PRELIMINARY PREPRINT VERSION: DO NOT CITE
The AAAI Digital Library will contain the published
version some time after the conference.

# Inferring Lexicographically-Ordered Rewards from Preferences

**Alihan Hüyük,[1] William R. Zame,[2] Mihaela van der Schaar[1 2 3]**

[1]University of Cambridge, [2]University of California, Los Angeles, [3]The Alan Turing Institute
ah2075@cam.ac.uk, zame@econ.ucla.edu, mv472@cam.ac.uk

## Abstract

Modeling the preferences of agents over a set of alternatives is a principal concern in many areas. The dominant approach has been to find a single reward/utility function with the property that alternatives yielding higher rewards are preferred over alternatives yielding lower rewards. However, in many settings, preferences are based on multiple—often competing—objectives; a single reward function is not adequate to represent such preferences. This paper proposes a method for inferring multi-objective reward-based representations of an agent's observed preferences. We model the agent's priorities over different objectives as entering *lexicographically*, so that objectives with lower priorities matter only when the agent is indifferent with respect to objectives with higher priorities. We offer two example applications in healthcare—one inspired by cancer treatment, the other inspired by organ transplantation—to illustrate how the lexicographically-ordered rewards we learn can provide a better understanding of a decision-maker's preferences and help improve policies when used in reinforcement learning.

## Introduction

Modeling the preferences of agents over a set of alternatives plays an important role in many areas, including economics, marketing (Singh, Hansen, and Gupta 2005), politics (Bräuninger, Müller, and Stecker 2016), and healthcare (Mühlbacher and Johnson 2016). One common approach to modeling preferences is to find a *utility function* over alternatives with the property that alternatives with higher utility are preferred over the ones with lower utility. This approach has been studied extensively in the machine learning (ML) literature as well—although the ML literature uses the term *reward function* rather than utility function. In reinforcement learning particularly, inferring reward function from the observed behavior of an agent (viz. inverse reinforcement learning) has proved an effective method of replicating their policy (Ng and Russell 2000; Abbeel and Ng 2004). Moreover, as Brown et al. (2019); Brown, Goo, and Niekum (2019) have recently shown, if reward functions are inferred from the preferences of an expert instead, then policies can even be improved.

In many circumstances, agent behavior is based on multiple—often competing—objectives. Healthcare in particular is replete with such circumstances. In treating cancer

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

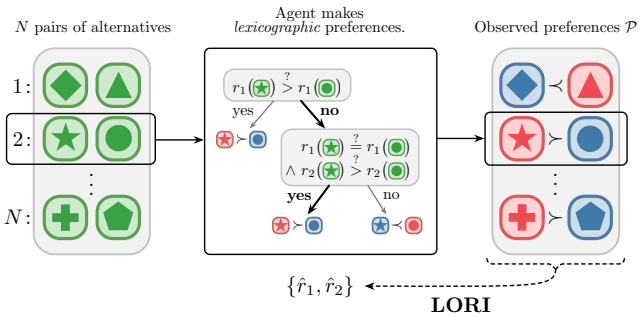



Figure 1: *Setting of LORI.* We model the preferences of agents lexicographically through reward functions $r_1, r_2, ..., r_k$ while allowing for errors and indifference to small differences. LORI aims to infer these reward functions from a dataset of observed preferences over pairs of alternatives.

(and many other diseases), clinicians aims for treatment that is the most effective but also has the fewest harmful side-effects. This is especially true in radiation therapy (Wilkens et al. 2007; Jee, McShan, and Fraass 2007), where high doses are needed to be effective against tumors but also cause damage to surrounding tissue. In organ transplantation, clinicians aim to make the best match between the organ and the patient but also to give priority to patients who have been waiting the longest and/or have the most urgent need (Coombes and Trotter 2005; Schaubel et al. 2009). In the allocation of resources in a pandemic, clinicians hope to minimize the spread of infection but also to safeguard the most vulnerable populations (Koyuncu and Erol 2010; Gutjahr and Nolz 2016). In these situations, and many others, the preferences of decision-makers reflect the priorities they assign to various criteria.

This paper provides a method for inferring multi-objective reward-based representations of a decision-maker's observed preferences. Such representations provide a better understanding of the decision-maker's preferences and promote reinforcement learning. We model priorities over different objectives as entering *lexicographically*, so that objectives that have lower priority matter only when the decision-maker is indifferent with respect to objective that have higher priority. While lexicographic ordering is certainly not the only way an agent might prioritize different objectives, it is a prevalent one: there is plenty of evidence showing that humans use

lexicographic reasoning when making decisions (Kohli and Jedidi 2007; Slovic 1975; Tversky, Sattah, and Slovic 1988; Colman and Stirk; Drolet and Luce 2004; Yee et al. 2007). In modeling priorities lexicographically, we take into account that the decision-maker may be indifferent to "small" differences. We allow for the possibility that the decision-maker may be an expert consultant, a treating clinician, the patient, or a population of experts, clinicians or patients. As we shall see, these considerations shape our model.

**Contributions** We introduce a new stochastic preference model based on multiple lexicographically-ordered reward functions. We formulate Lexicographically-Ordered Reward Inference (LORI) as the problem of identifying such models from observed preferences of an agent and provide a Bayesian algorithm to do so. We offer two examples—one inspired by cancer treatment, the other inspired by organ transplantation—to illustrate how the lexicographically-ordered reward functions we learn can be used to interpret and understand the preferences of a decision-maker, and demonstrate how inferring lexicographically-ordered reward functions from preferences of an expert can help improve policies when used in reinforcement learning.

## Related Work

As a method for learning reward-based representations of lexicographic preferences, LORI is related to *preference learning* (PL), and as a tool for discovering lexicographically-ordered reward functions for reinforcement learning purposes, it is related to *inverse reinforcement learning* (IRL).

**Preference learning** Preference learning is the problem of finding a model that best explains a given set of observed preferences over a set of alternatives. It can be tackled either with an *ordinal* approach, where a binary preference relation between alternatives is learned directly (e.g. Boutilier et al. 2004), or with a *cardinal/numerical* approach, where such relations are induced through reward functions (e.g. Chu and Ghahramani 2005; Brochu, Freitas, and Ghosh 2008; Bonilla, Guo, and Sanner 2010; Salvatore et al. 2013).

Lexicographic preferences have been primarily considered from an ordinal perspective. Schmitt and Martignon (2006); Dombi, Imreh, and Vincze (2007); Yaman et al. (2008); Kohli, Boughanmi, and Kohli (2019) assume that each alternative has an unordered set of attributes (i.e. features in ML literature) and preferences are made by comparing those attributes lexicographically. They aim to learn in what order the attributes are compared. Kohli and Jedidi (2007); Jedidi and Kohli (2008) consider variations of this approach including binary-satisficing lexicographic preferences, which allows indifference when comparing attributes and relaxes the assumption that attributes are compared one at a time. Booth et al. (2010); Braüning and Hüllermeyer (2012); Liu and Truszczynski (2015); Fernandes, Cardoso, and Palacios (2016); Braüning et al. (2017); Fargier, Gimenez, and Mengin (2018) generalize this framework and learn lexicographic preference trees instead, where the priority order of attributes is not assumed to be static but allowed to change dynamically depending on the outcome of pairwise comparisons in higher-priority attributes. All of these methods are purely ordinal rather than cardinal.

