Dominance Criteria and Solution Sets for the Expected Scalarised Returns

Conor F. Hayes
School of Computer Science
National University of Ireland Galway
Ireland
c.hayes13@nuigalway.ie

Timothy Verstraeten
AI Lab
Vrije Universiteit Brussel
Belgium
timothy.verstraeten@vub.ac.be

Diederik M. Roijers
AI Lab, Vrije Universiteit Brussel (BE)
& Microsystems Technology,
HU Univ. of Appl. Sci. Utrecht (NL)
diederik.yamamoto-roijers@hu.nl

Enda Howley
School of Computer Science
National University of Ireland Galway
Ireland
enda.howley@nuigalway.ie

Patrick Mannion
School of Computer Science
National University of Ireland Galway
Ireland
patrick.mannion@nuigalway.ie

ABSTRACT

In many real-world scenarios, the utility of a user is derived from the single execution of a policy. In this case, to apply multi-objective reinforcement learning, the expected utility of the returns must be optimised. Various scenarios exist where a user's preferences over objectives (also known as the utility function) are unknown or difficult to specify. In such scenarios, a set of optimal policies must be learned. However, settings where the expected utility must be maximised have been largely overlooked by the multi-objective reinforcement learning community and, as a consequence, a set of optimal solutions has yet to be defined. In this paper we address this challenge by proposing first-order stochastic dominance as a criterion to build solution sets to maximise expected utility. We also propose a new dominance criterion, known as expected scalarised returns (ESR) dominance, that extends first-order stochastic dominance to allow a set of optimal policies to be learned in practice. Finally, we define a new solution concept called the ESR set, which is a set of policies that are ESR dominant.

KEYWORDS

multi-objective; decision making; distributional; reinforcement learning; stochastic dominance

1 INTRODUCTION

In multi-objective reinforcement learning (MORL), there are two classes of algorithms: single-policy and multi-policy [26, 31]. Each MORL algorithm has two phases: the learning phase and the execution phase [26]. When using single-policy methods, an agent learns a single optimal policy that maximises a user's utility function where a user's preferences over objectives are represented by a utility function. The agent then executes the optimal policy during the execution phase. Single-policy methods require the utility function of a user to be known during the learning phase. In certain scenarios a user's preferences over objectives may be unknown; therefore, the utility function is unknown. In this case, a user is said to be in the unknown utility function or unknown weights scenario [13, 26]. In the unknown utility function scenario, multi-policy methods must be used to learn a set of optimal policies during the learning phase. We assume that the utility function of the user will become known during the execution phase. Once the utility function of the user is known, it is possible to select a policy, from the set of learned policies, that will maximise the user's utility function.

In contrast to single-objective reinforcement learning (RL), multiple optimality criteria exist for MORL [26]. In scenarios where the utility of the user is derived from multiple executions of a policy, the scalarised expected returns (SER) must be optimised. However, in scenarios where the utility of a user is derived from a single execution of a policy, the expected utility of the returns (or expected scalarised returns, ESR) must be optimised. The majority of MORL research focuses on the SER criterion and linear utility functions [22], which limits the applicability of MORL to real-world problems. In the real world, a user's utility function may be derived in a linear or non-linear manner. For known linear utility functions, single-objective methods can be used to learn an optimal policy [26]. Non-linear utility functions do not distribute across the sums of the immediate and future returns, which invalidates the Bellman equation [25]. Therefore, to learn optimal policies for non-linear utility functions, strictly multi-objective methods must be used.

For non-linear utility functions, a user can prefer significantly different policies depending on whether the SER or ESR criterion is optimised [22, 23]. Unfortunately, the ESR criterion has received very little attention, to date, in the MORL community. To learn optimal policies in many real-world scenarios where a policy will be executed only once, the ESR criterion must be optimised. For example, in a medical setting where a user has one opportunity to select a treatment, a user will want to maximise the expected utility of a single outcome. However, choosing the wrong optimisation criterion (SER) for such a scenario could potentially lead to a different policy than that which would be learned under ESR. In the real world, like in the aforementioned scenario, learning a sub-optimal policy could have catastrophic outcomes.

