tau, \tau') := \mathbf{u}(\tau) - \mathbf{u}(\tau')$ and $f : \mathbb{R}^d \rightarrow \Delta(\mathcal{C})$. We call f a *rationality model*.

We now define and discuss the desiderata that a rationality model f should satisfy.

³In CbRL, we say that the human expert is *rational* if their interaction with the learning agent can be modeled as an MDPC.

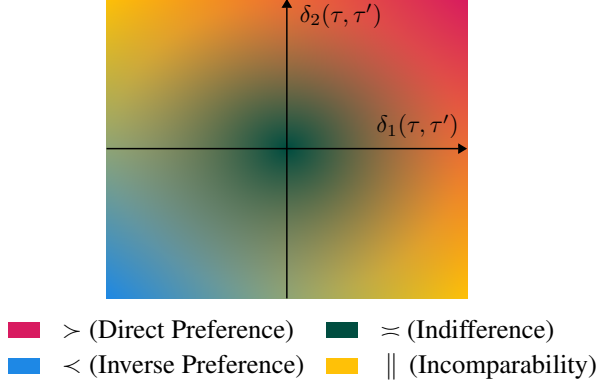


Figure 1. Desiderata for a 2D CbRL problem.

Definition 2.2 (MDPC Model Desiderata). *A function $f : \mathbb{R}^d \rightarrow \Delta(\mathcal{C})$ is a desirable rationality model for an MDPC if the following properties hold for every $\tau, \tau' \in \mathcal{T}$:*

$$\arg \max_{\circ \in \mathcal{C}} \lim_{\delta \rightarrow +\infty \mathbf{1}_d} f(\circ|\delta) = \{>\}, \quad (1)$$

$$\arg \max_{\circ \in \mathcal{C}} \lim_{\delta \rightarrow -\infty \mathbf{1}_d} f(\circ|\delta) = \{<\}, \quad (2)$$

$$\arg \max_{\circ \in \mathcal{C}} \lim_{\delta \rightarrow \mathbf{0}_d} f(\circ|\delta) = \{\asymp\}, \quad (3)$$

$$\arg \max_{\circ \in \mathcal{C}} \lim_{\delta \rightarrow +\infty \mathbf{t}} f(\circ|\delta) = \{||\}, \quad (4)$$

where $\delta = \delta(\tau, \tau')$ and $\mathbf{t} \in \{-1, 1\}^d \setminus \{\mathbf{1}_d, -\mathbf{1}_d\}$.

Definition 2.2 puts the focus of the rationality model on characterizing the mode outcome (i.e., the highest probability one) in various limits of utility differences. Intuitively, given two trajectories $\tau, \tau' \in \mathcal{T}$, if τ dominates τ' , i.e., is better w.r.t. all objectives, we expect that the most probable outcome that a human expert can provide is direct preference (Equation 1). Indeed, in this case, the utility difference vector δ has all positive components and is located in the positive direction along the main diagonal. A similar argument holds for the inverse preference, showing that clear dominance in either direction reflects the desiderata of PbRL (Equation 2). Additionally, since a rational expert implicitly evaluates the underlying utility of the alternatives, we expect that in case the two trajectories are similar, the most probable outcome is indifference. This is reflected by Equation (3), prescribing that indifference must be the most likely outcome when the utilities of the two trajectories are element-wise equal. Finally, we expect incomparability to be the most likely outcome when neither alternative dominates the other, i.e., when one is better in terms of one objective and worse in terms of another, e.g., “fast but risky” vs. “slow but safe” in autonomous driving. Equation (4) captures this behavior, prescribing that, along all the non-standard diagonals of the d -dimensional hyper-space of utility differences, the most probable outcome is incomparability ($||$). Intuitively, on such diagonals, the util-

ity difference vector $\delta(\tau, \tau')$ has at least two contrasting components, and thus non-dominance between the trajectories for at least one objective. In Figure 1, we report a graphical representation of the desiderata of Definition 2.2 for the 2-dimensional case.

Learning Goal. Given a class of models $\mathcal{F}$ complying with Definition 2.2, a hypothesis space $\mathcal{U} \subseteq \{\mathbf{u} : \mathcal{T} \rightarrow \mathbb{R}^d\}$ of utility functions, and a dataset $\mathcal{D} = \{(\tau_i, \tau'_i, \circ_i)\}_{i=1}^N$ of comparison triples, we aim to find the *maximum likelihood* pair $(\hat{f}, \hat{\mathbf{u}})$, made of a rationality model $\hat{f}$ and utility estimator $\hat{\mathbf{u}}$, obtained by minimizing the *empirical negative log-likelihood* (NLL) over the dataset, i.e.:

$$\begin{aligned} (\hat{f}, \hat{\mathbf{u}}) &\in \arg \min_{f \in \mathcal{F}, \mathbf{u} \in \mathcal{U}} \hat{\mathcal{L}}(f, \mathbf{u}|\mathcal{D}) \\ &:= - \sum_{(\tau_i, \tau'_i, \circ_i) \in \mathcal{D}} \log f(\circ_i | \mathbf{u}(\tau_i) - \mathbf{u}(\tau'_i)). \end{aligned} \quad (5)$$

In principle, we would aim to select a class of models $\mathcal{F}$ such that the NLL of Equation (5) is convex (Boyd & Vandenberghe, 2004). Unfortunately, this is unfeasible if $\mathcal{F}$ complies with Definition 2.2, as demonstrated by the following negative result.

Proposition 2.1 (Impossible Convex NLL). *There exists no function $f : \mathbb{R}^d \rightarrow \Delta(\mathcal{C})$ that satisfies Definition 2.2, and such that $-\log f(\circ|\delta)$ is convex in δ for every $\circ \in \mathcal{C}$.*

This result provides a strong insight into the complexity of CbRL. Intuitively, non-convexity arises because Definition 2.2 requires the most probable outcome to “change” twice along any non-standard diagonal, from incomparability at $\pm\infty$ to indifference in the origin. Indeed, this is necessary, as the problem is symmetric w.r.t. every pair of conflicting objectives, and showing the expert the pair (τ', τ) rather than (τ, τ') does not change the evaluation of the underlying utilities. As a consequence, it is more challenging to provide guarantees about reaching the global optimum of Equation (5).

3. Model and Theoretical Analysis

In this section, we propose the novel *multi-objective Bradley-Terry* model, show its compliance with Definition 2.2, and study the sample complexity of learning its parameters from a fixed dataset.

The Multi-Objective Bradley-Terry Model. Let $\tau, \tau' \in \mathcal{T}$ be two trajectories generated by interacting with an MDPC. We now define the novel *multi-objective Bradley-Terry* model (MOBT), which we employ to represent the comparison-generation process of a human expert. With a slight notation overload, we define the MOBT model class $\mathcal{F}$ as the set of differentiable *softmax* functions:

$$f(\circ|\tau, \tau') := \frac{\exp(h_\circ(\tau, \tau'))}{\sum_{\circ' \in \mathcal{C}} \exp(h_{\circ'}(\tau, \tau'))}, \quad (6)$$

defined through $h_\circ$, which we refer to as *scores*, for every $\circ \in \mathcal{C}$, defined as:

$$h_{>}(\tau, \tau') := \mathbf{1}_d^\top \delta(\tau, \tau') / (2d), \quad (7)$$

$$h_{<}(\tau, \tau') := -\mathbf{1}_d^\top \delta(\tau, \tau') / (2d). \quad (8)$$

$$h_{=}(\tau, \tau') := \alpha, \quad (9)$$

$$h_{\parallel}(\tau, \tau') := \sqrt{d} \text{std}(\delta(\tau, \tau')) + \beta, \quad (10)$$

where $\alpha, \beta \in \mathbb{R}$ are learnable parameters.⁴ Intuitively, $\delta(\tau, \tau')$ allows us to quantify the utility difference of the two trajectories as a point in $\mathbb{R}^d$. It clearly captures the concepts of dominance, i.e., all positive or all negative elements, and non-dominance, i.e., some positive elements and some negative ones. The standard deviation in Equation (10) is a convenient and intuitive choice to represent incomparability, as it quantifies the dispersion of the elements of the vector, which in our setting corresponds to quantifying how much the utility vectors differ across the various objectives. Let us now discuss MOBT's compliance with the desiderata of Definition 2.2.

Lemma 3.1. *Choosing $\alpha, \beta \in \mathbb{R}$ such that $\alpha > 0$ and $\alpha > \beta$, the MOBT model complies with the desiderata of Definition 2.2.*

We prove this result by tackling each desiderata separately. Compliance w.r.t. clear preferences holds by observing that the utility difference terms are all aligned along the standard diagonal. The choice of α and β ensures compliance with the indifference desiderata. Finally, Equation (4) prescribes a desideratum along the non-standard diagonals of the d -dimensional hyperspace. Intuitively, we model $h_{\parallel}(\tau, \tau')$ as a quantity that increases with the distance from the standard diagonal, and we show that the desideratum holds because $h_{\parallel}$ grows quicker than $h_{>}$ and $h_{<}$ along non-standard diagonals. We now show that the MOBT model generalizes BT and falls back to it in 1-dimensional problems.⁵

Remark 3.1 (Equivalence of MOBT and BT in PbRL). *Consider a PbRL instance where a human expert generates preferences according to the BT model, i.e., $\rho(\tau > \tau') = \sigma(u(\tau) - u(\tau'))$, where $\sigma(\cdot)$ is the sigmoid function. This corresponds to a CbRL instance with $d = 1$, since we cannot observe indifferences nor incomparabilities. In this case, MOBT falls back to BT. Setting $\alpha = \beta = -\infty$, we get a probability of zero of generating indifferences and incomparabilities. Thus, we can rewrite:*

$$f(> | \tau, \tau') = \frac{\exp(\delta(\tau, \tau')/2)}{\exp(\delta(\tau, \tau')/2) + \exp(-\delta(\tau, \tau')/2)}$$

⁴Throughout the paper, we assume knowledge of the number of objectives d . However, this assumption is not strictly necessary, as the number of dimensions can be overestimated based on the data (Drago et al., 2025).

⁵MOBT falls back to BT also when all objectives are aligned, i.e., when there exists a policy that jointly optimizes all.

$$= \frac{\exp(\delta(\tau, \tau'))}{\exp(\delta(\tau, \tau')) + 1} = \sigma(\delta(\tau, \tau')),$$

where $\sigma(\cdot)$ is the sigmoid function. Clearly, the same reasoning also holds for the inverse preference.

Sample Complexity. We now study the sample complexity of recovering a multi-dimensional reward model under the MOBT model in the case of a parametric utility function. Let $\Theta \subseteq \mathbb{R}^{d_\Theta}$ be the parameter space, with slight abuse of notation, we can rewrite the learning goal of Equation (5) as:

$$\hat{\theta} \in \arg \min_{\theta \in \Theta} \hat{\mathcal{L}}(\theta | \mathcal{D}). \quad (11)$$

For the sake of the analysis, we consider the case of a linear utility model. This assumption is widely adopted in both theoretical (Jin et al., 2023) and practical RL (Silver et al., 2014).

Assumption 3.1 (Feature-based Utility Representation). *There exists a known feature map $\varphi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}^k$ and an unknown matrix $\mathbf{W} \in \mathbb{R}^{d \times k}$ such that the expert's rationality model is defined in terms of utility function $u(\tau) := \mathbf{W}\phi(\tau)$, where $\phi(\tau) = \sum_{h \in [H]} \varphi(s_h, a_h)$ for every $\tau \in \mathcal{T}$.*

Notably, this assumption is convenient for analytical purposes, but is not necessary, as the MOBT model can handle any function mapping (τ, τ') to $\delta(\tau, \tau')$. In practice, ϕ can be constructed via domain-specific feature engineering (Abbeel & Ng, 2004) or extracted from raw inputs through unsupervised pre-training (Lee et al., 2021b). Under Assumption 3.1, we can rewrite model f in terms of its feature representation, i.e., $f(\circ | \mathbf{W}(\phi(\tau) - \phi(\tau')))$ for every $\circ \in \mathcal{C}$. As customary in the literature, we consider a boundedness assumption w.r.t. parameter θ^* (Sadigh et al., 2017; Yang & Wang, 2020; Zhao et al., 2024) and feature vector difference (Zhang et al., 2021; Wang et al., 2020).

Assumption 3.2 (Boundedness). *There exist $R, \Lambda \geq 0$ such that:*

- The expert's model is parametrized by $\theta^* := (\text{vec}(\mathbf{W}^*), \alpha^*, \beta^*) \in \Theta \subseteq [-R, R]^{dk+2}$.
- The known feature map $\phi : \mathcal{T} \rightarrow \mathbb{R}^k$ satisfies $\|\phi(\tau) - \phi(\tau')\|_2 \leq \Lambda$, for every $\tau, \tau' \in \mathcal{T}$.

Assumptions 3.1 and 3.2 together ensure that the utility $u(\tau) = \mathbf{W}\phi(\tau)$ is bounded for every $\tau \in \mathcal{T}$. We can now state our sample complexity result.