We take a different approach to modeling lexicographic preferences. In our model, each alternative has an associated multi-dimensional reward and a preference relation over alternatives is induced by the standard lexicographic preference relation between their associated rewards. The goal of LORI is to infer the reward functions that determine the reward vector of each alternative. Remember that existing methods for inferring lexicographic relations assume the set of relevant attributes to be specified beforehand (except for their priority order, which need to be inferrred). Inferring the latent reward functions in our model can be conceptualized as learning those attributes from scratch.

**Inverse reinforcement learning** Given the demonstrated behavior of an agent, IRL aims to find a reward function that makes the demonstrated behavior appear optimal (Abbeel and Ng 2004; Ramachandran and Amir 2007; Ziebart et al. 2008; Boularias, Kober, and Peters 2011; Levine, Popovic, and Koltun 2011; Finn, Levine, and Abbeel 2016). When the demonstrated behavior is in fact optimal, the learned reward function can guide (forward) reinforcement learning to reproduce optimal policies. However, agents do not always behave optimally according to the judgement of others—or even according to their own judgement. In that case, conventional IRL (followed by reinforcement learning) can, at best, lead to policies that mimic the demonstrated suboptimal behavior.

Preference-based IRL is an alternative approach to conventional IRL, where a reward function is inferred from the preferences of an expert over various demonstrations instead (Sugiyama, Meguro, and Minami 2012; Wirth, Fürnkranz, and Neumann 2016; Christiano et al. 2017; Ibarz et al. 2018). Recently, Brown et al. (2019) showed that this alternative preference-based approach can identify the intended reward function of the expert and lead to optimal policies even when the demonstrations provided are suboptimal. Brown, Goo, and Niekum (2019) showed that these expert preferences can be generated synthetically in scenarios where it is possible to interact with the environment.

Both conventional and preference-based IRL methods focus almost exclusively on inferring a single reward function to represent preferences. However, as we have discussed in the introduction, many important tasks are not readily evaluated in terms of a single reward function. Task representations that go beyond single reward functions has been considered in IRL. Most notably, Vazquez-Chanlatte et al. (2018) propose non-Markovian and Boolean specifications to describe more complex tasks, and Chou, Berenson, and Özay (2018); Scobee and Sastry (2020) infer the constraints that a task might have when a secondary goal is also provided. (We view tasks with constraints as the special case of lexicographically-prioritized objectives in which the reward function describing the secondary goal is only maximized when all constraints are satisfied.) To the best of our knowledge, LORI is the first reward inference method that can learn general lexicographically-ordered reward functions solely from observed preferences (as can be seen in Table 1).

Table 1: *Comparison with related work.* LORI is the only method that can infer lexicographically-ordered (lex.-ordered) reward functions solely from observed preferences. Preferences, demonstrations, features/attributes, and reward/utility functions are denoted by $\mathcal{P}$, $\mathcal{D}$, $\{\phi_i\}$ and $r$; the space of alternatives, states, actions, and state-action histories are denoted by $X$, $S$, $A$, and $H$.

| **Method** | **Prototype** | **Given** | **Inferred** |
|---|---|---|---|
| Ordinal PL | Boutilier et al. (2004) | $\mathcal{P}$ | $\succ$ |
| Ordinal lexicographic PL | Yaman et al. (2008) | $\mathcal{P}$, *Unordered* $\{\phi_i \in \mathbb{R}^X\}$ | *Lex.-ordered* $\{\phi_{i'} \in \mathbb{R}^X\}$ |
| Cardinal single-dimensional PL | Chu and Ghahramani (2005) | $\mathcal{P}$ | $r \in \mathbb{R}^X$ |
| Conventional IRL | Ziebart et al. (2008) | $\mathcal{D}$ | $r \in \mathbb{R}^{S \times A}$ |
| Preference-based IRL | Brown et al. (2019) | $\mathcal{P}$ | $r \in \mathbb{R}^{S \times A}$ |
| IRL w/ specifications | Vazquez-Chanlatte et al. (2018) | $\mathcal{D}$ | $r \in \{0,1\}^H$ |
| IRL w/ constraints | Scobee and Sastry (2020) | $\mathcal{D}$, $r_2 \in \mathbb{R}^{S \times A}$ | $r_1 \in \{0,1\}^{S \times A}$ |
| **LORI** | **Ours** | $\mathcal{P}$ | *Lex.-ordered* $\{r_i \in \mathbb{R}^X\}$ |

## Problem Formulation

We assume that we are given observations about the preferences of a decision-maker or a set of decision-makers in the form of a pair $\mathcal{P} = (X, n)$, where $X$ is a set of alternatives and $n : X \times X \to \mathbb{Z}_+$ is a function. We interpret $n(x^\star, x^\circ)$ as the *number of times* that alternative $x^\star$ was observed to be preferred to alternative $x^\circ$. Note that $n(x^\star, x^\circ)$ may be 0 because $x^\star$ was *never* observed to be preferred to $x^\circ$. We allow for the possibility that both $n(x^\star, x^\circ) > 0$ and $n(x^\circ, x\star) > 0$, either because we are observing the preferences of a single decision-maker who is not completely consistent or because we are observing the preferences of a *population* of decision-makers who do not entirely agree. Write $A = \{(x^\star, x^\circ) \in X \times X : n(x^\star, n^\circ) > 0\}$; this is the set of pairs $(x^\star, x^\circ)$ for which $x^\star$ is observed to be preferred to $x^\circ$ at least once. If $(x^\star, x^\circ) \in A$, we often write $x^\star \succ x^\circ$. (Note that we *do not* assume that $\succ$ is a preference relation in the sense used in economics; in particular, we *do not* assume that the preference relation is asymmetric or transitive.)

We seek to explain these observations in terms of an ordered set of $k$ reward functions $\{r_i \in \mathbb{R}^X\}_{i=1}^k$ (numbered so that $r_1$ is prioritized over $r_2$, $r_2$ is prioritized over $r_3$, and so on), in the sense that $x^\star \succ x^\circ$ tends to be observed if the reward vector $\boldsymbol{r}(x^\star) = \{r_1(x^\star), \ldots, r_k(x^\star)\}$ lexicographically dominates the reward vector $\boldsymbol{r}(x^\circ)$, in which case we write $\boldsymbol{r}(x^\star) >_{\text{lex}} \boldsymbol{r}(x^\circ)$. Formally, $\boldsymbol{r}(x^\star) >_{\text{lex}} \boldsymbol{r}(x^\circ)$ if and only if there exists $i \in \{1, \ldots, k\}$ such that $r_i(x^\star) > r_i(x^\circ)$ and $r_j(x^\star) = r_j(x^\circ)$ for all $j < i$. Because we allow for the possibility that $x^\star \succ x^\circ$ and $x^\circ \succ x^\star$, we incorporate randomness; we also allow for indifference in the presence of small differences. Our objective can be summarized as:

**Objective** Infer reward functions $\{r_i\}_{i=1}^k$ from the observed preferences $\mathcal{P}$ of a decision-maker.

## Lexicographically-ordered Reward Inference

A substantial literature views $x^\star \succ x^\circ$ as a *random event* and models the probability of this event in terms of a noisy comparison between the rewards of $x^\star$ and $x^\circ$. This idea is the foundation of the *logistic choice model* originated by McFadden (1973); recent ML work includes Brown et al. (2019); Brown, Goo, and Niekum (2019). Formally, given a reward function $r \in \mathbb{R}^X$ and a parameter $\alpha \geq 0$, this literature defines

$$\begin{aligned} \mathbb{P}(x^\star \succ x^\circ | r) &= \frac{e^{\alpha r(x^\star)}}{e^{\alpha r(x^\star)} + e^{\alpha r(x^\circ)}} \\ &= \frac{1}{1 + e^{-\alpha(r(x^\star) - r(x^\circ))}} \ . \end{aligned} \tag{1}$$

The parameter $\alpha$ represents the extent to which preference is *random*. (Randomness of the preference may reflect inconsistency on the part of the decision-maker, or reflect the aggregate preferences of some population of decision-makers—experts or clinicians or patients.) Note that if $\alpha = 0$ then preference is uniformly random; at the opposite extreme, as $\alpha \to \infty$ then preference becomes perfectly deterministic.

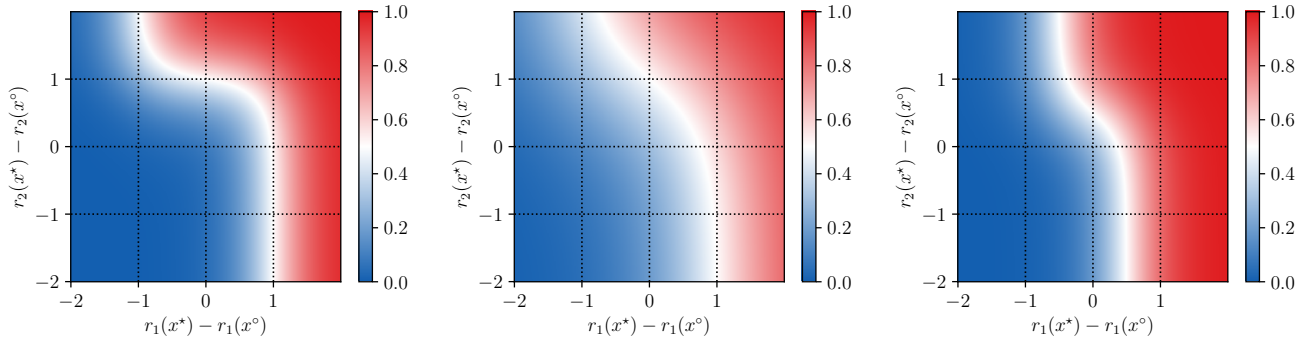
Given a reward function $r$, parameter $\alpha$, observed preferences $\mathcal{P} = (X, n)$ and alternatives $x^\star, x^\circ \in X$, we can ask how likely it is that $r, \alpha$ would generate the same relationship between $x^\star, x^\circ$ that we observe in $\mathcal{P}$. By definition, $n(x^\star, x^\circ)$ is the number of times that $x^\star$ is observed to be preferred to $x^\circ$ and $n(x^\circ, x^\star)$ is the number of times that $x^\circ$ is observed to be preferred to $x^\star$, so $N(x^\star, x^\circ) = n(x^\star, x^\circ) + n(x^\circ, x^\star)$ is the total number of times that $x^\star, x^\circ$ are observed to be compared. The probability that the preference generated by $r$ agrees with the observations in $\mathcal{P}$ with respect to $x^\star, x^\circ$ is

$$\begin{aligned} \mathbb{P}(\mathcal{P}_{x^\star, x^\circ} | r) = &\tbinom{N(x^\star, x^\circ)}{n(x^\star, x^\circ)} \\ &\times \mathbb{P}(x^\star \succ x^\circ | r)^{n(x^\star, x^\circ)} \\ &\times \mathbb{P}(x^\circ \succ x^\star | r)^{n(x^\circ, x^\star)} \ . \end{aligned} \tag{2}$$

Hence, the probability that the preference generated by $r$ agrees with $\mathcal{P}$ (i.e. the likelihood of $r$ with respect to $\mathcal{P}$) is

$$\begin{aligned} \mathcal{L}(r; \mathcal{P}) &= \mathbb{P}(\mathcal{P} | r) \\ &= \left[\textstyle\prod_{(x^\star, x^\circ) \in X \times X} \mathbb{P}(\mathcal{P}_{x^\star, x^\circ} | r)\right]^{1/2} \ . \end{aligned} \tag{3}$$

(The exponent $1/2$ is needed because each of the terms $\mathbb{P}(\mathcal{P}_{x^\star, x^\circ} | r)$ appears twice: once indexed by $(x^\star, x^\circ)$ and again indexed by $(x^\circ, x^\star)$.) If the "true" reward function is known/assumed to belong to a given family $R$, it can be inferred by finding the maximum likelihood estimate (MLE) $\hat{r} = \operatorname{argmax}_{r \in R} \mathcal{L}(r; \mathcal{P})$.




**(a) Baseline preferences** ($\alpha_1 = \alpha_2 = 3, \varepsilon_1 = \varepsilon_2 = 1$)

**(b) Less consistent preferences** ($\alpha_1 = \alpha_2 = 1.5, \varepsilon_1 = \varepsilon_2 = 1$)

**(c) Less tolerant preferences** ($\alpha_1 = \alpha_2 = 3, \varepsilon_1 = \varepsilon_2 = 0.5$)

Figure 2: *Probability distribution* $\mathbb{P}(x^\star \succ x^\circ | r_1, r_2)$ *over preferences induced by reward functions* $r_1$, $r_2$. Smaller $\alpha_i$'s lead to preferences that are less consistent while smaller $\varepsilon_i$'s lead to ones less tolerant of differences.