Therefore, it is crucial that the MORL community focuses on developing both single-policy and multi-policy methods that can learn optimal policies under the ESR criterion. Recently, a number of single-policy methods have been implemented that can learn optimal policies under the ESR criterion [12, 25]. Based on the findings of Hayes et al. [11, 12], a distribution over the expected utility of the returns must be used to learn an optimal policy under

the ESR criterion in realistic settings where rewards are stochastic[1]. Traditionally, a single expected value of the returns is used to make decisions. However, the expected value cannot account for the range of positive or adverse effects a decision might have [12]. In the current MORL literature, no multi-policy methods exist for the ESR criterion. In fact, a set of optimal policies for the ESR criterion has yet to be defined.

This paper aims to fill the aforementioned research gaps that exist for the ESR criterion. Due to the lack of existing research for the ESR criterion, a formal definition of the requirements to satisfy the ESR criterion has yet to be defined. In Section 3, we define the requirements necessary to satisfy the ESR criterion. The applicability of MORL to many real-world scenarios under the ESR criterion is limited because no solution set has been defined for scenarios when a user's utility function is unknown. In Section 4, we show how first-order stochastic dominance can be used to define sets of optimal policies under the ESR criterion. However, using FSD in practice, when the utility function of a user is unknown, determining a set of optimal policies is difficult because FSD relies on having access to a utility function. We address this challenge in Section 5 and expand first-order stochastic dominance to define a new dominance criterion, called expected scalarised returns (ESR) dominance. This work proposes that ESR dominance can be used to learn a set of optimal solutions, which we define as the ESR set.

2 BACKGROUND

2.1 Multi-Objective Reinforcement Learning

In multi-objective reinforcement learning, we deal with decision making problems with multiple objectives, often modelled as a multi-objective Markov decision process. An MOMDP represents a tuple, $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{T}, \gamma, \mathcal{R})$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ is a probabilistic transition function, γ is a discount factor determining the importance of future rewards and $\mathcal{R} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}^n$ is an n-dimensional vector-valued immediate reward function. In multi-objective reinforcement learning, $n > 1$.

2.2 Utility Functions

In MORL, utility functions are used to model a user's preferences, and are used in both single-objective and multi-objective RL. Utility functions are functions that map returns to a scalar value which represents the user's preferences over the returns,

$$u : \mathbb{R}^n \rightarrow \mathbb{R}, \tag{1}$$

where u is a utility function and $\mathbf{R}^n$ is an n-dimensional vector. Linear utility functions are widely used to represent a user's preferences,

$$u = \sum_{i=1}^{n} w_i r_i, \tag{2}$$

where w_i is the preference weight and r_i is the value at position i of the return vector. However, certain scenarios exist where linear utility functions cannot accurately represent a user's preferences. In this case, the user's preferences must be represented using a non-linear utility function.

In this paper, we consider monotonically increasing utility functions [26], i.e.,

$$(\forall\, i, V_i^{\pi} \geq V_i^{\pi'} \wedge \exists\, i, V_i^{\pi} > V_i^{\pi'}) \implies (\forall\, u, u(\mathbf{V}^{\pi}) > u(\mathbf{V}^{\pi'})), \tag{3}$$

where $\mathbf{V}^{\pi}$ and $\mathbf{V}^{\pi'}$ are the values of executing policies π and π' respectively.

A monotonically increasing utility function includes linear utility functions of the form in Equation 2. It is important to note that in certain scenarios the utility function may be unknown, therefore we do not know the shape of the utility function. If we assume the utility function is monotonically increasing we know that, if the value of one of the objectives in the return vector increases, then the utility also increases [26]. This assumption makes it possible to determine an ordering over policies when the shape of the utility function is unknown.

2.3 Scalarised Expected Returns and Expected Scalarised Returns

For MORL, the ability to express a user's preferences over objectives as a utility function is essential when learning a single optimal policy. In MORL different optimality criteria exist [26]. In MORL, the utility function can be applied to the expectation of the returns, or the utility function can be applied directly to the returns before computing the expectation. Calculating the expected value of the return of a policy before applying the utility function leads to the scalarised expected returns (SER) optimisation criterion:

$$V_u^{\pi} = u\left(\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathbf{r}_t \,\middle|\, \pi, \mu_0\right]\right), \tag{4}$$

where μ_0 is the probability distribution over possible starting states.