Theorem 3.2. *Let $\theta^* \in \Theta$ be the expert's parametrization, and $\mathcal{D} = \{(\tau_i, \tau'_i, \circ_i)\}_{i=1}^N$ be a dataset collected i.i.d. from an unknown distribution $\mathcal{P}_{\theta^*} \in \Delta(\mathcal{T} \times \mathcal{T} \times \mathcal{C})$ such that $\mathcal{P}_{\theta^*}(\tau, \tau', \circ) = f(\circ | \tau, \tau'; \theta^*) Q((\tau, \tau'))$, for some $Q \in \Delta(\mathcal{T}^2)$. Let $\hat{\theta}$ be the solution to Equation (11), and define $\mathcal{P}_{\hat{\theta}}$ as its induced distribution under Q . Under As-*

assumptions 3.1 and 3.2, for every $\delta \in (0, 1)$, it holds that:

$$D_{\text{KL}}(\mathcal{P}_{\theta^*} \parallel \mathcal{P}_{\hat{\theta}}) \leq \tilde{\mathcal{O}} \left(R\Lambda dk \sqrt{\frac{\log(1/\delta)}{N}} \right). \quad (12)$$

Some comments are in order. First, we observe that this result bounds the KL divergence between the expert’s *induced distribution* $\mathcal{P}_{\theta^*}$ and the estimated one $\mathcal{P}_{\hat{\theta}}$. Notice that we are indeed controlling the divergence between the true and learned comparison-generation models, since $D_{\text{KL}}(\mathcal{P}_{\theta^*} \parallel \mathcal{P}_{\hat{\theta}}) = \mathbb{E}_{\tau, \tau' \sim Q} [D_{\text{KL}}(f(\cdot | \tau, \tau'; \theta^*) \parallel f(\cdot | \tau, \tau'; \hat{\theta}))]$. Due to the non-convexity of the NLL, no guarantees on the parameter distance $\|\hat{\theta} - \theta^*\|_2$ can be derived. The bound, as expected, decreases with the number of samples and increases linearly with the bound of the feature difference Λ , as well as the dimension and bound of the parameter space, i.e., dk and R . Second, our result holds even when the dataset is generated by a model $\mathcal{P}^*$ that falls outside the MOBT class. In such a case, we define $\theta^* \in \arg \min_{\theta \in \Theta} \mathbb{E}_{\mathcal{P}^*} [\hat{\mathcal{L}}(\theta | \mathcal{D})]$, i.e., the KL-projection of $\mathcal{P}^*$ onto our class and add to the bound of Equation (12) an additive term $D_{\text{KL}}(\mathcal{P}^* \parallel \mathcal{P}_{\theta^*})$ to quantify this approximation.

Notably, Theorem 3.2 defines $\hat{\theta}$ as the *solution* to Equation (11), i.e., its global optimum. However, this might be hard to obtain in practice due to the non-convexity of the optimization manifold. Nonetheless, we can recover a local minimum of $\hat{\mathcal{L}}(\theta | \mathcal{D})$ by resorting to numerical optimization methods (Nocedal & Wright, 2006; Paszke et al., 2019). In the following theorem, we bound the KL divergence between the distribution induced by *any* attainable local optimum and $\mathcal{P}_{\theta^*}$.

Theorem 3.3. *Let $\theta^* \in \Theta$ be the expert’s parametrization, and $\mathcal{D} = \{(\tau_i, \tau'_i, \circ_i)\}_{i=1}^N$ be a dataset collected i.i.d. from an unknown distribution $\mathcal{P}_{\theta^*}$. Let $\hat{\theta}_{\text{loc}}$ be a local minimum of Equation (11), and define as $\mathcal{P}_{\hat{\theta}_{\text{loc}}}$ its induced distribution under the same $Q \in \Delta(\mathcal{T}^2)$. Under Assumptions 3.1 and 3.2, for every $\delta \in (0, 1)$, it holds that:*

$$D_{\text{KL}}(\mathcal{P}_{\theta^*} \parallel \mathcal{P}_{\hat{\theta}_{\text{loc}}}) \leq \tilde{\mathcal{O}} \left(R\Lambda dk \sqrt{\frac{\log(1/\delta)}{N}} + \mathcal{P}_{\theta^*}(\parallel) \Lambda d \sqrt{k} R \right), \quad (13)$$

where $\mathcal{P}_{\theta^*}(\parallel) := \int_{\tau, \tau'} \mathcal{P}_{\theta^*}(d\tau, d\tau', \parallel)$ is the marginal probability of observing an incomparability.

We observe that moving from the global optimum, which we are not guaranteed to reach, to a local one results only in an additive error term depending on the probability of observing incomparabilities, which is an intrinsic property of the problem and, as already discussed, the source of non-convexity of the model. We can further bound this quantity by observing that, whenever $\beta \geq 0$, we have $\mathcal{P}_{\theta^*}(\tau, \tau', \parallel) \leq \exp(h_{\parallel}(\tau, \tau')) / (1 + \exp(h_{\parallel}(\tau, \tau')))$, which allows us to write

$\mathcal{P}_{\theta^*}(\parallel) \leq \sqrt{d} \mathbb{E}_{(\tau, \tau') \sim Q} [\text{std}(\delta(\tau, \tau'))] + \beta$. Here, the expected standard deviation of the utility difference vector acts as an *index of conflict* between the objectives, and β represents a minimum level of incomparability. Notice that when $\beta \rightarrow -\infty$, no incomparability is present, i.e., $\mathcal{P}_{\theta^*}(\parallel) = 0$.

4. Discussion

In this work, we proposed the novel CbRL, generalizing PbRL by allowing for *incomparability* feedback and defining the human-feedback-based counterpart of MORL. We formalized this setting as an *MDP with comparisons*, defined the desiderata a rationality model should satisfy, and demonstrated that learning in this setting requires solving a non-convex optimization problem. Then, we devised the MOBT model and studied its sample complexity.

4.1. Open Challenges

We now discuss some challenges that remain open for future research. First, we considered a scenario in which we are provided with an *offline* dataset of comparisons. An important future research direction is to tackle the *online* scenario. Indeed, this requires balancing (i) the minimization of the number of human interactions and (ii) the maximization of the accuracy within the portion of the trajectory space explored by Pareto-optimal policies. This, in turn, requires devising a principled method for selecting trajectory pairs to compare. Second, we observe that the fundamental modeling choice of this work, i.e., considering a 4-class comparison feedback, has the advantage of putting the least cognitive weight on the human expert. However, this comes at a cost, as the informativeness of the samples decreases with the number of objectives, which is a well-known scaling problem in the literature (Hayes et al., 2022, Section 10.2). Indeed, in high-dimensional problems, pairwise feedback, such as incomparability, may introduce ambiguity in assigning the *credit* of the feedback across different objectives. In light of this, we believe that a relevant research direction is to study the trade-off between the *methodological representational power* and the *human cognitive strain* of the preference elicitation method, aiming to strike a balance that achieves near-optimal performance while avoiding overburdening the human expert. Finally, a third, orthogonal future research direction concerns the availability of real data. To the best of our knowledge, there are currently no datasets that comprise human-generated labels and include both indifference and incomparability, as the available data allows at most for surrogate comparison generation based on annotations across multiple dimensions (see, e.g., Wang et al., 2025, for an example in RLHF). Indeed, although complex, collecting such data for tasks from various fields would enable the definition of principled benchmarks for CbRL.

References

- Abbeel, P. and Ng, A. Y. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 69 of *ACM International Conference Proceeding Series*. ACM, 2004.
- Abdelkareem, Y., Shehata, S., and Karray, F. Advances in preference-based reinforcement learning: A review. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 2527–2532. IEEE, 2022.
- Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.
- Bartlett, P. L. and Wegkamp, M. H. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9(59):1823–1840, 2008.
- Bemporad, A., Morari, M., Dua, V., and Pistikopoulos, E. N. The explicit linear quadratic regulator for constrained systems. *Automatica*, 38(1):3–20, 2002.
- Bıyık, E., Huynh, N., Kochenderfer, M. J., and Sadigh, D. Active preference-based gaussian process regression for reward learning. *arXiv preprint arXiv:2005.02575*, 2020.
- Boyd, S. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, 2004.
- Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.
- Chakraborty, S., Qiu, J., Yuan, H., Koppel, A., Manocha, D., Huang, F., Bedi, A., and Wang, M. MaxMin-RLHF: Alignment with diverse human preferences. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6116–6135. PMLR, 2024.
- Chen, Y.-T., Hung, W., Wu, B.-S., Hong, Z.-W., and Hsieh, P.-C. A reward-free viewpoint on multi-objective reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2026.
- Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems (NIPS)*, volume 30, 2017.
- Davidson, R. R. On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, 65(329): 317–328, 1970.
- Dewey, D. Reinforcement learning and the reward engineering principle. In *AAAI Spring Symposium Series*, 2014.
- Drago, S., Mussi, M., and Metelli, A. M. Towards theoretical understanding of sequential decision making with preference feedback. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 14499–14514. PMLR, 2025.
- Felten, F., Alegre, L. N., Nowe, A., Bazzan, A., Talbi, E. G., Danoy, G., and C. da Silva, B. A toolkit for reliable benchmarking and research in multi-objective reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 23671–23700, 2023.
- Friedman, M. and Savage, L. J. The expected-utility hypothesis and the measurability of utility. *Journal of Political Economy*, 60(6):463–474, 1952.
- Fürnkranz, J., Hüllermeier, E., Cheng, W., and Park, S.-H. Preference-based reinforcement learning: a formal framework and a policy iteration algorithm. *Machine Learning*, 89(1):123–156, 2012.
- Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1861–1870. PMLR, 2018.
- Hardy, G. H., Littlewood, J. E., and Pólya, G. *Inequalities*. Cambridge University Press, 2nd edition, 1964.
- Hayes, C. F., Rădulescu, R., Bargiacchi, E., Källström, J., Macfarlane, M., Reymond, M., Verstraeten, T., Zintgraf, L. M., Dazeley, R., Heintz, F., Howley, E., Irissappane, A. A., Mannion, P., Nowé, A., Ramos, G., Restelli, M., Vamplew, P., and Roijers, D. M. A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents and Multi-Agent Systems*, 36, 2022.
- Hendrickx, K., Perini, L., Van der Plas, D., Meert, W., and Davis, J. Machine learning with a reject option: A survey. *Machine Learning*, 113(5):3073–3110, 2024.
- Hong, Y., Zhang, H., Bao, J., Jiang, H., and Song, Y. Energy-based preference model offers better offline alignment than the Bradley-Terry preference model. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 23787–23804. PMLR, 2025.
- Horn, R. A. and Johnson, C. R. *Matrix Analysis*. Cambridge University Press, 1985.

- Jeon, H. J., Milli, S., and Dragan, A. Reward-rational (implicit) choice: A unifying formalism for reward learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 4415–4426, 2020.
- Jin, C., Krishnamurthy, A., Simchowitz, M., and Yu, T. Reward-free exploration for reinforcement learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4870–4879. PMLR, 2020.
- Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. *Mathematics of Operations Research*, 48(3):1496–1521, 2023.
- Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.
- Lattimore, T. and Szepesvári, C. *Bandit Algorithms*. Cambridge University Press, 2020.
- LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., and Huang, F. J. A tutorial on energy-based learning. *Predicting Structured Data*, 1(0), 2006.
- Lee, K., Smith, L., Dragan, A., and Abbeel, P. B-Pref: Benchmarking preference-based reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021a.
- Lee, K., Smith, L. M., and Abbeel, P. Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6152–6163. PMLR, 2021b.
- Luce, R. D. *Individual Choice Behavior*, volume 4. Wiley, 1959.
- Mu, N., Luan, Y., and Jia, Q.-S. Preference-based multi-objective reinforcement learning. *IEEE Transactions on Automation Science and Engineering*, 2025.
- Munos, R., Valko, M., Calandriello, D., Azar, M. G., Rowland, M., Guo, Z. D., Tang, Y., Geist, M., Mesnard, T., Fiegel, C., Michi, A., Selvi, M., Girgin, S., Momchev, N., Bachem, O., Mankowitz, D. J., Precup, D., and Piot, B. Nash learning from human feedback. In *International Conference on Machine Learning (ICML)*, 2024.
- Ng, A. Y. and Russell, S. Algorithms for inverse reinforcement learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 663–670. Morgan Kaufmann, 2000.
- Nocedal, J. and Wright, S. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, 2006.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 27730–27744, 2022.
- Pan, A., Bhatia, K., and Steinhardt, J. The effects of reward misspecification: Mapping and mitigating misaligned models. In *International Conference on Learning Representations (ICLR)*, 2022.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.
- Plackett, R. L. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2):193–202, 1975.
- Qiu, S., Zhang, D., Yang, R., Lyu, B., and Zhang, T. Traversing pareto optimal policies: Provably efficient multi-objective reinforcement learning. *arXiv preprint arXiv:2407.17466*, 2024.
- Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 53728–53741, 2023.
- Rao, P. V. and Kupper, L. L. Ties in paired-comparison experiments: A generalization of the Bradley-Terry model. *Journal of the American Statistical Association*, 62(317): 194–204, 1967.
- Rojers, D. M., Vamplew, P., Whiteson, S., and Dazeley, R. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48:67–113, 2013.