**Lexicographic setting** Adapting this probabilistic viewpoint to our context requires two changes. The first is that we make probabilistic comparisons for *multiple* reward functions. The second is that we allow for the possibility that the rewards $r(x^\star), r(x^\circ)$ may not be *exactly* equal but the difference between them may be regarded as negligible. We therefor begin by defining:

$$\begin{aligned}
\mathbb{P}(x^\star \succ_i x^\circ | r_i) &= \frac{1}{1 + e^{-\alpha_i(r_i(x^\star) - r_i(x^\circ) - \varepsilon_i)}} \,, \\
\mathbb{P}(x^\star \prec_i x^\circ | r_i) &= \frac{1}{1 + e^{-\alpha_i(r_i(x^\circ) - r_i(x^\star) - \varepsilon_i)}} \,, \\
\mathbb{P}(x^\star \equiv_i x^\circ | r_i) &= 1 - \mathbb{P}(x^\star \succ_i x^\circ | r_i) - \mathbb{P}(x^\star \prec_i x^\circ | r_i) \,.
\end{aligned} \tag{4}$$

Respectively, these are the probabilities that $x^\star$ is regarded as significantly better than, significantly worse than or not significantly different from $x^\circ$ when measured in terms of the reward function $r_i$. As before, the parameter $\alpha_i \geq 0$ represents the extent to which the reward is random. The parameter $\varepsilon_i \geq 0$ measures the extent to which the rewards of $x^\star$ and $x^\circ$ must differ for the difference to be regarded as "significant." Our model is then given by

$$\begin{aligned}
&\mathbb{P}(x^\star \succ x^\circ | \{r_i\}_{i=1}^k) \\
&\quad = \sum_{i=1}^{k} \left[ \mathbb{P}(x^\star \succ_i x^\circ | r_i) \cdot \prod_{j=1}^{i-1} \mathbb{P}(x^\star \equiv_j x^\circ | r_j) \right] .
\end{aligned} \tag{5}$$

Figure 2 offers a visual description of how the parameters $\alpha_i$ and $\varepsilon_i$ control the properties of this preference distribution.

As before, the probability that the preference generated by $\{r_i\}$ agrees with $\mathcal{P}$ with respect to $x^\star, x^\circ$ is

$$\begin{aligned}
\mathbb{P}(\mathcal{P}_{x^\star, x^\circ} | \{r_i\}) = &\tbinom{N(x^\star, x^\circ)}{n(x^\star, x^\circ)} \\
&\times \mathbb{P}(x^\star \succ x^\circ | \{r_i\})^{n(x^\star, x^\circ)} \\
&\times \mathbb{P}(x^\circ \succ x^\star | \{r_i\})^{n(x^\circ, x^\star)} \,.
\end{aligned} \tag{6}$$

Hence, the probability that the preference generated by $\{r_i\}$ agrees with $\mathcal{P}$ (i.e. the likelihood of $\{r_i\}$ w.r.t. $\mathcal{P}$) is

$$\begin{aligned}
\mathcal{L}(\{r_i\}; \mathcal{P}) &= \mathbb{P}(\mathcal{P} | \{r_i\}) \\
&= \left[ \textstyle\prod_{(x^\star, x^\circ) \in X \times X} \mathbb{P}(\mathcal{P}_{x^\star, x^\circ} | \{r_i\}) \right]^{1/2} .
\end{aligned} \tag{7}$$

And, as before, if the "true" reward functions are known/assumed to belong to a given family $R$, they can be inferred by finding the maximum likelihood estimate $\{\hat{r}_i\} = \operatorname{argmax}_{\{r_i \in R\}} \mathcal{L}(\{r_i\}; \mathcal{P})$.

Now, suppose we consider a parameterized family of reward functions such that $r_i = r_{\theta_i}$ for $\theta_i \in \Theta$, where $\Theta$ is the space of parameters. Then, the (approximate) MLE of the parameters $\{\theta_i\}_{i=1}^k$ can simply be found via gradient descent using the negative log-likelihood $\lambda = -\log \mathcal{L}(\{r_{\theta_i}\}; \mathcal{P})$ as the loss function. This is because $\lambda$ is differentiable with respect to $\{\theta_i\}$ as long as $r_\theta$ is a differentiable parameterization with respect to $\theta$ (which we show in Appendix A). In our exposition so far, we have chosen to treat $\alpha_i$'s and $\varepsilon_i$'s as hyperparameters for simplicity. However, in practice, they can easily be treated as free variables and learned alongside with parameters $\{\theta_i\}$, which we will be doing in all of our experiments. (We show that $\lambda$ is differentiable with respect to $\{\alpha_i\}$ and $\{\varepsilon_i\}$ in Appendix A as well.)

Finally, it needs to be emphasized that LORI is inherently capable of identifying the priority order of reward functions as well as which reward functions are relevant to modeling preferences. This is because different permutations of a given set of parameters $\{\theta_i\}_{i=1}^k$ are all present in our search space $\Theta^k$. In contrast, ordinal models of lexicographic preferences assume relevant attributes to be specified beforehand and only learn the priority order of those attributes.

## Analysis of LORI

Our model satisfies the following desirable properties:

(i) Setting $k = 1$ and $\varepsilon_1 = 0$ reproduces the logistic model given in (1),

(ii) Taking $\alpha_i \to \infty$, $\varepsilon_i \to 0$, and $\alpha_i \varepsilon_i \to \infty$ yields the no-errors, deterministic case,

(iii) The parameters $\varepsilon_i$ have natural interpretations. They are the thresholds above which a reward difference is considered significant: $\mathbb{P}(x^\star \succ_i x^\circ | r_i) > 1/2$ if and only if $r_i(x^\star) - r_i(x^\circ) > \varepsilon_i$.

It is worth noting that lexicographic representations using *multiple* reward functions are not only convenient, but (often) *necessary*. For the simplest example, consider the ordinary

lexicographic preference relation on $\mathbb{R}^2$: $(x_1, x_2) \succ (x_1', x_2')$ if $x_1 > x_1'$ or $x_1 = x_1'$ and $x_2 > x_2'$. It is well-known that there does not exist any reward function $r : \mathbb{R}^2 \to \mathbb{R}$ that represents $\succ$. That is, there is no reward function $r$ with the property that $(x_1, x_2) \succ (x_1', x_2')$ if and only if $r(x_1, x_2) > r(x_1', x_2')$. Similarly, the ordinary lexicographic ordering $\succ$ on $\mathbb{R}^{k+1}$ cannot be represented by a lexicographic ordering involving only $k$ reward functions.

When we consider probabilistic orderings and allow for errors, we can again find simple situations in which preferences that employ $k+1$ reward functions cannot be represented in terms of $k$ reward functions. For example, consider $X = \mathbb{R}^2$; let the reward functions $r_1, r_2 : \mathbb{R}^2 \to \mathbb{R}$ be the coordinate projections and take $\alpha_1 = \alpha_2 = 1$ and $\varepsilon_1 = \varepsilon_2 = 1$. The probabilistic preference relation $\succ$ defined by these reward functions and parameters is *intransitive*. For example, consider the points $x, x', x'' \in \mathbb{R}^2$ defined by: $x = (-0.6, 2)$, $x' = (0, 0)$, $x'' = (0.6, -2)$. Direct calculation shows that $\mathbb{P}(x \succ x') > 0.5$, $\mathbb{P}(x' \succ x'') > 0.5$, and $\mathbb{P}(x'' \succ x) > 0.5$. On the other hand, suppose we are given a reward function $r : \mathbb{R}^2 \to \mathbb{R}$, and parameters $\alpha \geq 0$ and $\varepsilon \geq 0$. If we define $\succ'$ to be the probabilistic preference relation defined by $r, \alpha, \varepsilon$, then $\succ'$ is necessarily *transitive*. To see this, note in order that if $\mathbb{P}(x^\star \succ' x^\circ) > 0.5$, we must have $\alpha(r(x^\star) - r(x^\circ) - \varepsilon) > 0$; in particular, we must have $r(x^\star) - r(x^\circ) > \varepsilon$. Hence, if $\mathbb{P}(x \succ' x') > 0.5$ and $\mathbb{P}(x' \succ' x'') > 0.5$ then we must have $r(x) - r(x') > \varepsilon$ and $r(x') - r(x'') > \varepsilon$, whence $r(x) - r(x'') > 2\varepsilon > \varepsilon$ and $\mathbb{P}(x \succ' x'') > 0.5$.