SER is the most commonly used criterion in the multi-objective (single agent) planning and reinforcement learning literature [26]. For SER, a coverage set is defined as a set of optimal solutions for all possible utility functions. If the utility function is instead applied before computing the expectation, this leads to the expected scalarised returns (ESR) optimisation criterion [12, 25, 26]:

$$V_u^{\pi} = \mathbb{E}\left[u\left(\sum_{t=0}^{\infty} \gamma^t \mathbf{r}_t\right) \,\middle|\, \pi, \mu_0\right]. \tag{5}$$

ESR is the most commonly used criterion in the game theory literature on multi-objective games [22].

2.4 Stochastic Dominance

Stochastic dominance [3, 10] gives a partial order between distributions and can be used when making decisions under uncertainty. Stochastic dominance is particularly useful when a distribution must be taken into consideration rather than an expected value when making decisions. Stochastic dominance is a prominent dominance criterion in finance, economics and decision theory. When making decisions under uncertainty, Stochastic dominance can be used to determine the most risk averse decision. Various degrees of stochastic dominance exist, however, in this paper we focus on first-order stochastic dominance (FSD). FSD can be used to give a

[1]We note that distributional methods also work well for simple problems with deterministic rewards. In such cases, the value distribution only has a single value vector per state-action pair that occurs with probability 1.0.

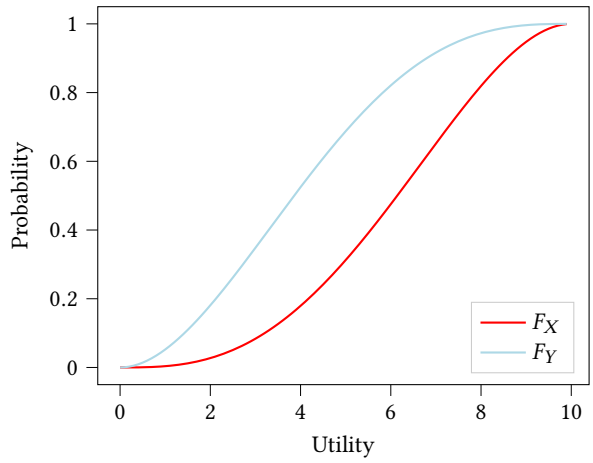

Figure 1: For random variables X and Y, $X \geq_{FSD} Y$, where F_X and F_Y are the cumulative distribution functions (CDFs) of X and Y respectively. In this case, X is preferable to Y because higher utilities occur with greater frequency in F_X.

partial ordering over random variables or random vectors to give an FSD dominant set.

In Definition 2.1 we present the necessary conditions for FSD and in Theorem 2.2 we prove that if a random variable is FSD dominant it has at least as high an expected value as another random variable [34]. We use the work of Wolfstetter [34] to prove Theorem 2.2.

Definition 2.1. For random variables X and Y, X $\geq_{FSD}$ Y if:

$$P(X > z) \geq P(Y > z), \forall z$$

If we consider the cumulative distribution function (CDF) of X, F_X, and the CDF of Y, F_Y, we can say that X $\geq_{FSD}$ Y if:

$$F_X(z) \leq F_Y(z), \forall z.$$

Theorem 2.2. *If $X \geq_{FSD} Y$, then X has a greater than or equal expected value as Y.*

$$X \geq_{FSD} Y \implies E(X) \geq E(Y).$$

Proof. By a known property of expected values the following is true for any random variable:

$$\mathbb{E}(X) = \int_0^{+\infty} (1 - F_X(x))\, dx$$

$$\mathbb{E}(Y) = \int_0^{+\infty} (1 - F_Y(x))\, dx$$

Therefore, if X $\geq_{FSD}$ Y then:

$$\int_0^{+\infty} (1 - F_X(x))\, dx \geq \int_0^{+\infty} (1 - F_Y(x))\, dx$$

Which gives,

$$\mathbb{E}(X) \geq \mathbb{E}(Y)$$

[34] □

3 EXPECTED SCALARISED RETURNS

In contrast to single-objective reinforcement learning, different optimality criteria exist for MORL. In scenarios where the utility of a user is derived from multiple executions of a policy, the agent should optimise over the SER criterion. In scenarios where the utility of a user is derived from a single execution of a policy, the agent should optimise over the ESR criterion. Let us consider, as an example, a power plant that generates electricity for a city and emits harmful CO_2 and greenhouse gases. City regulations have been imposed which limit the amount of pollution that the power plant can generate. If the regulations require that the emissions from the power plant do not exceed a certain amount over an entire year, the SER criterion should be optimised. In this scenario, the regulations allow for the pollution to vary day to day, as long as the emissions do not exceed the regulated level for a given year. However, if the regulations are much stricter and the power plant is fined every day it exceeds a certain level of pollution, it is beneficial to optimise under the ESR criterion.

The majority of MORL research focuses on linear utility functions. However, in the real world, a user's utility function can be non-linear. For example, a utility function is non-linear in situations where a minimum value must be achieved on each objective [20]. Focusing on linear utility functions limits the applicability of MORL in real-world decision making problems. For example, linear utility functions cannot be used to learn policies in concave regions of the Pareto front [32]. Furthermore, if a user's preferences are non-linear, these are fundamentally incompatible with linear utility functions. In this case, strictly multi-objective methods must be used to learn optimal policies for non-linear utility functions. In MORL, for non-linear utility functions, significantly different policies are preferred when optimising under the ESR criterion versus the SER criterion [23]. It is important to note that, for linear utility functions, the distinction between ESR and SER does not exist [22].

For example, a decision maker has to choose between the following lotteries, L_1 and L_2, which are highlighted in Table 1.

L_1	
P(L_1= **R**)	**R**
0.5	(4, 3)
0.5	(2, 3)

L_2	
P(L_2=**R**)	**R**
0.9	(1, 3)
0.1	(10, 2)

Table 1: A lottery, L_1, has two possible returns, (4, 3) and (2, 3), each with a probability, p, of 0.5. A lottery, L_2, has two possible returns, (1, 3) with a probability, p of 0.9 and (10, 2) with a probability of 0.1.

The decision maker has the following non-linear utility function:

$$u(\mathbf{x}) = x_1^2 + x_2^2, \tag{6}$$

where $\mathbf{x}$ is a vector returned from **R** in Table 1, and x_1 and x_2 are the values of two objectives. Note that this utility function is monotonically increasing for $x_1 \geq 0$ and $x_2 \geq 0$. Under the SER criterion, the decision maker will compute the expected value of each lottery, apply the utility function, and select the lottery that

maximises their utility function. Let us consider which lottery the decision maker will play under the SER criterion:

$$L_1 : E(L_1) = 0.5(4,3) + 0.5(2,3) = (2,1.5) + (1,1.5)$$
$$L_1 : u(E(L_1)) = (2^2 + 1.5^2) + (1^2 + 1.5^2) = 6.25 + 3.25 = 9.5$$
$$L_2 : E(L_2) = 0.9(1,3) + 0.1(10,2) = (0.9,2.7) + (1,0.2)$$
$$L_2 : u(E(L_2)) = (0.9^2 + 2.7^2) + (1^2 + 0.2^2) = 8.1 + 1.04 = 9.14$$

Therefore, a decision maker with the utility function in Equation 6 will prefer to play lottery L_1 under the SER criterion.

Under the ESR criterion, the decision maker will first apply the utility function to the return vectors, compute the expectation, and select the lottery to maximise their utility function. Let us consider how a decision maker will choose which lottery to play under the ESR criterion:

$$L_1 : u(L_1) = u(4,3) + u(2,3) = (4^2 + 3^2) + (2^2 + 3^2) = (25) + (13)$$
$$L_1 : \mathbb{E}(u(L_1)) = 0.5(25) + 0.5(13) = 12.5 + 6.5 = 19$$
$$L_2 : u(L_2) = u(1,3) + u(10,2) = (1^2 + 3^2) + (10^2 + 2^2) = (10) + (104)$$
$$L_2 : \mathbb{E}(u(L_2)) = 0.9(10) + 0.1(104) = 9 + 10.4 = 19.4$$