- Roy, B. and Vincke, P. Relational systems of preference with one or more pseudo-criteria: Some new concepts and results. *Management Science*, 30(11):1323–1335, 1984.
- Sadigh, D., Dragan, A., Sastry, S., and Seshia, S. Active preference-based learning of reward functions. In *Robotics: Science and Systems*, 2017.
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- Sen, A. K. *Collective Choice and Social Welfare*. Elsevier, 1970.
- Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. Deterministic policy gradient algorithms. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 32 of *Proceedings of Machine Learning Research*, pp. 387–395. PMLR, 2014.
- Siththaranjan, A., Laidlaw, C., and Hadfield-Menell, D. Distributional preference learning: Understanding and accounting for hidden context in RLHF. In *International Conference on Learning Representations (ICLR)*, 2024.
- Snedecor, G. W. *Statistical Methods*. Iowa State University Press, 1967.
- Sorensen, T., Moore, J., Fisher, J., Gordon, M. L., Miresghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., Althoff, T., and Choi, Y. Position: A roadmap to pluralistic alignment. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46280–46302. PMLR, 2024.
- Sutton, R. S. The reward hypothesis. <http://incompleteideas.net/rlai.cs.ualberta.ca/RLAI/rewardhypothesis.html>, 2004.
- Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT Press, 2018.
- Thurstone, L. L. A law of comparative judgment. *Psychological Review*, 34(4):273–286, 1927.
- Towers, M., Kwiatkowski, A., Balis, J., De Cola, G., Deleu, T., Goulão, M., Andreas, K., Krimmel, M., KG, A., Perez-Vicente, R., Terry, J., Pierré, A., Schulhoff, S., Tai, J. J., Tan, H., and Younis, O. G. Gymnasium: A standard interface for reinforcement learning environments. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025.
- Tsoukias, A. and Vincke, P. A new axiomatic foundation of partial comparability. *Theory and Decision*, 39(1): 79–114, 1995.
- von Neumann, J. and Morgenstern, O. *Theory of games and economic behavior*. Princeton University Press, 1947.
- Wang, R., Du, S., Yang, L., and Salakhutdinov, R. R. On reward-free reinforcement learning with linear function approximation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 17816–17826, 2020.
- Wang, Z., Bukharin, A., Delalleau, O., Egert, D., Shen, G., Zeng, J., Kuchaiev, O., and Dong, Y. Helpsteer2-preference: Complementing ratings with preferences. In *International Conference on Learning Representations (ICLR)*, 2025.
- Wirth, C., Akrou, R., Neumann, G., and Fürnkranz, J. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18(136): 1–46, 2017.
- Yang, L. and Wang, M. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10746–10756. PMLR, 2020.
- Zhang, W., Zhou, D., and Gu, Q. Reward-free model-based reinforcement learning with linear function approximation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 1582–1593, 2021.
- Zhang, Y., Zhang, G., Wu, Y., Xu, K., and Gu, Q. Beyond bradley-terry models: A general preference model for language model alignment. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 76939–76965. PMLR, 2025.
- Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J., and Wen, J. A survey of large language models. *Frontiers of Computer Science*, 2026.
- Zhao, Y., Escamill, J. E. A., Lu, W., and Wang, H. RA-PbRL: Provably efficient risk-aware preference-based reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 60835–60871, 2024.
- Zitzler, E. *Evolutionary Algorithms for Multiobjective Optimization: Methods and Applications*, volume 63. 1999.

A. Extended Related Works

In this appendix, we summarize the relevant literature, focusing on PbRL, rationality models alternative to BT, and proposed methods for handling non-preference.

Multi-Objective Reinforcement Learning. MORL addresses the challenge of optimizing multiple, often conflicting, criteria to discover a set of Pareto-optimal policies. Recent research has emphasized the controllability of this frontier. Qiu et al. (2024) investigates how a specific optimization target influences the *steerability* of the resulting policies, proposing a two-stage approach for finite *multi-objective MDPs* (MOMDP), ensuring a broad coverage of the Pareto frontier. More recently, Chen et al. (2026) introduced a reward-free perspective to MORL, by utilizing *reward-free reinforcement learning* (Jin et al., 2020) as an auxiliary task to enhance learning in continuous MOMDPs, providing also an algorithm that returns a weight-conditioned⁶ policy that can adapt to different weights at inference time. The CbRL setting has the same learning goal, i.e., recovering the Pareto frontier. However, it operates under significantly weaker signals, namely pairwise trajectory-level comparisons instead of the dense, multi-dimensional instantaneous rewards of MORL. This leads to an inherently higher sample complexity, which we aim to minimize through CbRL. Towards lowering this, strategies coming from active learning and preference elicitation (Sadigh et al., 2017; Bıyık et al., 2020) might be used to strategically select the most informative trajectory pairs, ensuring that the learning process remains feasible for human experts.

Preference-Based Reinforcement Learning. Preference-based Reinforcement Learning (Förnkrantz et al., 2012) was first formalized to integrate the *learning from preferences* and RL fields, and has since gained renewed attention with the rise of LLMs (Zhao et al., 2026). PbRL revolves around the mathematical framework of the *Markov decision process with preferences* (MDPP, Wirth et al., 2017), combining an MDP\R with a probability distribution ρ to model the probability of observing a clear preference given two trajectories. In this work, we take inspiration from the MDPP to formalize its multi-objective counterpart, the *Markov decision process with comparisons* (MDPC). Indeed, our framework combines the MDP\R with a probability distribution ρ defined, in our case, over a set of *four* possible comparison feedback classes, considering also *indifference* and *incomparability*. There are two main approaches to PbRL. The first, known as *reward learning* (Christiano et al., 2017), comprises two steps: (i) inferring a reward model from observed preferences and (ii) learning an optimal policy via standard RL methods. The second, *direct policy optimization* (DPO, Rafailov et al., 2023), directly works on the policy space. In this work, we focus on the first approach and propose a method to infer a multi-objective reward model from comparisons. Recently, a preliminary approach to PbRL in a multi-objective setting has been proposed (Mu et al., 2025). This work, however, does not formalize the concept of incomparability, instead eliciting standard PbRL preference feedback under a weighting of the objectives provided to the expert alongside the two trajectories. In our opinion, this approach trades the use of a convenient rationality model for an additional burden on the human expert. Indeed, eliciting a preference under a specific numerical scalarization of the problem requires not only that the expert combine the utilities but that they can also quantify such utilities in the first place. For this reason, we propose a novel framework to explicitly model incomparability and avoid numerical reasoning on behalf of the human expert.

Limitations of PbRL from the Social Choice Theory Perspective. The standard assumption in PbRL that human preferences can be effectively represented by a scalar utility function has recently come under scrutiny. Chakraborty et al. (2024) derives a fundamental impossibility result, showing that a single scalar representation is insufficient for capturing the heterogeneity of preferences across a diverse set of users. Learning a single scalar reward can thus lead to a *representational failure*, where the reward model aligns with a majority group of users while ignoring the conflicting preferences of the minority. In the direction of overcoming this, Munos et al. (2024) introduced *Nash learning from human feedback* to align with diverse preferences by learning the Nash equilibrium of the preference model. This can be seen as a shift of the paradigm towards a pluralistic alignment, emphasizing the need for an agent that can represent and align with a multitude of human values simultaneously, rather than converging to a single, potentially biased average (Sorensen et al., 2024). Shifting from scalar to vector-based rewards allows accounting for the inherent multi-objective nature of human judgment.

Alternative Rationality Models. The most widely considered rationality model in the PbRL literature is the Bradley-Terry (BT, Bradley & Terry, 1952) model, which models the probability of observing a preference as the sigmoid of the difference in utility of the two trajectories. The *Plackett-Luce* (Plackett 1975; Luce 1959) model generalizes BT to model *rankings* of objects. A classical alternative to BT is the *Thurstone-Mosteller* (Thurstone, 1927) model, which considers an additive Gaussian noise (Siththaranjan et al., 2024) in the preference generation process. In recent years, novel rationality models have been proposed to tackle specific limitations of BT. A first approach directly models the expert’s rationality, i.e., the

⁶We use the term “weight” to refer to the scalarization weight of the multiple objectives rather than the term “preference” commonly used in MORL to avoid ambiguity.

determinism of their decisions, via a rationality parameter that rescales each trajectory’s utility (Jeon et al., 2020). Then, the *infinite preference model* (Hong et al., 2025) tackles the problem of BT not guaranteeing a unique optimum to the MLE problem in *direct preference optimization* (Rafailov et al., 2023), proposing an approach based on *energy-based models* (LeCun et al., 2006). Finally, the *general preference optimization* (Zhang et al., 2025) addresses intransitive and/or cyclic preferences, proposing a model based on preference embeddings to capture complex structures that evade the limitations of BT’s scalar utility.

Handling Non-Preference Feedback and the Role of Abstension. The standard PbRL approach prescribes that the BT model characterizes only direct and inverse preferences, without accounting for *non-preference* feedback, i.e., indifference and incomparability. Several approaches have been proposed throughout the year to model *indifference* at varying levels of complexity. The most common one does not modify the BT model, defining the expert’s preference as a distribution over the two trajectories instead, and allowing for indifference by dividing this probability between the two trajectories (Christiano et al., 2017). An alternative approach modifies the BT model, characterizing the probability of observing an indifference as well as the probabilities of observing the two preferences. (Rao & Kupper, 1967) defines a “threshold of the sensory perception of the judge”, i.e., a value such that, if the difference in utility is lower than this value, the expert is unable to state a clear preference. (Davidson, 1970), instead, proposes the probability of indifference as inversely proportional to the extent to which the two trajectories are distinguishable. The concept of *incomparability* was first introduced in decision theory to model the choice of not taking a definite position w.r.t. a comparison (Roy & Vincke, 1984). Incomparability can be linked to abstention and the refusal to provide a definitive preference, which have been increasingly recognized as rational expert choices rather than technical failures, both in social choice theory and in machine learning (see, e.g., Hendrickx et al., 2024, which discusses reject options in machine learning). (Sen, 1970) formulates the inability to choose among alternatives as a rational response to a conflict between competing values. The standard approach in the presence of incomparability in PbRL is to discard the query and consider the sample as erroneous (Christiano et al., 2017). However, as argued in the context of robust learning, allowing the expert to “refuse” a query can improve the reliability and safety of the resulting model (Bartlett & Wegkamp, 2008). Incomparability can be linked to both MORL and order theory, indicating that its presence induces a partial order among trajectories (Förnkrantz et al., 2012). In recent years, the role of incomparability has again attracted the interest of the community, advancing the idea of exploiting it to learn a Pareto frontier of policies (Abdelkareem et al., 2022). Our work explicitly models this feedback, treating it as a primary signal to identify the presence of multiple, conflicting objectives.

B. Derivation of the Incomparability Score Function

In this appendix, we provide the complete derivation of Equation (10), i.e., the incomparability score function $h_{\parallel}$.

As noted in the main paper, Equation (4) requires incomparability to be the feedback with the highest probability of being generated along all the non-standard diagonals of the d -dimensional hyperspace. Our modeling approach for the incomparability score function is based on the idea that points along the non-standard diagonals, i.e., all the diagonals of the hypercube that can be formulated as xs for $x \in \mathbb{R}$ and $s \in \{-1, 1\}^d \setminus \{\mathbf{1}_d, -\mathbf{1}_d\}$, *grow farther* from the standard diagonal. Thus, we can quantify our incomparability score based on the distance of the point represented by $\delta(\tau, \tau')$ from its projection on the standard diagonal.

Let $\mathbf{t} := \frac{1}{\sqrt{d}} \cdot \mathbf{1}_d$ be a choice of vector such that the standard diagonal in $\mathbb{R}^d$ can be defined as the span of $\mathbf{t}$, i.e., $\mathcal{V} = \{\mathbf{v} \in \mathbb{R}^d : \mathbf{v} = c \cdot \mathbf{t}, c \in \mathbb{R}\}$. Then, denoting as $\mathbf{p} = \bar{c}\mathbf{t} \in \mathcal{V}$ the projection of $\delta(\tau, \tau')$ onto the standard diagonal, we choose to define the incomparability score function as:

$$h_{\parallel}(\tau, \tau') = \|\delta(\tau, \tau') - \mathbf{p}\|_2. \quad (14)$$

By definition of $\mathbf{p}$, it follows that $\mathbf{t}$ is orthogonal to the vector defined as $\mathbf{p} - \delta(\tau, \tau')$, i.e.:

$$\mathbf{t}^\top (\mathbf{p} - \delta(\tau, \tau')) = 0. \quad (15)$$

By applying the definition of $\mathbf{p}$, we can rewrite Equation (15) as $\bar{c} - \mathbf{t}^\top \delta(\tau, \tau') = 0$. Applying the definition of $\mathbf{t}$ and solving for $\bar{c}$, we derive that:

$$\bar{c} = \frac{1}{\sqrt{d}} \mathbf{1}_d^\top \delta(\tau, \tau'). \quad (16)$$

We can then rewrite the definition of $\mathbf{p}$ as:

$$\mathbf{p} = \bar{c}\mathbf{t} = \frac{1}{d} \mathbf{1}_d \mathbf{1}_d^\top \delta(\tau, \tau'). \quad (17)$$

Notice that we can define the quantity:

$$\bar{\delta}(\tau, \tau') := \frac{1}{d} \sum_{i=1}^d \delta_i(\tau, \tau') = \frac{1}{d} \mathbf{1}_d^\top \boldsymbol{\delta}(\tau, \tau'), \quad (18)$$

where $\delta_i(\tau, \tau')$ represents the i -th component of $\boldsymbol{\delta}(\tau, \tau')$, as the *mean* of the vector $\boldsymbol{\delta}(\tau, \tau')$. By plugging Equation (18) into Equation (17), we can rewrite $\mathbf{p}$ as:

$$\mathbf{p} = \bar{\delta}(\tau, \tau') \mathbf{1}_d, \quad (19)$$

i.e., as the d -dimensional vector having all components equal to $\bar{\delta}(\tau, \tau')$. Finally, by plugging Equation (19) into Equation (14) and applying the definition of the population standard deviation, we obtain:

$$h_{\parallel}(\tau, \tau') = \|\boldsymbol{\delta}(\tau, \tau') - \bar{\delta}(\tau, \tau') \mathbf{1}_d\|_2. \quad (20)$$

Simply adding a learnable parameter $\beta \in \mathbb{R}$ to Equation (20) results in Equation (10). Clearly, this summation does not alter the satisfiability of the desiderata in Definition 2.2, as β can be set to zero. Instead, having this additional degree of freedom allows us to handle situations in which the expert may present a slight irrationality, or may be more or less prone to evaluate two trajectories as incomparable.

C. Omitted Proofs

In this appendix, we present auxiliary lemmas and proofs of the results in the main paper.