Finally, a reasonable question to ask is how to determine the number of reward functions $k$ when using LORI. For lexicographic models, even when the number of potential criteria considered by the agent is large, the number of criteria that actually enter into preference is likely to remain small. Remember that, in a lexicographic model, the $i$-th most important criterion will only be relevant for a particular decision if the $i-1$ more important criteria are all deemed equivalent. For most decision, this would be unlikely for even moderately large $i$. This means using a lexicographic model that employs only a few criteria (i.e. a model with small $k$) would be enough in most cases; increasing $k$ any further would have little to no benefit in terms of accuracy. We demonstrate this empirically in Appendix B.

## Illustrative Examples in Healthcare

Inferring lexicographically-ordered reward functions from preferences can be used either to improve decision-making behavior or to understand it. Here, we give examples of each in healthcare. However, it is worth noting that applications of LORI are not limited to the healthcare setting; it can be applied in any decision-making environment where multiple objectives are at play.

### Improving Behavior: Cancer Treatment

Consider the problem of treatment planning for cancer patients. For a given patient, write $a_t \in A = \{0, 1\}$ for the binary application of a treatment such as chemotherapy, $z_t \in Z = \mathbb{R}$ for tumor volume (as a measure of the treatment's efficacy), and $w_t \in W = \mathbb{R}$ for the white blood cell (WBC) count (as a measure of the treatment's toxicity) at time $t \in \{1, 2, \ldots\}$. In our experiments, we will simulate the tumor volume $z_t$ and the WBC count $w_t$ of a patient given a treatment plan $a_t$ according to the pharmacodynamic models of Iliadis and Barbolosi (2000):

$$\begin{aligned} z_{t+1} &= z_t + 0.003 z_t \log(1000/z_t) - 0.15 z_t a_t + \nu_t \\ w_{t+1} &= w_t + 1.2 - 0.15 w_t - 0.4 w_t a_t + \eta_t \ , \end{aligned} \tag{8}$$

where $\nu_t \sim \mathcal{N}(0, 0.5^2)$ and $\eta_t \sim \mathcal{N}(0, 0.5^2)$ are noise variables, $z_1 \sim \mathcal{N}(30, 5^2)$, and $w_1 = 8$. Notice that both the tumor volume and the WBC count decrease when the treatment is applied and increase otherwise. We assume that clinicians aim to minimize the tumor volume while keeping the average WBC count above a threshold of five throughout the treatment. We define the set of alternatives to be all possible treatment trajectories: $X = \bigcup_{\tau=1}^{\infty} (A \times Z \times W)^\tau$. Then, the treatment objective can be represented in terms of lexicographic reward functions:

$$\begin{aligned} r_1(\boldsymbol{a}_{1:\tau}, \boldsymbol{z}_{1:\tau}, \boldsymbol{w}_{1:\tau}) &= \min\left\{5, \tfrac{1}{\tau}\textstyle\sum_{t=1}^{\tau} w_t\right\} \ , \\ r_2(\boldsymbol{a}_{1:\tau}, \boldsymbol{z}_{1:\tau}, \boldsymbol{w}_{1:\tau}) &= -\tfrac{1}{\tau}\textstyle\sum_{t=1}^{\tau} z_t \ . \end{aligned} \tag{9}$$

A *policy* $\pi \in \Delta(A)^{Z \times W}$ is a function from the features of a patient at time $t$ to a distribution over actions such that $a_t \sim \pi(\cdot | z_t, w_t)$. By definition, the optimal policy is given by $\pi^* = \operatorname{argmax}_{\text{lex},\pi} \mathbb{E}[\boldsymbol{r}(\boldsymbol{a}_{1:\tau}, \boldsymbol{z}_{1:\tau}, \boldsymbol{w}_{1:\tau})]$. We *do not* assume that clinicians follow the optimal policy, but rather follow some policy that approximates the optimal policy: $\pi_b(a | \boldsymbol{z}_{1:t}, \boldsymbol{w}_{1:t}) = (1-\epsilon)\pi^*(a | \boldsymbol{z}_{1:t}, \boldsymbol{w}_{1:t}) + \epsilon/|A|$ for some $\epsilon > 0$, which we will call the behavior policy. This means that the policy actually followed by clinicians is improvable.

Now, suppose we want to improve the policy followed by clinicians using reinforcement learning but we do not have access to their preferences explicitly in the form of reward functions $\{r_1, r_2\}$ so we cannot compute the optimal policy $\pi^*$ directly. Instead, we have access to some demonstrations $\mathcal{D} \subset X$ generated by clinicians as they follow the behavior policy $\pi_b$ and we query an expert or panel of experts (which might consist of the clinicians themselves) over the demonstrations/alternatives in $\mathcal{D}$ to obtain a set of observed preferences $\mathcal{P}$. We use LORI to estimate the reward functions $\{r_1, r_2\}$ from the observed preferences $\mathcal{P}$.

**Setup** For each experiment, we take $\epsilon = 0.5$ and generate 1000 trajectories with $\tau = 20$ to form the demonstration set $\mathcal{D}$. Then, we generate preferences by sampling 1000 trajectory pairs from $\mathcal{D}$ and evaluating according to the ground-truth reward functions $\{r_1, r_2\}$ and the model given in (5), where $\varepsilon_1 = \varepsilon_2 = 0.1$ and $\alpha_1 = \alpha_2 = 10\log(9)$ (ties are broken uniformly at random). These form the set of expert preferences $\mathcal{P}$.

We infer $k = 2$ reward functions from the expert preferences $\mathcal{P}$ using LORI; for comparison, we infer a single reward function using T-REX (Brown et al. 2019), which is the single-dimensional counterpart of LORI (with $k = 1$), and another single reward function from demonstrations $\mathcal{D}$ instead of preferences $\mathcal{P}$ using Bayesian IRL (Ramachandran and Amir 2007). (Keep in mind that LORI infers lexicographic reward functions but the two alternatives infer only a

Table 2: *Comparison of reward functions based on preference prediction performance.* LORI performs the best.

| Alg. | RMSE | Accuracy |
|---|---|---|
| BIRL | $0.323 \pm 0.005$ | $88.1\% \pm 0.95\%$ |
| T-REX | $0.214 \pm 0.007$ | $89.1\% \pm 0.61\%$ |
| LORI | $0.103 \pm 0.089$ | $92.4\% \pm 2.71\%$ |

Table 3: *Comparison of policies based on how often the row policies are preferred to the column policies.* LORI is the most frequently preferred to the behavior policy and its performance is comparable to that of the optimal policy.

| | Behavior | BC | BIRL | T-REX | LORI |
|---|---|---|---|---|---|
| BC | $56.7\% \pm 1.18\%$ | – | | | |
| BIRL | $65.6\% \pm 1.17\%$ | $56.0\% \pm 1.48\%$ | – | | |
| T-REX | $65.7\% \pm 1.20\%$ | $56.2\% \pm 1.66\%$ | $49.8\% \pm 0.93\%$ | – | |
| LORI | $75.2\% \pm 0.84\%$ | $69.7\% \pm 0.96\%$ | $82.4\% \pm 0.59\%$ | $81.8\% \pm 0.74\%$ | – |
| Optimal | $75.1\% \pm 1.31\%$ | $69.1\% \pm 1.11\%$ | $82.5\% \pm 0.27\%$ | $82.4\% \pm 0.55\%$ | $49.3\% \pm 1.30\%$ |

single reward function.) LORI also infers parameters $\varepsilon_1$ and $\varepsilon_2$ together with $r_1, r_2$. (We set $\alpha_1 = \alpha_2 = 1$; there is no loss of generality because the other variables simply scale.)