Therefore, a decision maker with the utility function in Equation 6 will prefer to play lottery L_2 under the ESR criterion. From the example, it is clear that users with the same non-linear utility function can prefer different policies, depending on which multi-objective optimisation criterion is selected. Therefore, it is critical that the distinction ESR and SER is taken into consideration when selecting a MORL algorithm to learn optimal policies in a given scenario. The majority of MORL research focuses on the SER criterion [22]. By comparison, the ESR criterion has received very little attention from the MORL community [12, 22, 25, 26]. Many of the traditional MORL methods cannot be used when optimising under the ESR criterion. The fact that non-linear utility functions in MOMDPs do not distribute across the sum of immediate and future returns invalidates the Bellman equation [25],

$$\max_{\pi} \mathbb{E}\left[u\left(\mathbf{R}_t^- + \sum_{i=t}^{\infty} \gamma^i \mathbf{r}_i\right) \middle| \pi, s_t\right] \neq u(\mathbf{R}_t^-) + \max_{\pi} \mathbb{E}\left[u\left(\sum_{i=t}^{\infty} \gamma^i \mathbf{r}_i\right) \middle| \pi, s_t\right], \tag{7}$$

where u is a non-linear utility function and $\mathbf{R}_\mathbf{t}^- = \sum_{i=0}^{t-1} \gamma^i \mathbf{r}_i$.

Hayes et al. [12] implement a Distributional Monte Carlo Tree Search (DMCTS) algorithm, which learns a posterior distribution over the expected utility of individual policy executions. DMCTS achieves state-of-the-art performance under the ESR criterion. Hayes et al. [12] demonstrate that, when optimising under the ESR criterion, making decisions based on a distribution over the expected utility of the returns is crucial to learn optimal policies in realistic problems where rewards are stochastic. Traditional RL approaches use the expected value of the future returns to make decisions. The expected value cannot provide the agent with sufficient critical information to avoid adverse outcomes and exploit positive outcomes when making a decision [12].

To understand why it is critical to make decisions when optimising under the ESR criterion using a distribution over the expected utility of the returns, let us consider the following example in Table 2 regarding a human decision maker.

L_3	
P(L_3= **R**)	**R**
0.5	(-20, 1)
0.5	(20, 3)

L_4	
P(L_4=**R**)	**R**
0.9	(0, 2)
0.1	(10, 2)

Table 2: A lottery, L_3, has two possible returns, (-20, 1) and (20, 3), each with a probability of 0.5. A lottery, L_4, has two possible returns, (0, 2) with a probability of 0.9 and (10, 2) with a probability of 0.1.

The decision maker has the following non-linear utility function:

$$u(\mathbf{x}) = x_1 + x_2^2 \tag{8}$$

where $\mathbf{x}$ is a vector returned from **R** in Table 2, and x_1 and x_2 are the values of two objectives. Note that this utility function is monotonically increasing for all values of x_1 and for values of $x_2 \geq 0$.

For the non-linear utility function in Equation 8, under the ESR criterion, both L_3 and L_4 have the same expected utility value of 5. It is important to note if an agent plays lottery L_3, there is 0.5 chance of receiving a utility of -19 and a 0.5 chance of receiving a utility of 29. For a human decision maker, receiving a utility of 29 is an ideal outcome. However, receiving a utility of -19 might represent a severely negative outcome that the decision maker would want to avoid, e.g. going into debt. Instead, the decision maker may prefer lottery L_4. As shown in this example, it is crucial that a distribution over the expected utility of the returns is used when making decisions under the ESR criterion.

The current MORL literature on the ESR criterion assumes a scalar expected utility (see Section 2.3) [12, 22, 25, 26]. As demonstrated above, using a single expected value to make decisions under the ESR criterion is not sufficient to avoid choosing policies with undesirable outcomes. Therefore, it is necessary to adopt a distributional approach to ESR problems.