C.1. Auxiliary Lemmas

Lemma C.1 (Lipschitz-Continuous Loss). *Let $\mathcal{D} = \{(\tau_i, \tau'_i, \circ_i)\}_{i=1}^N$ be a dataset of i.i.d. triples sampled from a distribution $\mathcal{P}_{\boldsymbol{\theta}^*} \in \Delta(\mathcal{T} \times \mathcal{T} \times \mathcal{C})$, defined such that $\mathcal{P}_{\boldsymbol{\theta}^*}((\tau, \tau', \circ)) = f(\circ|\tau, \tau'; \boldsymbol{\theta}^*)Q((\tau, \tau'))$, for every $\tau, \tau' \in \mathcal{T}$, $\circ \in \mathcal{C}$, and for some distribution $Q \in \Delta(\mathcal{T}^2)$. Let:*

$$\mathcal{L}(\boldsymbol{\theta}) := \mathbb{E}_{(\tau, \tau', \circ) \sim \mathcal{P}_{\boldsymbol{\theta}^*}} [-\log f(\circ|\tau, \tau'; \boldsymbol{\theta})], \quad (21)$$

and:

$$\hat{\mathcal{L}}(\boldsymbol{\theta}|\mathcal{D}) := \frac{1}{N} \sum_{i=1}^N -\log f(\circ_i|\tau_i, \tau'_i; \boldsymbol{\theta}). \quad (22)$$

Under Assumption 3.2, $\hat{\mathcal{L}}$ and $\mathcal{L}$ are L -Lipschitz continuous in L_∞ -norm with respect to $\boldsymbol{\theta}$, with:

$$L = 8\Lambda(\sqrt{d} + 2). \quad (23)$$

Proof. Let us denote as $\boldsymbol{\theta} = (\text{vec}(\mathbf{W}), \alpha, \beta)^\top \in \mathbb{R}^{dk+2}$, a vector representing the set of parameters $(\mathbf{W}, \alpha, \beta)$. We now compute the norm of the gradients of $h_\circ(\cdot|\boldsymbol{\theta})$ for every $\circ \in \mathcal{C}$ with respect to $\boldsymbol{\theta}$. By the Minkowski inequality (Hardy et al., 1964), we have that:

$$\|\nabla_{\boldsymbol{\theta}} h_\circ(\tau, \tau'|\boldsymbol{\theta})\|_F \leq \|\nabla_{\mathbf{W}} h_\circ(\tau, \tau'|\boldsymbol{\theta})\|_F + \|\nabla_{\alpha} h_\circ(\tau, \tau'|\boldsymbol{\theta})\|_F + \|\nabla_{\beta} h_\circ(\tau, \tau'|\boldsymbol{\theta})\|_F, \quad (24)$$

for every $\circ \in \mathcal{C}$. Thus, we can compute the gradients with respect to $\mathbf{W}$, α , and β individually and combine them.

With respect to $\mathbf{W}$, the gradients can be upper bounded as:

$$\|\nabla_{\mathbf{W}} h_{>}(\tau, \tau'|\boldsymbol{\theta})\|_F = \left\| \frac{1}{2d} \mathbf{1}_d (\phi(\tau) - \phi(\tau'))^\top \right\|_F \leq \frac{\Lambda}{2\sqrt{d}}, \quad (25)$$

$$\|\nabla_{\mathbf{W}} h_{<}(\tau, \tau'|\boldsymbol{\theta})\|_F = \|\nabla_{\mathbf{W}} h_{>}(\tau, \tau'|\boldsymbol{\theta})\|_F \leq \frac{\Lambda}{2\sqrt{d}},$$

$$\|\nabla_{\mathbf{W}} h_{=;}(\tau, \tau'|\boldsymbol{\theta})\|_F = 0,$$

where Equation (25) follows from Assumption 3.2. Regarding incomparability, let us first rewrite its definition as:

$$h_{\parallel}(\tau, \tau'|\boldsymbol{\theta}) = \sqrt{d} \sqrt{\frac{1}{d^2} \sum_{i,j \in [d]} \left((\mathbf{W}(\phi(\tau) - \phi(\tau')))_i - (\mathbf{W}(\phi(\tau) - \phi(\tau')))_j \right)^2} + \beta \quad (26)$$

$$= \sqrt{\frac{1}{d} \sum_{i,j \in [d]} \left((\mathbf{e}_i - \mathbf{e}_j)^\top \mathbf{W}(\phi(\tau) - \phi(\tau')) \right)^2} + \beta, \quad (27)$$

where, in Equation (27), e_i represents the i -th element of the d -dimensional canonical base, i.e., the vector obtained by replacing the i -th element of $\mathbf{0}_d$ with 1. Then, we can write the Frobenius norm of the gradient of $h_{\parallel}$ as:

$$\|\nabla_{\mathbf{W}} h_{\parallel}(\tau, \tau | \boldsymbol{\theta})\|_F \leq \frac{2}{\sqrt{d}} \frac{\sum_{i,j \in [d]} \|(e_i - e_j)^\top \mathbf{W}(\phi(\tau) - \phi(\tau'))\| \|(e_i - e_j)(\phi(\tau) - \phi(\tau'))^\top\|_F}{\sqrt{\sum_{i,j \in [d]} \left((e_i - e_j)^\top \mathbf{W}(\phi(\tau) - \phi(\tau')) \right)^2}} \quad (28)$$

$$\leq \frac{2}{\sqrt{d}} \sqrt{\sum_{i,j \in [d]} \|(e_i - e_j)(\phi(\tau) - \phi(\tau'))^\top\|_F^2} \quad (29)$$

$$\leq \frac{2}{\sqrt{d}} \sqrt{\sum_{i,j \in [d]: i \neq j} (2\Lambda)^2} \quad (30)$$

$$\leq 4\Lambda\sqrt{d}, \quad (31)$$

where Equation (29) is obtained by applying the Cauchy-Schwartz inequality, and Equation (30) follows by Assumption 3.2 and by observing that $e_i - e_j = \mathbf{0}_d$ if $i = j$.

With respect to α , we have:

$$\begin{aligned} \|\nabla_{\alpha} h_{>}(\tau, \tau' | \boldsymbol{\theta})\|_2 &= 0, & \|\nabla_{\alpha} h_{<}(\tau, \tau' | \boldsymbol{\theta})\|_2 &= 0, \\ \|\nabla_{\alpha} h_{=;}(\tau, \tau' | \boldsymbol{\theta})\|_2 &= 1, & \|\nabla_{\alpha} h_{\parallel}(\tau, \tau' | \boldsymbol{\theta})\|_2 &= 0. \end{aligned}$$

With respect to β , we have:

$$\begin{aligned} \|\nabla_{\beta} h_{>}(\tau, \tau' | \boldsymbol{\theta})\|_2 &= 0, & \|\nabla_{\beta} h_{<}(\tau, \tau' | \boldsymbol{\theta})\|_2 &= 0, \\ \|\nabla_{\beta} h_{=;}(\tau, \tau' | \boldsymbol{\theta})\|_2 &= 0, & \|\nabla_{\beta} h_{\parallel}(\tau, \tau' | \boldsymbol{\theta})\|_2 &= 1. \end{aligned}$$

By combining these results into Equation (24), we obtain that:

$$\begin{aligned} \|\nabla_{\boldsymbol{\theta}} h_{>}(\tau, \tau' | \boldsymbol{\theta})\|_F &\leq \frac{\Lambda}{2\sqrt{d}}, & \|\nabla_{\boldsymbol{\theta}} h_{<}(\tau, \tau' | \boldsymbol{\theta})\|_F &\leq \frac{\Lambda}{2\sqrt{d}}, \\ \|\nabla_{\boldsymbol{\theta}} h_{=;}(\tau, \tau' | \boldsymbol{\theta})\|_F &= 1, & \|\nabla_{\boldsymbol{\theta}} h_{\parallel}(\tau, \tau' | \boldsymbol{\theta})\|_F &\leq 4\Lambda\sqrt{d} + 1. \end{aligned} \quad (32)$$

We can now bound the norm of the gradient of $-\log f(\circ | \tau, \tau'; \boldsymbol{\theta})$, for every $\circ \in \mathcal{C}$:

$$\|\nabla_{\boldsymbol{\theta}} - \log f(\circ | \tau, \tau'; \boldsymbol{\theta})\|_F = \left\| \nabla_{\boldsymbol{\theta}} - h_{\circ}(\tau, \tau' | \boldsymbol{\theta}) + \log \left(\sum_{\circ' \in \mathcal{C}} \exp(h_{\circ'}(\tau, \tau' | \boldsymbol{\theta})) \right) \right\|_F \quad (33)$$

$$= \|\nabla_{\boldsymbol{\theta}} - h_{\circ}(\tau, \tau' | \boldsymbol{\theta})\|_F + \left\| \nabla_{\boldsymbol{\theta}} \log \left(\sum_{\circ' \in \mathcal{C}} \exp(h_{\circ'}(\tau, \tau' | \boldsymbol{\theta})) \right) \right\|_F \quad (34)$$

$$= \|\nabla_{\boldsymbol{\theta}} - h_{\circ}(\tau, \tau' | \boldsymbol{\theta})\|_F + \frac{\sum_{\circ' \in \mathcal{C}} \exp(h_{\circ'}(\tau, \tau' | \boldsymbol{\theta})) \|\nabla_{\boldsymbol{\theta}} h_{\circ'}(\tau, \tau' | \boldsymbol{\theta})\|_F}{\sum_{\circ' \in \mathcal{C}} \exp(h_{\circ'}(\tau, \tau' | \boldsymbol{\theta}))}. \quad (35)$$

By upper bounding the gradients of h with the highest upper bound of Equation (32), we finally get:

$$\|\nabla_{\boldsymbol{\theta}} - \log f(\circ | \tau, \tau'; \boldsymbol{\theta})\|_F \leq 8\Lambda\sqrt{d} + 2 \quad (36)$$

Finally, by observing that Equation (36) holds for every comparison label $\circ \in \mathcal{C}$ and that $\hat{\mathcal{L}}$ and $\mathcal{L}$ are a sample mean and an expectation, respectively, of $-\log f(\circ | \tau, \tau'; \boldsymbol{\theta})$ function, and recalling that $\|\mathbf{A}\|_{\infty} \leq \|\mathbf{A}\|_F$ for every matrix $\mathbf{A}$ (Horn & Johnson, 1985), it follows that both functions are L -Lipschitz, with $L = 8\Lambda(\sqrt{d} + 2)$. $\square$

Lemma C.2 (Subgaussian Scores). *Let $f : \mathbb{R}^d \rightarrow \Delta(\mathcal{C})$ be the MOBT model defined in Equation (6), and let $l : \mathbb{R} \rightarrow [0, +\infty)^4$ be the function defined as $l(\circ | \tau, \tau'; \boldsymbol{\theta}) = -\log f(\circ | \tau, \tau'; \boldsymbol{\theta})$. Under Assumption 3.2, l is ν^2 -subgaussian, with:*

$$\nu = 1 + 10R\Lambda\sqrt{dk}. \quad (37)$$

Proof. First, observe that, for every $\tau, \tau' \in \mathcal{T}$, $\boldsymbol{\theta} \in \Theta$, and $\circ \in \mathcal{C}$, it holds that:

$$| -\log f(\circ|\tau, \tau'; \boldsymbol{\theta}) | = \left| -h_{\circ}(\tau, \tau'|\boldsymbol{\theta}) + \log \left(\sum_{\circ' \in \mathcal{C}} \exp(h_{\circ'}(\tau, \tau'|\boldsymbol{\theta})) \right) \right| \quad (38)$$

$$\leq |h_{\circ}(\tau, \tau'|\boldsymbol{\theta})| + \left| \log \left(\sum_{\circ' \in \mathcal{C}} \exp(h_{\circ'}(\tau, \tau'|\boldsymbol{\theta})) \right) \right|. \quad (39)$$

Now, we bound each component independently. Starting from $h_{>}$, we have that:

$$|h_{>}(\tau, \tau'|\boldsymbol{\theta})| = \frac{1}{2} |\mathbf{1}_d^{\top} \mathbf{W}(\phi(\tau) - \phi(\tau'))| \quad (40)$$

$$\leq \frac{1}{2} \|\mathbf{1}_d^{\top} \mathbf{W}\|_2 \|\phi(\tau) - \phi(\tau')\|_2 \quad (41)$$

$$\leq \frac{1}{2} R\Lambda\sqrt{k}, \quad (42)$$

under Assumption 3.2, where Equation (41) follows from the Cauchy-Schwarz inequality. Clearly, the same holds for $h_{<}$, since $h_{<} = -h_{>}$. Then, it is straightforward to verify that $|h_{\pm}| \leq R$, under Assumption 3.2. Moving to incomparability, we have that:

$$|h_{\parallel}(\tau, \tau'|\boldsymbol{\theta})| = \sqrt{\frac{1}{d} \sum_{i,j \in \llbracket d \rrbracket} \left((e_i - e_j)^{\top} \mathbf{W}(\phi(\tau) - \phi(\tau')) \right)^2} + \beta \leq 3R\Lambda\sqrt{dk}. \quad (43)$$