Then, using the estimated reward functions, we train optimal policies using reinforcement learning. In the case of LORI, we use the algorithm proposed in Gábor, Kalmár, and Szapesvári (1998), which is capable of optimizing thresholded lexicographic objectives. We also learn a policy directly from the demonstration set $\mathcal{D}$ by performing behavioral cloning (BC), where actions are simply regressed on patient features. The resulting policy aims to mimic the behavior policy $\pi_b$ as closely as possible. Details regarding the implementation of algorithms can be found in Appendix C. We repeat each experiment five times to obtain error bars.

**Results** To evaluate the quality of reward functions learned by LORI, T-REX, and BIRL, we compare the prediction accuracy of the preference models they induce on a test set of 1000 additional preferences; the results are shown in Table 2. We see that LORI performs better than either T-REX and BIRL. This is because LORI is the only method to capture the multi-objective nature of the clinicians' goal (and the expert's preferences). BIRL performs the worst since the demonstrations in $\mathcal{D}$ are suboptimal.

Table 3 compares the performance of various policies: the behavior policy, the policy that is learned via BC, the policies that are trained on the basis of reward functions learned by BIRL, T-REX and LORI, and the true optimal policy. Each entry in Table 3 shows the frequency with which the row policy is preferred to the column policy. Letting $x \sim \pi$ be the distribution of trajectories $x$ generated by following policy $\pi$, the frequency that policy $\pi^\star$ is preferred to the policy $\pi^\circ$ is defined to be $\mathbb{P}(\pi^\star \succ \pi^\circ) = \mathbb{E}_{x^\star \sim \pi^\star, x^\circ \sim \pi^\circ} \mathbb{P}(x^\star \succ x^\circ | r_1, r_2)$. (This is estimated by sampling 1000 trajectories from both policies.) We use the frequency with which one policy is preferred to another as the measure for the improvement provided by the first policy over the second. (Note that because there is no single ground-truth reward function, it is not feasible to compare the *values* of the two policies, which would have been the usual measure used in reinforcement learning.) As can be seen, LORI improves on other methods, improves on the behavioral policy more often than other methods, and is almost indistinguishable from the true optimal policy.

Appendix B includes additional experiments where the ground-truth preferences are generated by a single reward function (rather than $r_1, r_2$ defined in (9)). Since our preference model is strictly a generalization of single-dimensional models, LORI (with $k = 2$) still performs the best—but it does not improve over T-REX as much.

## Understanding Behavior: Organ Transplantation

Consider the organ allocation problem for transplantation. Write $P$ for the space of all patients (or patient characteristics) and $O$ for the set of organs. At a given time $t \in T$, there is a set of patients $P_t \subset P$ who are waiting for an organ transplant. When an organ $o_t \in O$ becomes available at time $t$, an allocation policy matches that organ with a particular patient $p_t \in P_t$. In effect, the allocation policy expresses a preference relation: letting $X = P \times O$ be the set of alternatives, the pairing $(p_t, o_t) \in X$ is preferred over alternative pairings $\{(p', o_t)\}_{p' \in P_t \setminus \{p_t\}}$ that were also possible at time $t$. From these observations, we can infer a lexicographic reward representation $\{r_i\}_{i=1}^k$ that explains the decisions made by the allocation policy. (Notice that if we view elements of $P$ as patient characteristics, then the observed preferences need not be consistent over time.)

**Setup** We investigate the liver allocation preferences in the US in terms of the *benefit* a transplantation could provide and the *need* of patients for a transplant. These two objectives are at odds with each other since the patient in greatest need might not be the patient who would benefit the most from an available organ. As a measure of benefit, we consider the estimated additional number of days a patient would survive if they were to receive the available organ, and as a measure of need, we consider the estimated number of days they would survive without a transplant. We estimate both benefit and need via Cox models following the same methodology as TransplantBenefit, which is used in the current allocation policy of the UK (Neuberger et al. 2008). Our analysis is based on the Organ Procurement and Transplantation Network (OPTN) data for liver transplantations as of December 4, 2020. From the OPTN data, we only consider patients who were in the waitlist during 2019 and for whom benefit and need can be estimated from the recorded features. This leaves us with 4070 patients, 3942 organ arrivals, and 2,450,718 observations about the allocation policy's preferences over patient-organ pairings.

We use LORI (with $k = 2$) and T-REX (Brown et al. 2019), which is the single-dimensional counterpart of LORI ($k = 1$, $\varepsilon_1 = 0$), to infer reward functions that are linear with respect to benefit and need. (The restriction to linear reward functions means that preferences depend only on the benefit and need *differences* between the two pairings.) In the case of LORI,

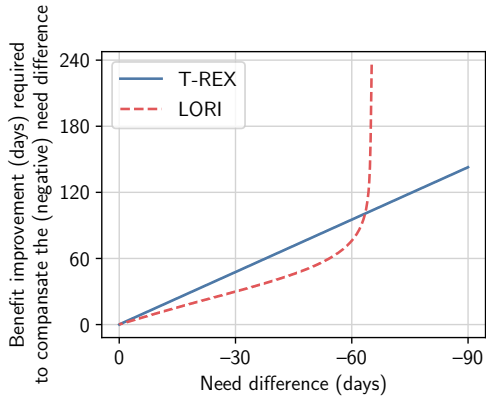



Figure 3: *Trade-off between benefit and need according to T-REX and LORI.* For both T-REX and LORI, there is a trade-off of benefit against need. For T-REX, this trade-off occurs at a constant rate. For LORI, the trade-off occurs at an almost constant rate when the need difference is below 60 days, but above 60 days the required trade-off of benefit against need increases sharply, and above 65 days it is almost impossible for benefit to compensate for the need difference.

we also infer $\varepsilon_1$ and $\varepsilon_2$ to determine what amount of benefit or need is considered to be significant by the allocation policy. For T-REX, we set $\alpha = 1$, and for LORI, we set $\alpha_1 = \alpha_2 = 1$ (without loss of generality because the remaining variables are free to be scaled arbitrarily).