Firstly, we define a multi-objective version of the value distribution [6], $\mathbf{Z}^\pi$, which gives the distribution over returns of a random vector [30] when a policy π is executed, such that,

$$\mathbb{E}\,\mathbf{Z}^\pi = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathbf{r}_t \middle| \pi, \mu_0\right]. \tag{9}$$

Moreover, a value distribution can be used to represent policies. Under the ESR criterion, the utility of the value distribution, Z_u^π, is defined as a distribution over the scalar utilities received from applying the utility function to each vector in the value distribution, $\mathbf{Z}^\pi$. Therefore, Z_u^π is a distribution over the scalar utility of vector returns of a random vector received from executing a policy, π, such that,

$$\mathbb{E}\,Z_u^\pi = \mathbb{E}\left[u\left(\sum_{t=0}^{\infty} \gamma^t \mathbf{r}_t\right) \middle| \pi, \mu_0\right]. \tag{10}$$

The utility of the value distribution can only be calculated when the utility function is known a priori.

In the examples used in Section 3, the utility function of the user is known. However, many scenarios exist where the user's utility function is unknown at the time of learning [26]. In this scenario, a set of policies that are optimal for all monotonically increasing utility functions must be learned. However, for the ESR criterion, a set of optimal solutions has yet to be defined. To learn a set of optimal policies under the ESR criterion we must develop new methods.

To address this challenge, in Section 4 we apply first-order stochastic dominance to determine a partial ordering over value distributions to satisfy the ESR criterion.

4 STOCHASTIC DOMINANCE FOR ESR

For MORL there are two classes of algorithms: single-policy and multi-policy algorithms [26, 31]. When the user's utility function is known a priori, it is possible to use a single-policy algorithm [12, 25] to learn an optimal solution. However, when the user's utility function is unknown we aim to learn a set of policies that are optimal for all monotonically increasing utility functions. The current literature on the ESR criterion focuses only on scenarios where the utility function of a user is known [12, 25], overlooking scenarios where the utility function of a user is unknown. Moreover, a set of solutions under the ESR criterion for the unknown utility function scenario [26] has yet to be defined.

Various algorithms have been proposed to learn solution sets under the SER criterion (see Section 2.3), for example [18, 27, 33]. Under the SER criterion, multi-policy algorithms determine optimality by comparing policies based on the utility of vector valued expectations (Equation 4). In contrast, under the ESR criterion it is crucial to maintain a distribution over the utility of possible vector-valued outcomes. SER multi-policy algorithms cannot be used to learn optimal policies under the ESR criterion because they compute expected value vectors. It is necessary to develop new methods that can generate solution sets for the ESR criterion with unknown utilities. The development of methods that determine an optimal partial ordering over value distributions is a promising avenue to address this challenge.

First-order stochastic dominance (see Section 2.4) is a method which gives a partial ordering over random variables [15, 34]. FSD compares the cumulative distribution functions of the underlying probability distributions of random variables to determine optimality. To satisfy the ESR criterion, it is essential that the expected utility is maximised. To use FSD for the ESR criterion, we must show the FSD conditions presented in Section 2.4 also hold when optimising the expected utility for unknown monotonically increasing utility functions.

For the single-objective case, Theorem 4.1 proves for random variables X and Y, if X $\geq_{FSD}$ Y, the expected utility of X is greater than, or equal to, the expected utility of Y for monotonically increasing utility functions. In Theorem 4.1, random variables X and Y are considered, and their corresponding CDFs F_X, F_Y. The work of Mas-Colell et al. [17] is used as a foundation for Theorem 4.1.

Theorem 4.1. *A random variable, X, is preferred to a random variable, Y, for all decision makers with a monotonically increasing utility function if, and only if, X $\geq_{FSD}$ Y.*

$$X \geq_{FSD} Y \implies \mathbb{E}(u(X)) \geq \mathbb{E}(u(Y))$$

Proof. If X $\geq_{FSD}$ Y, then[2],

$$F_X(z) \leq F_Y(z), \forall\, z$$

Since,

$$\mathbb{E}(u(X)) = \int_{-\infty}^{\infty} u(z) dF_X(z)$$
$$\mathbb{E}(u(Y)) = \int_{-\infty}^{\infty} u(z) dF_Y(z)$$