Then, we bound the log-sum-exp term, observing that $h_{<} = -h_{>}$ and, consequently, $\exp(h_{>}(\tau, \tau'|\boldsymbol{\theta})) + \exp(h_{<}(\tau, \tau'|\boldsymbol{\theta})) \geq 1$, we have:

$$\left| \log \left(\sum_{\circ' \in \mathcal{C}} \exp(h_{\circ'}(\tau, \tau'|\boldsymbol{\theta})) \right) \right| \quad (44)$$

$$= \log \left(\exp(h_{>}(\tau, \tau'|\boldsymbol{\theta})) + \exp(h_{<}(\tau, \tau'|\boldsymbol{\theta})) + \exp(h_{\pm}(\tau, \tau'|\boldsymbol{\theta})) + \exp(h_{\parallel}(\tau, \tau'|\boldsymbol{\theta})) \right) \quad (45)$$

$$\leq \log \left(1 + \exp \left(\frac{1}{2} R\Lambda\sqrt{k} \right) + \exp(R) + \exp(3R\Lambda\sqrt{dk}) \right) \quad (46)$$

$$\leq \log \left(1 + 3 \exp(3R\Lambda\sqrt{dk}) \right) \quad (47)$$

$$\leq 1 + 9R\Lambda\sqrt{dk}, \quad (48)$$

where Equation (46) follows by combining the upper bounds in Equations (42) and (43) and the upper bound on $h_{\pm}$. Finally, we can derive that, for every comparison label $\circ \in \mathcal{C}$, it holds that:

$$| -\log f(\circ|\tau, \tau'; \boldsymbol{\theta}) | \leq \max_{\circ' \in \mathcal{C}} |h_{\circ'}(\tau, \tau'|\boldsymbol{\theta})| + \left| \log \left(\sum_{\circ' \in \mathcal{C}} \exp(h_{\circ'}(\tau, \tau'|\boldsymbol{\theta})) \right) \right| \quad (49)$$

$$\leq 1 + 10R\Lambda\sqrt{dk} =: \nu, \quad (50)$$

thus concluding the proof. $\square$

Lemma C.3. Let $f, g : \mathbb{R}^n \rightarrow \mathbb{R}$ be two continuously differentiable functions such that, for all $\mathbf{x} \in \mathbb{R}^n$, it holds that:

$$|f(\mathbf{x}) - g(\mathbf{x})| \leq c \quad \text{and} \quad \|\nabla g(\mathbf{x}) - \nabla f(\mathbf{x})\|_2 \leq d, \quad (51)$$

where $c, d, > 0$. Suppose that g is λ -strongly convex for some $\lambda > 0$. Let $\mathbf{x}_{\text{loc}}$ be a local minimum of f and let $\mathbf{x}_{\text{glob}}$ be a global minimum of f . Then, it holds that:

$$f(\mathbf{x}_{\text{loc}}) - f(\mathbf{x}_{\text{glob}}) \leq 2c + \frac{d^2}{2\lambda}. \quad (52)$$

Proof. Let us denote as $\mathbf{x}_g^*$ the unique minimizer of g , which exists since g is λ -strongly convex. Since we have defined $\mathbf{x}_{\text{loc}}$ as a local minimum of the continuously differentiable function f , it follows that $\nabla f(\mathbf{x}_{\text{loc}}) = 0$. Hence, by the gradient closeness assumption, we have that:

$$\|\nabla g(\mathbf{x}_{\text{loc}})\|_2 = \|\nabla g(\mathbf{x}_{\text{loc}}) - \nabla f(\mathbf{x}_{\text{loc}})\|_2 \leq d. \quad (53)$$

By combining this with the strong convexity of g , we have that:

$$g(\mathbf{x}_{\text{loc}}) - g(\mathbf{x}_g^*) \leq \frac{\|\nabla g(\mathbf{x}_{\text{loc}})\|_2^2}{2\lambda} \leq \frac{d^2}{2\lambda}. \quad (54)$$

From the pointwise bound, we observe that:

$$f(\mathbf{x}_{\text{loc}}) - f(\mathbf{x}_{\text{glob}}) \leq g(\mathbf{x}_{\text{loc}}) - g(\mathbf{x}_{\text{glob}}) + 2c. \quad (55)$$

Moreover, since $\mathbf{x}_g^*$ is the global minimizer of g , it holds that:

$$g(\mathbf{x}_{\text{loc}}) - g(\mathbf{x}_{\text{glob}}) \leq g(\mathbf{x}_{\text{loc}}) - g(\mathbf{x}_g^*) \leq \frac{d^2}{2\lambda}. \quad (56)$$

Finally, combining Equations (55) and (56), we get that:

$$f(\mathbf{x}_{\text{loc}}) - f(\mathbf{x}_{\text{glob}}) \leq 2c + \frac{d^2}{2\lambda}, \quad (57)$$

thus concluding the proof. $\square$

C.2. Omitted Proofs of the Main Paper

Proposition 2.1 (Impossible Convex NLL). *There exists no function $f : \mathbb{R}^d \rightarrow \Delta(\mathcal{C})$ that satisfies Definition 2.2, and such that $-\log f(\circ|\delta)$ is convex in δ for every $\circ \in \mathcal{C}$.*

Proof. We demonstrate this result by contradiction. Assume that $f : \mathbb{R}^d \rightarrow \Delta(\mathcal{C})$ satisfies Definition 2.2, and that $-\log f(\circ|\delta)$ is convex in δ for every $\circ \in \mathcal{C}$.

First, observe that $f(\circ|\delta) \in [0, 1]$ for every $\circ \in \mathcal{C}$ and $\delta \in \mathbb{R}^d$. Then, observing that $\sum_{\circ \in \mathcal{C}} f(\circ|\delta) = 1$ and $|\mathcal{C}| = 4$, we have that $\max_{\circ \in \mathcal{C}} f(\circ|\delta) \geq 1/4$, for every $\delta \in \mathbb{R}^d$. Considering now $-\log f(\circ|\delta)$, our previous observations translate to $-\log f(\circ|\delta) \geq 0$, for every $\circ \in \mathcal{C}$ and $\delta \in \mathbb{R}^d$, and $\min_{\circ \in \mathcal{C}} -\log f(\circ|\delta) \leq \log 4$, for every $\delta \in \mathbb{R}^d$. Let us now consider one of the non-standard diagonals of the d -dimensional hyperspace. For ease of notation, fix a $\mathbf{t} \in \{-1, 1\}^d \setminus \{\mathbf{1}_d, -\mathbf{1}_d\}$, and denote $f(\circ|x\mathbf{t})$, i.e., the points along such a diagonal, as $f(\circ|x)$, with $x \in \mathbb{R}$. Equation (4) corresponds to requiring, along the non-standard diagonal, that:

$$\lim_{x \rightarrow +\infty} -\log f(\circ|x) = c_1 \text{ and } \lim_{x \rightarrow -\infty} -\log f(\circ|x) = c_2, \quad (58)$$

where $c_1, c_2 \in [0, \log 4]$. Since, by assumption, the function is convex (and, thus, also continuous), the only way to satisfy these constraints is that $c_1 = c_2 =: c$ and enforcing that $-\log f(\circ|x)$ is the constant function $-\log f(\circ|x) = c$. Indeed, if $c_1 \neq c_2$ to fulfill convexity, either function $-\log f(\circ|x)$ changes concavity (and, thus, it is not convex) or one between c_1 and c_2 is infinite (thus, violating $c_1, c_2 \in [0, \log 4]$). Thus, it must be that:

$$f(\circ|x) = \exp(-c), \quad (59)$$

for every $x \in \mathbb{R}$. Now, recalling that $\sum_{\circ \in \mathcal{C}} f(\circ|x) = 1$, and considering Equation (59), the remaining terms must satisfy:

$$f(>|x) + f(<|x) + f(\asymp|x) = 1 - \exp(-c) =: K, \quad (60)$$

for some constant $K > 0$. By the desideratum of Equation 3, we have that $f(\asymp|x)$ must be the mode at $x = 0$, and:

$$\lim_{x \rightarrow \pm\infty} f(\asymp|x) < \exp(-c) \text{ and } f(\asymp|0) > \exp(-c). \quad (61)$$

Combining these requirements, we have that $f(\asymp|x)$ presents two changes in concavity. To respect this constraint, and have $-\log f(\asymp|x)$ be convex, then $f(\asymp|\cdot)$ must be a *bell-shaped* log-concave function.

Let us now define the additional quantity $S(x) := f(>|x) + f(<|x) = K - f(\asymp|x)$. Recalling that a function $g(\cdot)$ is log-concave, then it holds that $g''(\cdot) \leq (g'(\cdot))^2/g(\cdot)$, by requiring that both $f(>|x)$ and $f(<|x)$ must be log-concave, then $S(x)$ must be log-concave in regions of the space where $f(\asymp|x)$ is low. In such regions, i.e., those where $x \rightarrow \infty$, we require that $f(\circ|x) = \exp(c)$ is the mode, and thus $f(>|x)$ and $f(<|x)$ are log-concave and bounded.

Finally, we get that, for $f(>|x)$, $f(<|x)$, and $f(\asymp|x)$ to be all log-concave simultaneously and sum to K , they must all be constant. This, however, contradicts the desideratum of Equation 3, thus concluding the proof. $\square$

Lemma 3.1. *Choosing $\alpha, \beta \in \mathbb{R}$ such that $\alpha > 0$ and $\alpha > \beta$, the MOBT model complies with the desiderata of Definition 2.2.*

Proof. We demonstrate compliance with respect to one desideratum at a time. First, regarding *direct preference* $>$, we

observe that, along the standard diagonal, $\delta(\tau, \tau')$ is composed of d equal components, and can be rewritten as:

$$\delta(\tau, \tau') = t\mathbf{1}_d, \quad (62)$$

for some $t \in \mathbb{R}$. Thus, it follows that the standard deviation of $\delta(\tau, \tau')$ along the standard diagonal is zero. Compliance to the desideratum of Equation (1) simply follows by observing that, for $t \rightarrow +\infty$, $h_{>} \rightarrow \infty$, $h_{<} \rightarrow -\infty$, $h_{\leq} = \alpha$, and $h_{\parallel} = 0$. By taking the softmax of the four functions, the direct preference is the probability distribution's most probable outcome. A similar reasoning holds for *inverse preference* $<$, by considering $t \rightarrow -\infty$.

Then, regarding *indifference* $\asymp$, we observe that, as $\delta(\tau, \tau') \rightarrow \mathbf{0}_d$, we have that:

$$h_{>} = 0, h_{<} = 0, h_{\leq} = \alpha, h_{\parallel} = \beta. \quad (63)$$

Thus, having selected $\alpha > 0$ such that $\alpha > \beta$, it is straightforward to verify that $f(\asymp | \mathbf{0}_d) = \exp(\alpha)/(2 + \exp \alpha + \exp \beta)$ is the maximal element of the probability distribution, ensuring compliance with the desideratum of Equation (3).

Finally, regarding *incomparability* $\parallel$, we observe that the desideratum of Equation (4) prescribes that incomparability must be the mode of the probability distribution over $\mathcal{C}$ as $\delta(\tau, \tau')$ goes to infinity along any non-standard diagonal of the d -dimensional hyperspace. Let us consider the case of $d \geq 2$ as otherwise, if $d = 1$, the standard deviation is zero by definition. Fixing the direction along a non-standard diagonal, i.e., $\mathbf{t} \in \{+1, -1\}^d \setminus \{\mathbf{1}_d, -\mathbf{1}_d\}$, we observe that vector $\mathbf{t}$ comprises exactly k components equal to $+1$ and $d - k$ components equal to -1 . By denoting the utility difference along the non-standard diagonal as $c \cdot \mathbf{t}$, with $c \in \mathbb{R}$, we first observe that we can write, with a slight abuse of notation, the *clear preference* scores as:

$$h_{>}(c) = \frac{c(2k - d)}{2d} \quad \text{and} \quad h_{<}(c) = \frac{-c(2k - d)}{2d}. \quad (64)$$

Then, denoting the mean of vector $c\mathbf{t}$ as $\mu = \frac{c(2k-d)}{d}$, and observing that $(c \cdot t_i)^2 = c^2$ for every $t_i \in \mathbf{t}$, we can write the standard deviation of the vector as:

$$\text{std}(c\mathbf{t}) = \sqrt{\frac{1}{d} \sum_{i=1}^d c^2 - \mu^2} = \sqrt{c^2 - c^2 \frac{(2k-d)^2}{d^2}} = \frac{2|c|\sqrt{k(d-k)}}{d}. \quad (65)$$

Then, we can rewrite the incomparability score as:

$$h_{\parallel}(c) = 2|c|\sqrt{\frac{k(d-k)}{d}} + \beta. \quad (66)$$

Compliance with the desideratum of Equation (4) then corresponds to proving that Equation (66) is greater than Equation (64) as $c \rightarrow \infty$.

By combining Equations (64) and (66) we can write:

$$2|c|\sqrt{\frac{k(d-k)}{d}} - \frac{|c|(2k-d)}{2d} + \beta = |c| \left(2\sqrt{\frac{k(d-k)}{d}} - \frac{2k-d}{2d} \right) + \beta \quad (67)$$

$$\geq |c| \left(2\sqrt{\frac{d-1}{d}} - \frac{d-2}{2d} \right) + \beta \quad (68)$$

$$\geq |c| \left(\sqrt{2} - \frac{1}{2} \right) + \beta \quad (69)$$

where Equation (68) is obtained by observing that the minimum of $k(d-k)$ is $d-1$, attained at either $k = 1$ or $k = d-1$, and the maximum value of $(2k-d)/(2d)$ is $(d-2)/(2d)$, attained by $k = d-1$, and Equation (69) holds by observing that $d \geq 2$.