**Results** The single-dimensional reward function learned by T-REX is

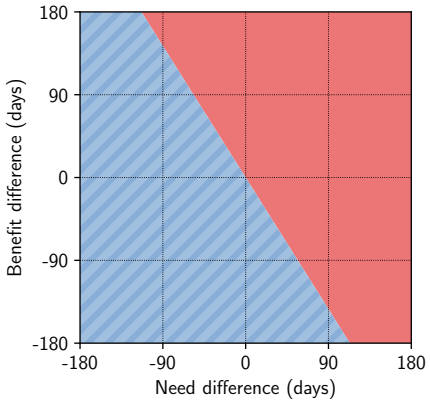
$$r = 0.0086 \cdot \textit{Benefit} + 0.0137 \cdot \textit{Need} \tag{10}$$

while the lexicographically-ordered reward functions learned by LORI are

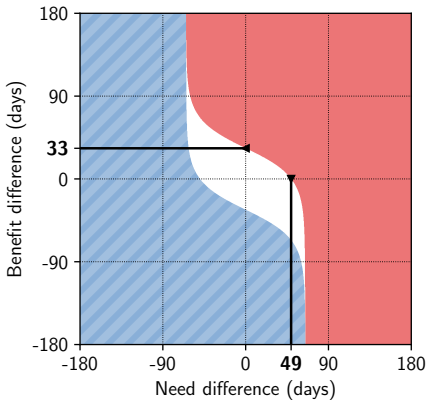
$$\begin{aligned} r_1 &= 0.0001 \cdot \textit{Benefit} + 0.0139 \cdot \textit{Need} \\ r_2 &= 0.0561 \cdot \textit{Benefit} + 0.0002 \cdot \textit{Need} \end{aligned} \tag{11}$$

with $\varepsilon_1 = 0.8955$ and $\varepsilon_2 = 1.8799$.

As Figure 3 shows, LORI reveals that need is *prioritized* over benefit. This can be seen by the weights in the primary reward function $r_1$ and secondary reward function $r_2$, but more specifically in the finding that a need difference greater than 65 days cannot be outweighed by *any* benefit difference; *guaranteeing* that the patient with greater need will receive the organ. By contrast, there is no such finding for T-REX: any need difference can be outweighed by a sufficiently large benefit difference. This finding is consistent with current allocation practices in the US. When an organ becomes available for transplantation, it is offered to a patient based on their MELD score (Wiesner et al. 2003), which is strictly a measure of how sick the patient is, and so considers only the patient's need and not the benefit they will obtain. However after an offer is made, clinicians might still choose to reject the offer in order to wait for a future offer that would be more beneficial for their patient. Since offers are first made on the basis of need and then accepted/rejected on the basis of benefit, it is reasonable for need to have priority over benefit in the representation learned by LORI.




**(a) Preferences learned by T-REX**




**(b) Preferences learned by LORI**

Figure 4: *Preferences with respect to benefit and need differences between a candidate pairing and an alternative pairing.* Red (solid) means the candidate pairing while blue (striped) means the alternative pairing is likely to be preferred with probability at least 0.5. For LORI, white denotes a region of indifference, where the probability of preferring either pairing is less than 0.5.

Figure 4 depicts visually the preferences induced by the two alternative representations. As pointed out above, in the preferences learned by LORI, a need difference greater than 65 days cannot be outweighed by any benefit difference, while in the preferences learned by T-REX, any need difference can be outweighed by a sufficiently large benefit difference (and the rate at which benefit trades-off for need is constant). Moreover, LORI identifies a region (indicated by white in Figure 4b), where differences in benefit and need become insignificantly small and preferences are made mostly at random (at least with respect to benefit and need).

## Conclusion

In this paper, we proposed LORI, which is a method for inferring lexicographically-ordered reward functions from the preferences of an expert or a decision-making agent. Through examples from healthcare, we showed that learning such reward functions can be useful in two ways: either to improve behavior or to understand behavior.

## Acknowledgements

This work was supported by the US Office of Naval Research (ONR) and the National Science Foundation (NSF, grant number 1722516). Part of our experimental results are based on the liver transplant data from OPTN, which was supported in part by Health Resources and Services Administration contract HHSH250-2019-00001C.

## References

Abbeel, P.; and Ng, A. Y. 2004. Apprenticeship learning via inverse reinforcement learning. In *Proc. 23rd AAAI Conf. Artif. Intell.*

Bonilla, E. V.; Guo, S.; and Sanner, S. 2010. Gaussian process preference elicitation. In *Proc. 24th Conf. Neural Inf. Process. Syst.*

Booth, R.; Chevaleyre, Y.; Lang, J.; Mengin, J.; and Sombattheera, C. 2010. Learning conditionally lexicographic preference relations. In *Proc. 19th Eur. Conf. Artif. Intell.*

Boularias, A.; Kober, J.; and Peters, J. 2011. Relative entropy inverse reinforcement learning. In *Proc. 14th Int. Conf. Artif. Intell. Statist.*

Boutilier, C.; Brafman, R. I.; Domshlak, C.; and Hoss, H. H. 2004. CP-nets: a tool for representing and reasoning with conditional ceteris paribus preference statements. *J. Artif. Intell. Res.*, 21: 135–191.

Braüning, M.; and Hüllermeyer, E. 2012. Learning conditional lexicographic preference trees. In *Proc. ECAI-12 Workshop Preference Learn.: Problems Appl. AI.*

Braüning, M.; Hüllermeyer, E.; Keller, T.; and Glaum, M. 2017. Lexicographic preferences for predictive modeling of human decision making: a new machine learning method with an application in accounting. *Eur. J. Oper. Res.*, 258(1): 295–306.

Bräuninger, T.; Müller, J.; and Stecker, C. 2016. Modeling preferences using roll call votes in parliamentary systems. *Political Anal.*, 24: 189–210.

Brochu, E.; Freitas, N.; and Ghosh, A. 2008. Active preference learning with discrete choice data. In *Proc. 22nd Conf. Neural Inf. Process. Syst.*

Brown, D.; Goo, W.; and Niekum, S. 2019. Better-than-demonstrator imitation learning via automatically-ranked demonstrations. In *Proc. 3rd Conf. Robot Learn.*

Brown, D.; Goo, W.; Prabhat, N.; and Niekum, S. 2019. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In *Proc. 36th Int. Conf. Mach. Learn.*

Chou, G.; Berenson, D.; and Özay, N. 2018. Learning constraints from demonstrations. In *Proc. 13th Int. Workshop Algorithmic Found. Robot.*

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. In *Proc. 31st Conf. Neural Inf. Process. Syst.*

Chu, W.; and Ghahramani, Z. 2005. Preference learning with Gaussian processes. In *Proc. 22nd Int. Conf. Mach. Learn.*

Colman, A. M.; and Stirk, J. A. ???? Singleton bias and lexicographic preferences among equally valued alternatives. *J. Econom. Behav. Organ.*, 40(4): 337–351.

Coombes, J. M.; and Trotter, J. F. 2005. Development of the allocation system for deceased donor liver transplantation. *Clin. Med. Res.*, 3(2): 87–92.

Dombi, J.; Imreh, C.; and Vincze, N. 2007. Learning lexicographic orders. *Eur. J. Oper. Res.*, 183: 748–756.

Drolet, A.; and Luce, M. F. 2004. Cognitive load and trade-off avoidance. *J. Consumer Res.*, 31: 63–77.

Fargier, H.; Gimenez, P.-F.; and Mengin, J. 2018. Learning lexicographic preference trees from positive examples. In *Proc. 32nd AAAI Conf. Artif. Intell.*

Fernandes, K.; Cardoso, J. S.; and Palacios, H. 2016. Learning and ensembling lexicographic preference trees with multiple Kernels. In *Int. Joint Conf. Neural Netw.*

Finn, C.; Levine, S.; and Abbeel, P. 2016. Guided cost learning: deep inverse optimal control via policy optimization. In *Proc. 33rd Int. Conf. Mach. Learning*.

Gábor, Z.; Kalmár, Z.; and Szapesvári, C. 1998. Multi-criteria reinfrocement learning. In *Proc. 15th Int. Conf. Mach. Learn.*

Gutjahr, W. J.; and Nolz, P. C. 2016. Multicriteria optimization in humanitarian aid. *Eur. J. Oper. Res.*, 252(2): 351–366.

Ibarz, B.; Leike, J.; Pohlen, T.; Irving, G.; Legg, S.; and Amodei, D. 2018. Reward learning from human preferences and demonstrations in Atari. In *Proc. 32nd Conf. Neural Inf. Process. Syst.*

Iliadis, A.; and Barbolosi, D. 2000. Optimizing drug regimens in cancer chemotherapy by an efficacy-toxicity mathematical model. *Comput. Biomed. Res.*, 33(3): 211–226.

Jedidi, K.; and Kohli, R. 2008. Inferring latent class lexicographic rules from choice data. *J. Math. Psychol.*, 52: 241–240.

Jee, K.-W.; McShan, D. L.; and Fraass, B. A. 2007. Lexicographic ordering: intuitive multicriteria optimization for IMRT. *Phys. Medicine Biol.*, 52(7): 1845–1861.

Kohli, R.; Boughanmi, K.; and Kohli, V. 2019. Randomized algorithms for lexicographic inference. *Oper. Res.*, 67(2): 357–375.

Kohli, R.; and Jedidi, K. 2007. Representation and inference of lexicographic preference models and their variants. *Marketing Sci.*, 26(3): 380–399.

Koyuncu, M.; and Erol, R. 2010. Optimal resource allocation model to mitigate the impact of pandemic influenza: a case study for Turkey. *J. Med. Syst.*, 34(1): 61–70.

Levine, S.; Popovic, Z.; and Koltun, V. 2011. Nonlinear inverse reinforcement learning with Gaussian processes. In *Proc. 24th Conf. Neural Inf. Process. Syst.*

Liu, X.; and Truszczynski, M. 2015. Learning partial lexicographic preference trees over combinatorial domains. In *Proc. 29th AAAI Conf. Artif. Intell.*

McFadden, D. 1973. Conditional logit analysis of qualitative choice behavior. In Zarembka, P., ed., *Frontiers in Econometrics*. New York: Academic Press.

Mühlbacher, A.; and Johnson, F. R. 2016. Choice experiments to quantify preferences for health and healthcare: state of the practice. *Appl. Health Econ. Health Policy*, 14(3): 253–266.

Neuberger, J.; Gimson, A.; Davies, M.; Akyol, M.; O'Grady, J.; Burroughs, A.; and Hudson, M. 2008. Selection of patients for liver transplantation and allocation of donated livers in the UK. *Gut*, 57(2): 252–257.

Ng, A. Y.; and Russell, S. J. 2000. Algorithms for inverse reinforcement learning. In *Proc. 17th Int. Conf. Mach. Learn.*

Ramachandran, D.; and Amir, E. 2007. Bayesian inverse reinforcement learning. In *Proc. 20th Int. Joint Conf. Artif. Intell.*

Salvatore, C.; Salvatore, G.; Kadziński, M.; and Słowiński, R. 2013. Robust ordinal regression in preference learning and ranking. *Mach. Learn.*, 93(2): 381–422.

Schaubel, D. E.; Guidinger, M. K.; Biggins, S. W.; Kalbfleisch, J. D.; Pomfret, E. A.; Sharma, P.; and Merion, R. M. 2009. Survival benefit-based deceased-donor liver allocation. *Amer. J. Transplantation*, 9: 970–981.

Schmitt, M.; and Martignon, L. 2006. On the complexity of learning lexicographic strategies. *J. Mach. Learn. Res.*, 7: 55–83.

Scobee, D. R. R.; and Sastry, S. S. 2020. Maximum likelihood constraint inference for inverse reinforcement learning. In *Proc. 8th Int. Conf. Learn. Representations*.

Singh, V. P.; Hansen, K. T.; and Gupta, S. 2005. Modeling preferences for common attributes in multicategory brand choice. *J. Marketing Res.*, 42: 195–209.

Slovic, P. 1975. Choice between equally values alternatives. *J. Experiment. Psych.: Human Perception Performance*, 1: 280–287.

Sugiyama, H.; Meguro, T.; and Minami, Y. 2012. Preference-learning based inverse reinforcement learning for dialog control. In *Proc. 13th Annu. Conf. Int. Speech Commun. Assoc.*

Tversky, A.; Sattah, S.; and Slovic, P. 1988. Contingent weighting in judgement and choice. *Psych. Rev.*, 95: 317–384.

Vazquez-Chanlatte, M.; Jha, S.; Tiwari, A.; Ho, M. K.; and Seshia, S. A. 2018. Learning task specifications from demonstrations. In *Proc. 32nd Conf. Neural Inf. Process. Syst.*

Wiesner, R.; Edwards, E.; Freeman, R.; Harper, A.; Kim, R.; Kamath, P.; Kremers, W.; Lake, J.; Howard, T.; Merion, R. M.; Wolfe, R. A.; Krom, R.; and The United Network for Organ Sharing Liver Disease Severity Score Committe. 2003. Model for end-stage liver disease (MELD) and allocation of donor livers. *Gastroenterology*, 124(1): 91–96.

Wilkens, J. J.; Alaly, J. R.; Zakarian, K.; Thorstad, W. L.; and Deasy, J. O. 2007. IMRT treatment planning based on prioritizing prescription goals. *Phys. Medicine and Biol.*, 52(6): 1675–1692.

Wirth, C.; Fürnkranz, J.; and Neumann, G. 2016. Model-free preference-based reinforcement learning. In *Proc. 30th AAAI Conf. Artif. Intell.*

Yaman, F.; Walsh, T. J.; Littman, M. L.; and Desjardins, M. 2008. Democratic approximation of lexicographic preference models. In *Proc. 25th Int. Conf. Mach. Learn.*

Yee, M.; Dahan, E.; Hauser, J. R.; and Orlin, J. 2007. Greedoid-based noncompansatory consideration-then-choice inference. *Marketing Sci.*, 26(4): 532–549.

Ziebart, B. D.; Maas, A.; Bagnell, J. A.; and Day, A. K. 2008. Maximum entropy inverse reinforcement learning. In *Proc. 23rd AAAI Conf. Artif. Intell.*