By observing that $\sqrt{2} - 0.5 > 0$ and recalling that β is a constant value, it clearly follows that:

$$\lim_{c \rightarrow \pm\infty} |c|(\sqrt{2} - 0.5) + \beta = +\infty, \quad (70)$$

thus showing the compliance of the MOBT model with respect to all the desiderata and concluding the proof. $\square$

Theorem 3.2. Let $\theta^* \in \Theta$ be the expert's parametrization, and $\mathcal{D} = \{(\tau_i, \tau'_i, \circ_i)\}_{i=1}^N$ be a dataset collected i.i.d. from an unknown distribution $\mathcal{P}_{\theta^*} \in \Delta(\mathcal{T} \times \mathcal{T} \times \mathcal{C})$ such that $\mathcal{P}_{\theta^*}(\tau, \tau', \circ) = f(\circ | \tau, \tau'; \theta^*)Q((\tau, \tau'))$, for some $Q \in \Delta(\mathcal{T}^2)$. Let $\hat{\theta}$ be the solution to Equation (11), and define $\mathcal{P}_{\hat{\theta}}$ as its induced distribution under Q . Under Assumptions 3.1 and 3.2, for

every $\delta \in (0, 1)$, it holds that:

$$D_{\text{KL}}(\mathcal{P}_{\theta^*} \parallel \mathcal{P}_{\hat{\theta}}) \leq \tilde{\mathcal{O}} \left(R \Lambda dk \sqrt{\frac{\log(1/\delta)}{N}} \right). \quad (12)$$

Proof. Under Assumption 3.2, let us consider an ϵ -cover Θ_ϵ of $\Theta \subseteq [-R, R]^{dk+2}$ in L_∞ -norm. We have that $|\Theta_\epsilon| \leq \left(\frac{6R}{\epsilon}\right)^{dk+2}$ (Lattimore & Szepesvári, 2020). Let now $\theta^* \in \arg \min_{\theta \in \Theta} \mathcal{L}(\theta)$ and $\hat{\theta} \in \arg \min_{\theta \in \Theta} \hat{\mathcal{L}}(\theta|\mathcal{D})$ as defined in Lemma C.1. Take $\tilde{\theta} \in \Theta_\epsilon$, we have for every $\delta \in (0, 1)$, with probability at least $1 - \delta$:

$$\mathcal{L}(\hat{\theta}) - \mathcal{L}(\theta^*) = \mathcal{L}(\hat{\theta}) - \mathcal{L}(\theta^*) \pm \hat{\mathcal{L}}(\hat{\theta}|\mathcal{D}) \quad (71)$$

$$\leq \mathcal{L}(\hat{\theta}) - \hat{\mathcal{L}}(\hat{\theta}|\mathcal{D}) + \hat{\mathcal{L}}(\theta^*|\mathcal{D}) - \mathcal{L}(\theta^*) \quad (72)$$

$$\leq 2 \sup_{\theta \in \Theta} |\hat{\mathcal{L}}(\theta|\mathcal{D}) - \mathcal{L}(\theta)| \quad (73)$$

$$= 2 \sup_{\theta \in \Theta} |\hat{\mathcal{L}}(\theta) - \mathcal{L}(\theta) \pm (\hat{\mathcal{L}}(\tilde{\theta}|\mathcal{D}) - \mathcal{L}(\tilde{\theta}))| \quad (74)$$

$$\leq 2 \sup_{\theta \in \Theta} |\hat{\mathcal{L}}(\theta|\mathcal{D}) - \mathcal{L}(\theta) - \hat{\mathcal{L}}(\tilde{\theta}|\mathcal{D}) + \mathcal{L}(\tilde{\theta})| + 2 |\mathcal{L}(\tilde{\theta}) - \hat{\mathcal{L}}(\tilde{\theta}|\mathcal{D})| \quad (75)$$

$$\leq 2 \sup_{\theta \in \Theta} \inf_{\tilde{\theta} \in \Theta_\epsilon} |\hat{\mathcal{L}}(\theta|\mathcal{D}) - \mathcal{L}(\theta) - \hat{\mathcal{L}}(\tilde{\theta}|\mathcal{D}) + \mathcal{L}(\tilde{\theta})| + 2 \sup_{\tilde{\theta} \in \Theta_\epsilon} |\hat{\mathcal{L}}(\tilde{\theta}|\mathcal{D}) - \mathcal{L}(\tilde{\theta})| \quad (76)$$

$$\leq 4L \sup_{\theta \in \Theta} \inf_{\tilde{\theta} \in \Theta_\epsilon} \|\theta - \tilde{\theta}\|_\infty + 2 \sup_{\tilde{\theta} \in \Theta_\epsilon} |\hat{\mathcal{L}}(\tilde{\theta}|\mathcal{D}) - \mathcal{L}(\tilde{\theta})| \quad (77)$$

$$\leq 4\epsilon L + 2\nu \sqrt{\frac{2 \log(|\Theta_\epsilon|/\delta)}{N}} \quad (78)$$

having observed that $\hat{\mathcal{L}}(\hat{\theta}|\mathcal{D}) \leq \hat{\mathcal{L}}(\theta^*|\mathcal{D})$, exploited the Lipschitzianity of $\mathcal{L}$ and $\hat{\mathcal{L}}$ in Equation (77) as in Lemma C.1, and having applied Hoeffding's inequality with the subgaussian constant derived in Lemma C.2 in Equation (78), recalling that the dataset is made of i.i.d. triples, and having performed a union bound over the cover Θ_ϵ . Now, recalling that $|\Theta_\epsilon| \leq \left(\frac{6R}{\epsilon}\right)^{dk+2}$ and choosing $\epsilon = \frac{\nu}{2L\sqrt{N}}$ which is a viable choice whenever $N \geq \left(\frac{\nu}{2LR}\right)^2$, we have, with probability at least $1 - \delta$:

$$\mathcal{L}(\hat{\theta}) - \mathcal{L}(\theta^*) \leq \frac{2\nu}{\sqrt{N}} + 2\sqrt{2}\nu \sqrt{\frac{(dk+2) \log\left(\frac{12R\sqrt{N}}{\nu}\right) + \log\left(\frac{1}{\delta}\right)}{N}} \quad (79)$$

$$\leq 5\nu \sqrt{\frac{(dk+2) \log\left(\frac{12RL\sqrt{N}}{\nu}\right) + \log\left(\frac{1}{\delta}\right)}{N}}. \quad (80)$$

By replacing the values of ν and L and recalling that $\mathcal{L}(\hat{\theta}) - \mathcal{L}(\theta^*) = D_{\text{KL}}(\mathcal{P}_{\theta^*} \parallel \mathcal{P}_{\hat{\theta}})$, we get the result. $\square$

Theorem 3.3. Let $\theta^* \in \Theta$ be the expert's parametrization, and $\mathcal{D} = \{(\tau_i, \tau'_i, \circ_i)\}_{i=1}^N$ be a dataset collected i.i.d. from an unknown distribution $\mathcal{P}_{\theta^*}$. Let $\hat{\theta}_{\text{loc}}$ be a local minimum of Equation (11), and define as $\mathcal{P}_{\hat{\theta}_{\text{loc}}}$ its induced distribution under the same $Q \in \Delta(\mathcal{T}^2)$. Under Assumptions 3.1 and 3.2, for every $\delta \in (0, 1)$, it holds that:

$$\begin{aligned} D_{\text{KL}}(\mathcal{P}_{\theta^*} \parallel \mathcal{P}_{\hat{\theta}_{\text{loc}}}) \\ \leq \tilde{\mathcal{O}} \left(R \Lambda dk \sqrt{\frac{\log(1/\delta)}{N}} + \mathcal{P}_{\theta^*}(\parallel) \Lambda d \sqrt{k} R \right), \end{aligned} \quad (13)$$

where $\mathcal{P}_{\theta^*}(\parallel) := \int_{\tau, \tau'} \mathcal{P}_{\theta^*}(d\tau, d\tau', \parallel)$ is the marginal probability of observing an incomparability.

Proof. Recalling the empirical negative log-likelihood for a given parameter $\theta \in \Theta$:

$$\hat{\mathcal{L}}(\theta|\mathcal{D}) := \frac{1}{N} \sum_{(\tau, \tau', \circ) \in \mathcal{D}} (-\log f(\circ|\tau, \tau'; \theta)), \quad (81)$$

we observe that we can decompose it as:

$$\hat{\mathcal{L}}(\boldsymbol{\theta}|\mathcal{D}) = \frac{1}{N} \sum_{(\tau, \tau', \circ) \in \mathcal{D}} (-\log f(\circ|\tau, \tau'; \boldsymbol{\theta}) \mathbb{1}\{\circ \neq \|\}) + \frac{1}{N} \sum_{(\tau, \tau', \circ) \in \mathcal{D}} (-\log f(\circ|\tau, \tau'; \boldsymbol{\theta}) \mathbb{1}\{\circ = \|\}) , \quad (82)$$

where $\mathbb{1}\cdot$ is the indicator function. Notice that the first term in Equation (82) is convex, since the standard deviation as a function of $\mathbf{u}(\tau|\boldsymbol{\theta}) - \mathbf{u}(\tau'|\boldsymbol{\theta})$ is convex (while its negative counterpart is not) and the log-sum-exp of convex functions is convex.

Let us now define the following function, for some $\lambda > 0$:

$$\tilde{\mathcal{L}}(\boldsymbol{\theta}|\mathcal{D}) := \frac{1}{N} \sum_{(\tau, \tau', \circ) \in \mathcal{D}} (-\log f(\circ|\tau, \tau'; \boldsymbol{\theta}) \mathbb{1}\{\circ \neq \|\}) + \frac{\lambda}{2} \|\boldsymbol{\theta}\|_2^2, \quad (83)$$

which is λ -strongly convex.

We can bound the absolute difference between $\hat{\mathcal{L}}(\boldsymbol{\theta}|\mathcal{D})$ and $\tilde{\mathcal{L}}(\boldsymbol{\theta}|\mathcal{D})$ as:

$$\left| \hat{\mathcal{L}}(\boldsymbol{\theta}|\mathcal{D}) - \tilde{\mathcal{L}}(\boldsymbol{\theta}|\mathcal{D}) \right| = \left| \frac{1}{N} \sum_{(\tau, \tau', \circ) \in \mathcal{D}} (-\log f(\circ|\tau, \tau'; \boldsymbol{\theta}) \mathbb{1}\{\circ = \|\}) - \frac{\lambda}{2} \|\boldsymbol{\theta}\|_2^2 \right| \quad (84)$$

$$\leq \left| \frac{1}{N} \sum_{(\tau, \tau', \circ) \in \mathcal{D}} (-\log f(\circ|\tau, \tau'; \boldsymbol{\theta}) \mathbb{1}\{\circ = \|\}) \right| + \left| \frac{\lambda}{2} \|\boldsymbol{\theta}\|_2^2 \right| \quad (85)$$

$$\leq \hat{\mathcal{P}}_{\boldsymbol{\theta}^*}(\|)(1 + 10R\Lambda\sqrt{dk}) + \frac{\lambda}{2}(dk + 2)R^2, \quad (86)$$

where the first term of Equation (86) is obtained by denoting the relative frequency of incomparability $\hat{\mathcal{P}}_{\boldsymbol{\theta}^*}(\|) = \sum_{i=1}^N \mathbb{1}\{\circ_i = \|\}/N$ in the dataset and applying Lemma C.2, and the second term is obtained by bounding $\|\boldsymbol{\theta}\|_2 \leq \sqrt{dk + 2}\|\boldsymbol{\theta}\|_\infty$.

Concerning the gradient, with the same reasoning as above and applying Lemma C.1, we derive that:

$$\|\nabla \hat{\mathcal{L}}(\boldsymbol{\theta}|\mathcal{D}) - \nabla \mathcal{L}(\boldsymbol{\theta}|\mathcal{D})\|_2 = \left\| \frac{1}{N} \sum_{(\tau, \tau', \circ) \in \mathcal{D}} (-\nabla \log f(\circ|\tau, \tau'; \boldsymbol{\theta}) \mathbb{1}\{\circ = \|\}) - \lambda \boldsymbol{\theta} \right\|_2 \quad (87)$$

$$\leq \hat{\mathcal{P}}_{\boldsymbol{\theta}^*}(\|)(8\Lambda\sqrt{d} + 2) + \lambda\sqrt{dk + 2}R. \quad (88)$$

Denoting, for conciseness of presentation, $C = \tilde{\mathcal{O}}\left(R\Lambda dk\sqrt{\frac{\log(1/\delta)}{N}}\right)$, we now derive:

$$\mathcal{L}(\hat{\boldsymbol{\theta}}_{\text{loc}}) - \mathcal{L}(\boldsymbol{\theta}^*) = \mathcal{L}(\hat{\boldsymbol{\theta}}_{\text{loc}}) - \mathcal{L}(\boldsymbol{\theta}^*) \pm \mathcal{L}(\hat{\boldsymbol{\theta}}) \quad (89)$$

$$\leq \mathcal{L}(\hat{\boldsymbol{\theta}}_{\text{loc}}) - \mathcal{L}(\hat{\boldsymbol{\theta}}) + C \pm \hat{\mathcal{L}}(\hat{\boldsymbol{\theta}}_{\text{loc}}|\mathcal{D}) \pm \hat{\mathcal{L}}(\hat{\boldsymbol{\theta}}|\mathcal{D}) \quad (90)$$

$$\leq \hat{\mathcal{L}}(\hat{\boldsymbol{\theta}}_{\text{loc}}|\mathcal{D}) - \hat{\mathcal{L}}(\hat{\boldsymbol{\theta}}|\mathcal{D}) + 2C \quad (91)$$

$$\leq 2\hat{\mathcal{P}}_{\boldsymbol{\theta}^*}(\|)(1 + 10R\Lambda\sqrt{dk}) + \frac{3}{2}\lambda(dk + 2)R^2 \quad (92)$$

$$\begin{aligned} &+ \frac{1}{2\lambda} \hat{\mathcal{P}}_{\boldsymbol{\theta}^*}(\|)^2 (8\Lambda\sqrt{d} + 2)^2 + 2C \\ &= 2\hat{\mathcal{P}}_{\boldsymbol{\theta}^*}(\|)(1 + 10R\Lambda\sqrt{dk}) + 2C \end{aligned} \quad (93)$$

$$+ \sqrt{3}\hat{\mathcal{P}}_{\boldsymbol{\theta}^*}(\|)(8\Lambda\sqrt{d} + 2)\sqrt{dk + 2}R \quad (94)$$

$$\leq 60\hat{\mathcal{P}}_{\boldsymbol{\theta}^*}(\|)\Lambda d\sqrt{k + 1}R + 2C \quad (94)$$

$$\leq 60\left(\mathcal{P}_{\boldsymbol{\theta}^*}(\|) + \sqrt{\frac{\log(2/\delta)}{2N}}\right)\Lambda d\sqrt{k + 1}R + 2C, \quad (95)$$

where Equation (90) is obtained by applying Theorem 3.2, Equation (91) is obtained by observing that we can repeat the derivation of Theorem 3.2 from Equation (73) onward, Equation (92) is obtained by applying Lemma C.3, observing that $\hat{\boldsymbol{\theta}}_{\text{loc}}$ and $\hat{\boldsymbol{\theta}}$ are a local and the global minimizer of $\hat{\mathcal{L}}$, respectively, that the two conditions in Equation (51) to apply the

lemma are satisfied by Equations (86) and (88), and by bounding $(a + b)^2 \leq 2a^2 + 2b^2$, Equation (93) is obtained by minimizing over λ , Equation (94) follows by observing that $d \geq 2$ otherwise no incomparabilities can be observed, and the problem is convex. Equation (95) follows by applying Hoeffding’s inequality, recalling that $|\mathcal{D}| := N$. $\square$

D. Numerical Validation

In this appendix, we provide a numerical validation of our MOBT model. This appendix is structured as follows. We first discuss the methodology of our experiments in Appendix D.1, covering the procedures for generating trajectories and comparison labels. We then report the specifications of the environment employed in our experiments in Appendix D.2, namely, GridWorld, MO-Hopper, and LQR. We discuss experimental results in Appendix D.3. We discuss the implementation details in Appendix D.4

D.1. Methodology and Experimental Pipeline

Given a dataset $\mathcal{D} = \{(\tau_i, \tau'_i, \circ_i)\}_{i=1}^N$, comprising N triples of two trajectories and a comparison label, and generated under the model of Equation (6), our goal is to the parameters $\hat{\theta} \in \mathbb{R}^{dk+2}$, i.e., the matrix $\hat{W} \in \mathbb{R}^{dk}$ and parameters $\hat{\alpha}, \hat{\beta} \in \mathbb{R}$, by minimizing the empirical NLL as in Equations (11). As demonstrated in Proposition 2.1, such a minimization cannot be achieved via convex optimization methods, due to the presence of incomparability feedback. As such, we resort to numerical optimization methods (Nocedal & Wright, 2006; Paszke et al., 2019).

The pipeline employed throughout the experiments of this paper comprises three stages: *trajectory generation*, *comparison label generation*, and *model optimization*. Let us now discuss each stage in detail.

Trajectory Generation. In order to generate dataset $\mathcal{D}$, we first generate a pool of trajectories by executing one or more behavioral policies in the target environment for fixed-horizon episodes. For each trajectory, we store the full sequence of states and actions, together with the true per-step multi-dimensional reward vectors generated by the environment. Crucially, this reward is never observed by our model and is stored solely to be provided to the synthetic expert for label-generation purposes.

The rationale behind this choice is to employ diverse policies to generate trajectories that *span the trajectory space*. The behavioral policies differ per environment:

- *GridWorld*: a single exploratory policy trained with implicit rewards to cover the state space as uniformly as possible.
- *MO-Hopper*: a suite of policies generated by first sampling a set of scalarization weights, then training for each weight in the true environment, and including in the suite both a near-optimal and a suboptimal policy.
- *LQR*: a single policy that generates trajectories by selecting a random feedback gain and producing a roll-out of fixed length. Notably, due to the linearity of the dynamics, this is sufficient to span the cost space.

We defer the definition of the environment-specific feature maps employed for each environment to Appendix D.2.

Comparison Label Generation. Once a pool of trajectories is generated, we select trajectory pairs by sampling from such a set. If the *number of labels per pair* is greater than one, we sample K different pairs, and query a *rational expert* to generate M labels for each trajectory pair, such that $N = KM$.

We generate labels by instantiating a synthetic expert adapted from `SimTeacher` (Lee et al., 2021a). We chose it as the basis for our synthetic expert as it natively provides support for parametrizable degrees of irrationality. We will evaluate varying degrees of irrationality concerning the probability of making a mistake, which we adapted from `SimTeacher` to provide a random label with probability $\epsilon \in [0, 1]$. Regarding comparison-label generation, we have modeled our synthetic expert to use our MOBT as its rationality model. When shown a pair of trajectories, we provide the expert also with the step-wise reward vectors for each trajectory, which the expert then employs to compute the score functions according to Equations (7) to (10), and finally generating a label by sampling from the distribution defined as the soft-max of the scores.

Model Optimization. Given the dataset $\mathcal{D} = \{(\tau_i, \tau'_i, \circ_i)\}_{i=1}^N$ and an environment-specific feature map $\phi : \mathcal{T} \rightarrow \mathbb{R}^k$, we minimize the empirical NLL of Equation (11) using the ADAM optimizer (Kingma & Ba, 2015). As demonstrated in Proposition 2.1, the objective is non-convex; we therefore rely on multiple random initializations to mitigate sensitivity to local optima. We report an additional experiment to quantify the spread of the training loss when using multiple initializations in Appendix D.3.

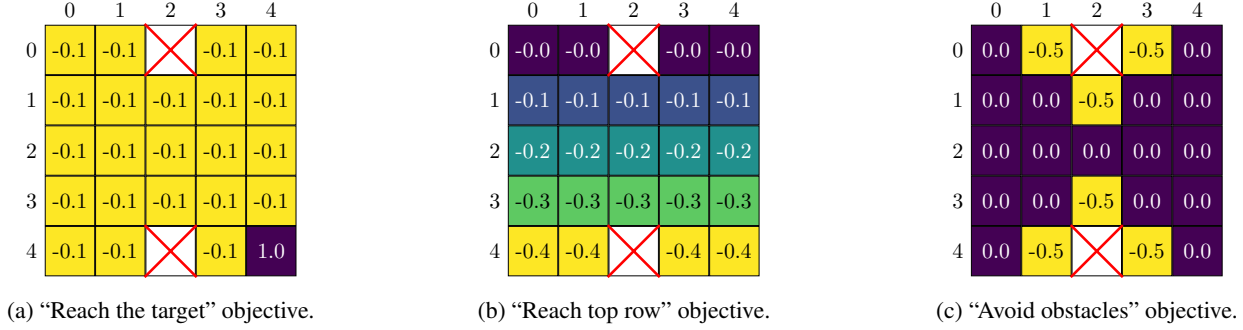


Figure 2. True 3-dimensional GridWorld reward matrices.

We reserve a randomly-sampled 20% of $\mathcal{D}$ as a held-out test set. The remaining 80% is used for training. After optimization, we evaluate the learned model by computing the KL divergence $D_{\text{KL}}(P_{\theta^*} \| P_{\hat{\theta}})$ on the test set, where P_{θ^*} is the distribution induced by the true expert parameters, and each reported result is the mean over 10 independent, seeded runs.

D.2. Experimental Environments Specification

We now provide detailed descriptions of the environments employed in our experiments.

GridWorld. We consider a custom 5×5 GridWorld with two obstacles placed in the topmost and bottommost cells of the middle column (column index 2). States are cell coordinates $(x, y) \in \mathcal{S} := \llbracket 0, 4 \rrbracket^2$, where $(0, 0)$ denotes the top-left cell and $(4, 4)$ the bottom-right cell. The agent starts each episode in the bottom-left cell $(4, 0)$. The action space is $\mathcal{A} := \{\text{U, D, L, R, S}\}$ (*up, down, left, right, stay*). The environment is implemented such that actions that would make the agent exit the grid or move into an obstacle are overwritten to make the agent stay in place.

At each step, the environment emits a three-dimensional reward vector $\mathbf{r}_{h+1} \in \mathbb{R}^3$ representing three objectives:

- (i) *Reach target*: $+1$ if $(x_{h+1}, y_{h+1}) = (4, 4)$, and -0.1 otherwise;
- (ii) *Reach top row*: $-0.1 \cdot x_{h+1}$, where x is the row index;
- (iii) *Avoid obstacles*: -0.5 if (x_{h+1}, y_{h+1}) is adjacent to an obstacle, 0 otherwise.

The true reward matrices for all three objectives are visualized in Figure 2. The feature map counts state visitation frequencies: $\phi(\tau) \in \mathbb{R}^{25}$, where the i -th entry is the number of time steps the agent occupies cell i .

MO-Hopper. We use the `mo-hopper-2obj-v5` environment from the MO-Gymnasium suite (Felten et al., 2023), the multi-objective extension of `Hopper-v5` (Towers et al., 2025). We consider two objectives: (i) forward velocity along the x -axis and (ii) height of the robot along the z -axis. The base environment provides an 11-dimensional observation vector at each step. We define the feature map as the episode-cumulative observation: $\phi(\tau) = \sum_{h=1}^H s_h \in \mathbb{R}^{11}$.

Linear Quadratic Regulator. We consider a one-dimensional discrete-time Linear Quadratic Regulator (Bemporad et al., 2002) with separated state and control costs, yielding a two-objective environment. Both the state $x_h \in \mathbb{R}$ and action $u_h \in \mathbb{R}$ are scalar. The dynamics follow:

$$x_{h+1} = a x_h + b u_h + \eta_h, \quad \eta_h \sim \mathcal{N}(0, 1), \quad (96)$$

and the per-step cost vector is:

$$\mathbf{c}_h = (q x_h^2, r u_h^2)^\top \in \mathbb{R}^2. \quad (97)$$

The system parameters are set to $a = 0.9$, $b = 1.0$, $q = 1.5$, and $r = 2.0$. The feature map returns the mean squared state and action over the episode: $\phi(\tau) = (\frac{1}{H} \sum_{h=1}^H x_h^2, \frac{1}{H} \sum_{h=1}^H u_h^2)^\top \in \mathbb{R}^2$. This choice is consistent with the quadratic structure of the cost, enabling exact recovery of q and r from the estimated utility weight matrix $\widehat{W}$. The Pareto frontier admits a closed-form solution under the LQR formalism, providing an exact reference for policy-level evaluation.

D.3. Experimental Results

We now report experiments to discuss several aspects of our model. First, we discuss our model’s ability to reconstruct a reward function that is *maximally compliant* with the expert. Second, we show why standard PbRL are not sufficient to

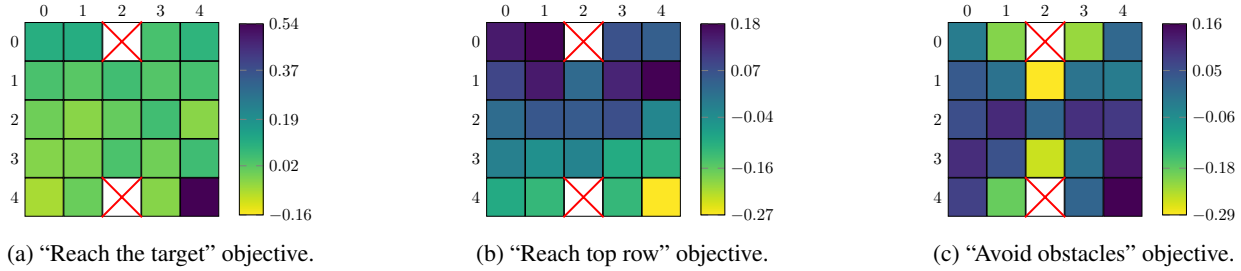


Figure 3. Estimated GridWorld reward matrices (mean over 10 runs, $K=1000$, $M=1$).

Pairs K	$M = 1$	$M = 5$	$M = 10$	Pairs K	$M = 1$	$M = 5$	$M = 10$
500	0.17 ± 0.01	0.08 ± 0.01	0.06 ± 0.01	1000	0.15 ± 0.02	0.06 ± 0.01	0.05 ± 0.01
1000	0.11 ± 0.01	0.06 ± 0.01	0.05 ± 0.01	5000	0.06 ± 0.01	0.05 ± 0.01	0.05 ± 0.01
5000	0.06 ± 0.01	0.05 ± 0.01	0.04 ± 0.01	10000	0.06 ± 0.01	0.05 ± 0.01	0.05 ± 0.01

(a) GridWorld Environment. (b) MO-Hopper Environment.

Table 2. Test KL divergence between estimated and true probabilities (10 runs, mean $\pm$ 95% C.I.).

address the additional complexity of the CbRL setting. Then, we evaluate our model’s ability to reconstruct the Pareto frontier of policies using the estimated reward function, demonstrating our model’s ability to solve multi-objective problems. Additionally, we provide a robustness analysis of the MOBT model to expert irrationality, considering the *probability of making a mistake*. Finally, we provide an analysis of the spread in terms of optimization quality induced by the non-convexity of the problem.

Reward Model Reconstruction. We evaluate our model’s ability to recover a meaningful reward function in a GridWorld environment with a dataset of $N = 1000$ samples. We report the inferred reward function in Figure 3. Clearly, our model approximates the true reward, although the reward range is shifted. This is expected, as our model is BT-based and thus translation-invariant.

Second, we evaluate the KL divergence on the test set in GridWorld and MO-Hopper for increasing numbers of pairs and samples per pair, i.e., K ranging from 500 to 10000 and M ranging from 1 to 10, respectively, and we report the results in Table 2. We observe that performance increases with the total number of comparisons N , i.e., K times M . This is expected, as both effectively counterbalance the epistemic uncertainty in the observed comparisons. We conjecture that increasing the number of pairs provides the same improvement as increasing the number of samples per pair (as can be seen in the secondary diagonal of Table 2a) only if the trajectories effectively span the feature space, as otherwise the model would be accurate only in a small portion of this space, with poor generalization.

Limitations of Standard PbRL. We now discuss why our MOBT model is necessary to address CbRL. As we thoroughly discussed throughout the paper, CbRL is the human-feedback counterpart to MORL. Solving a MORL problem requires handling multiple, possibly conflicting, objectives. The same requirement holds for CbRL. Crucially, PbRL approaches are defined to work under the *scalar utility* assumption. Indeed, the straightforward way to employ a PbRL approach in a CbRL problem is to either discard incomparable samples or treat them as indifferences. Thus, standard PbRL approaches are not sufficient to address CbRL settings.

To support this claim, we compare our MOBT model against several PbRL baselines, namely, standard Bradley-Terry (BT, Bradley & Terry, 1952), Thurstone-Mosteller (TM, Thurstone, 1927), Rao-Kupper (RK, Rao & Kupper, 1967), and Davidson (Davidson, 1970) models. Considering a train dataset of $N = 2000$ randomly sampled trajectory pairs, each labeled once, in the GridWorld environment, we adapted the dataset to fit the representational capacity of each model, by either *discarding* or *transforming* data. This corresponds to discarding both indifferences and incomparabilities for the BT and TM models, and incomparabilities for the RK and Davidson models. Additionally, we also consider the case in which both RK and Davidson models surrogate the problem by considering incomparabilities as if they were indifference labels.

First, we trained each model on its corresponding dataset. We report the recovered reward function for each baseline model in Figure 4.

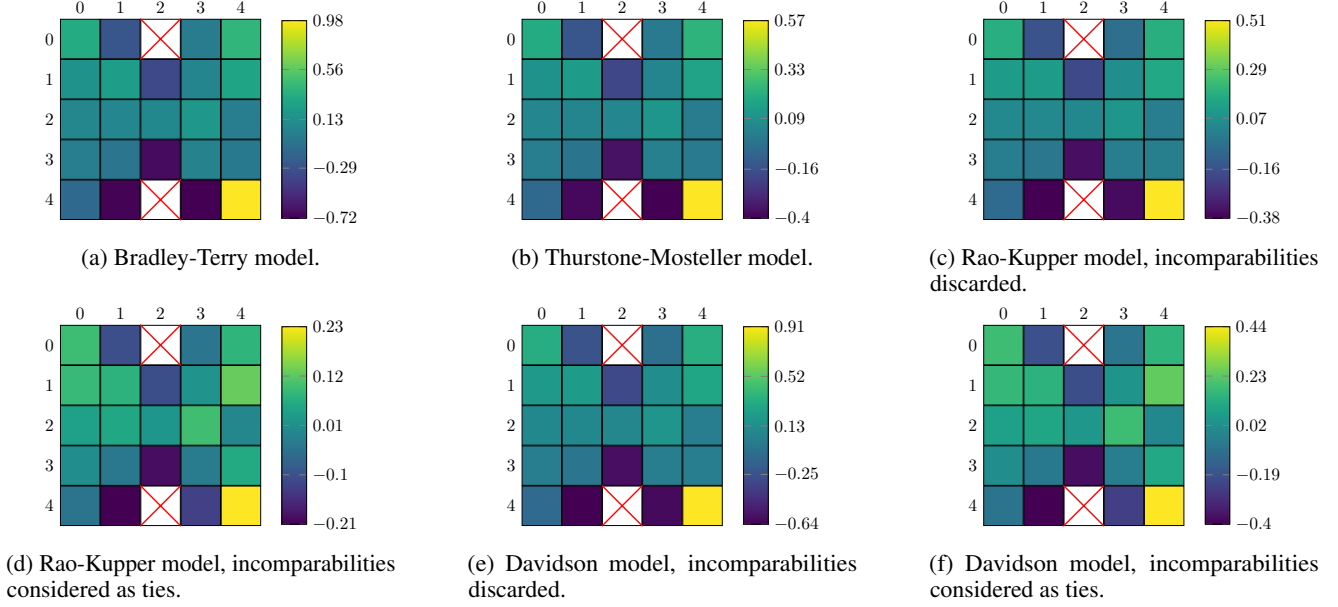


Figure 4. GridWorld reward matrix estimated using different single-objective rationality models (mean over 10 runs, each using 2000 trajectory pairs and 1 label per pair).

Model	D_{TV}
BT	0.6113 ± 0.0036
TM	0.6115 ± 0.0036
RK ($\parallel$ discarded)	0.4022 ± 0.0040
RK ($\parallel$ as $\asymp$)	0.4613 ± 0.0046
Davidson ($\parallel$ discarded)	0.4014 ± 0.0040
Davidson ($\parallel$ as $\asymp$)	0.4594 ± 0.0046
MOBT (<i>ours</i>)	0.0435 ± 0.0139

Table 3. Total Variation distance between the expert’s and model’s probability distribution over $\mathcal{C}$ computed on the test set (10 runs, mean $\pm$ 95% C.I.).

Clearly, all baseline models recover a reward function that minimizes the empirical NLL. However, due to their formulation, they are limited to learning a 1-dimensional reward, and are thus unable to generalize to a multi-objective problem.

Then, we evaluated each model on a different randomly sampled test set. The test set comprises all four comparison label types, implying that each model assigns zero probability to feedback classes it cannot handle. Crucially, this is necessary, as discarding samples at test time would be formally incorrect for two main reasons: (i) it would introduce a bias in the data-generating distribution and (ii) it would require knowing the label before it is observed. Due to their lower representational power, we cannot evaluate the KL divergence of the baseline models over the test set. Indeed, for any trajectory pair (τ, τ') for which $f(\parallel | \tau, \tau') > 0$, the KL divergence of the baselines takes value $+\infty$. This is due to the baselines-induced probability distributions over the comparison labels lacking the necessary support for incomparability (and, in the case of BT and TM models, also for indifference). This clearly highlights why standard PbRL models are insufficient to address CbRL problems. Nevertheless, to provide a meaningful, quantitative comparison of MOBT against the baselines, we consider the *total variation* (TV) distance D_{TV} , which always lies in $[0, 1]$ and remains finite even if the two distributions do not share the same support. Given two distributions $P, Q \in \Delta(\mathcal{X})$, we define their TV distance as $D_{TV}(P||Q) := (1/2) \sum_{x \in \mathcal{X}} |P(x) - Q(x)|$. In Table 3, we report the TV distance between the true probability distribution of the expert and the probability distribution over $\mathcal{C}$ induced by each model, averaged over 10 independent seeded runs.

We observe that our MOBT model effectively approximates the true expert comparison-generation probability distribution,

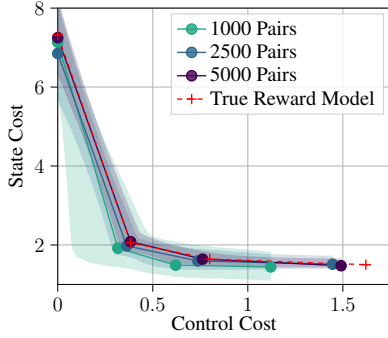


Figure 5. Pareto frontier in LQR (10 runs, mean $\pm$ 95% C.I.).

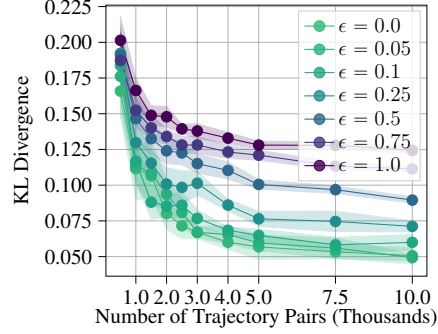


Figure 6. KL divergence for increasing *mistake probabilities* ϵ (10 runs, mean $\pm$ 95% C.I.).

achieving a TV distance that is an order of magnitude lower than that of the baselines. Recalling that the baseline models are *structurally unable* to predict incomparability labels, the evaluation via TV distance may underestimate their failure, as TV can only provide a bounded penalization in the case of support mismatch, thus underestimating the lack of representation power of the baseline models. Moreover, as expected, we observe that the more data is “unusable”, in the sense of the model not being able to learn from that data, the worse the performance, e.g., BT and TM compared to RK and Davidson discarding incomparabilities. Finally, we also observe that *mislabeling* the data, i.e., considering incomparabilities as if they were ties, does not provide an improvement down the line, resulting instead in a disproportionate probability distribution.

Regarding the behaviors that can be learned from the recovered reward functions, we notice that under the MOBT model, we can combine the three matrices under different weight vectors to traverse the Pareto frontier. The baselines, on the other hand, can only recover a single, scalar reward function, and are thus unable to deal with the multi-objective nature of CbRL.

Policy-Level Evaluation. We now discuss the ability of our MOBT model to recover a multi-dimensional reward that can be employed to reconstruct the Pareto frontier of policies. We consider a 1-dimensional *linear quadratic regulator* (LQR, Bemporad et al., 2002), as it provides a closed form for the Pareto frontier, enabling us to evaluate our MOBT model meaningfully. In LQR, the agent suffers two costs after every action: the *state cost*, which penalizes the distance of the agent from the target state 0, and the *control cost*, which penalizes the magnitude of the agent’s action.

For increasing values of K , with $M = 1$, we estimate parameter $\hat{\theta}$, compute estimated Pareto-optimal policies, and evaluate them in the true environment. We report the true and estimated Pareto frontiers in Figure 5. Qualitatively, we observe that, as K increases, our model induces policies that closely match the true Pareto frontier. We now consider the ratio between the *hypervolume* (HV, Zitzler, 1999; Felten et al., 2023) of the induced Pareto frontier and the HV of the true Pareto frontier. This ratio takes values in $[0, 1]$, with higher values representing better policy-level performance. In our experiment, it varies from $0.86(\pm 0.28)$ with 1000 samples to $0.96(\pm 0.21)$ with 5000 samples, demonstrating our model’s ability to solve a MORL problem from pairwise comparison feedback.

Robustness to Irrationality. We now evaluate the robustness of MOBT to expert irrationality in GridWorld, considering the *probability of making a mistake*, as modeled by SimTeacher (Lee et al., 2021a). To correctly cast it to CbRL, we redefine it as *returning a random label w.p. ϵ* . In Figure 6, we report the KL divergence between the distribution induced by the recovered model and the true, *noiseless* one, for $\epsilon \in [0, 1]$ and for different values of K , with $M = 1$. As expected, the KL divergence grows with ϵ , since MOBT cannot discern *mistakes* from rationally generated samples.

Multiple Initialization Analysis. As discussed in Section 3, in CbRL we have no guarantees of reaching the global optimum as a result of the optimization process. This is due to the non-convexity of the NLL shown in Proposition 2.1. Indeed, the presence of incomparabilities in the data introduces non-convexity in the optimization manifold. Nonetheless, as is customary in the Deep RL literature (Schulman et al., 2017; Haarnoja et al., 2018), convex first-order methods such as ADAM (Kingma & Ba, 2015) can be employed with good empirical performance, though with convergence guarantees only to local optima.

One common practical refinement to improve the overall performance is to employ an ensemble of predictors trained from randomized initializations. When using such an ensemble, the performance of the model then depends on the *spread of the optimization quality* of the predictors in the ensemble.

We now present an additional experiment to quantify the spread of the training loss when performing multiple initializations

Incomparability Ratio	Training Loss	Relative Standard Deviation
0.0	0.84 ± 0.0016	0.0010
0.2	1.02 ± 0.0014	0.0006
0.4	1.01 ± 0.0298	0.0147
0.6	0.91 ± 0.0600	0.0329

Table 4. Training loss (mean $\pm$ 95% C.I.) and relative standard deviation for $p \in \{0.0, 0.2, 0.4, 0.6\}$.

Hyperparameter	Default Value	Description
Learning rate	10^{-3}	Step size for the ADAM optimizer (Kingma & Ba, 2015).
Batch size	256	Number of comparison triples per gradient update.
Max. epochs	5000	Upper bound on training iterations; in practice, training terminates earlier via early stopping (see below).
Validation split	0.2	Fraction of the dataset held out for model evaluation (consistent with Section 4).
Early stopping patience	10	Number of epochs without improvement on the validation NLL before training is halted. It mitigates overfitting and reduces unnecessary computation once convergence is reached.

Table 5. Hyperparameters used for MOBT training.

in our setting, for varying degrees of incomparability in the dataset, i.e., the relative frequency of incomparability labels. As discussed in Theorem 3.3, the ratio of incomparability is a quantity that depends on both the dataset construction, i.e., the distribution $Q \in \Delta(\mathcal{T}^2)$ from which trajectory pairs are sampled, and on the problem instance itself, i.e., on the number of objectives and on how much they are in conflict with one another. Intuitively, a higher ratio of incomparability represents a more complex problem.

We consider a dataset of $N = 5000$ samples, on which we control the ratio of incomparabilities p , which represents the source of non-convexity. To quantify the *spread of optimization quality*, we employ the *relative standard deviation* (Snedecor, 1967), defined as the ratio of the standard deviation over the mean of a set of measurements. In Table 4, we report the training loss and relative standard deviation for varying values of $p \in \{0.0, 0.2, 0.4, 0.6\}$. We perform 100 independent optimizations per ratio, averaging over 10 seeds and 10 random initializations per seed.

First, we observe that the complexity of the optimization landscape increases with p , which we conjecture is due to the higher number of local optima. Second, the overall "spread" in the quality remains within an acceptable range, even when p reaches 60%. This can be interpreted as a consistent optimization performance during training.

D.4. Implementation Details

All experiments are implemented in Python 3.10 using PyTorch (Paszke et al., 2019). The codebase is organized into three modules: (i) `expert`, implementing the synthetic MOBT-based label generator; (ii) `optimizer`, wrapping the ADAM-based NLL minimizer; and (iii) `environment`, providing the GridWorld, MO-Hopper, and LQR environments. Experiment configurations are specified via `.yaml` files covering the four pipeline stages: policy generation, trajectory generation, label generation, and model evaluation. The LQR experiment is additionally available as a self-contained Jupyter notebook. A `README.md` with installation instructions and bash scripts for reproducing all results is provided in the supplementary code. All experiments have been conducted on a 10-core ARM CPU with 16GB of RAM.

Hyperparameters. Table 5 reports the default hyperparameters used to fit the MOBT model in the experiments. Notably, given the non-convexity result of Proposition 2.1, early stopping serves a dual purpose: (i) it acts as an implicit regularizer and (ii) it avoids unnecessary computation after the optimizer has effectively settled into a local minimum. The patience of 10 epochs was chosen empirically to balance convergence speed and stability across the three environments. The default learning rate of 10^{-3} is the standard recommended value for ADAM and was not tuned per environment.