When integrating both $\mathbb{E}(u(X))$ and $\mathbb{E}(u(Y))$ by parts, the following results is generated:

$$\mathbb{E}(u(X)) = [u(z)F_X(z)]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} u'(z)F_X(z)\, dz$$
$$\mathbb{E}(u(Y)) = [u(z)F_Y(z)]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} u'(z)F_Y(z)\, dz$$

Given $F_X(-\infty) = F_Y(-\infty) = 0$ and $F_X(\infty) = F_Y(\infty) = 1$, the first terms in $\mathbb{E}(u(X))$ and $\mathbb{E}(u(Y))$ are equal, and thus

$$\mathbb{E}(u(X)) - \mathbb{E}(u(Y)) = \int_{-\infty}^{\infty} u'(z)F_Y(z)\, dz - \int_{-\infty}^{\infty} u'(z)F_X(z)\, dz$$

Since $F_X(z) \leq F_Y(z)$ and $u'(z) \geq 0$ for all monotonically increasing utility functions, then

$$\mathbb{E}(u(X)) - \mathbb{E}(u(Y)) \geq 0$$

and thus,

$$\mathbb{E}(u(X)) \geq \mathbb{E}(u(Y))$$

□

A utility function maps an input (scalar or vector return) to an output (scalar utility). Since the probability of receiving some utility is equal to the probability of receiving some return for a random variable, X, we can write the following:

$$P(X > c) = P(u(X) > u(c)), \tag{11}$$

where c is a constant. Using the results shown in Theorem 4.1 and Equation 11, the FSD conditions highlighted in Section 2.4 can be rewritten to include monotonically increasing utility functions:

$$P(u(X) > u(z)) \geq P(u(Y) > u(z)) \tag{12}$$

Definition 4.2. Let X and Y be random variables. X dominates Y for all decision makers with a monotonically increasing utility function if the following is true:

$$X \geq_{FSD} Y \Leftrightarrow$$
$$\forall u : \forall v : P(u(X) > u(v)) \geq P(u(Y) > u(v)).$$

In MORL, the return from the reward function is a vector, where each element in the return vector represents an objective. To apply FSD to MORL under the ESR criterion, random vectors must be considered. In this case, a random vector (or multi-variate random variable) is a vector whose components are scalar-valued random variables on the same probability space. For simplicity, this paper focuses on the case in which a random vector has two random variables, known as the bi-variate case. FSD conditions have been proven to hold for random vectors with n random variables in the works of Sriboonchitta et al. [29], Levhari et al. [14], Nakayama et al. [19] and Scarsini [28]. In Theorem 4.3, the work of Atkinson

[2]CDFs with lower probability values for a given z are preferable. Figure 1 explains why this is the case.

and Bourguignon [2] is distilled into a suitable Theorem for MORL. Theorem 4.3 highlights how the conditions for FSD hold for random vectors while satisfying the ESR criterion for a monotonically increasing utility function, u, where $\frac{d^2u}{dx_1dx_2} \leq 0$[24]. It is important to note Atikson and Bourguignon [2] have proven Theorem 4.3 for utility functions where $\frac{d^2u}{dx_1dx_2} \geq 0$. We plan to extend these conditions for MORL in a future work. In Theorem 4.3, $\mathbf{X}$ and $\mathbf{Y}$ are random vectors where each random vector consists of two random variables, $\mathbf{X} = [X_1, X_2]$ and $\mathbf{Y} = [Y_1, Y_2]$. $F_{X_1X_2}$ and $F_{Y_1Y_2}$ are the corresponding CDFs.

Theorem 4.3. *A random vector, $\boldsymbol{X}$, is preferred to a random vector, $\boldsymbol{Y}$, by all decision makers with a monotonically increasing utility function if, and only if, $\boldsymbol{X} \geq_{FSD} \boldsymbol{Y}$.*

$$\boldsymbol{X} \geq_{FSD} \boldsymbol{Y} \implies \mathbb{E}(u(\boldsymbol{X})) \geq \mathbb{E}(u(\boldsymbol{Y}))$$

Proof. Since $\mathbf{X} \geq_{FSD} \mathbf{Y}$ means,

$$F_{X_1X_2}(t,z) \leq F_{Y_1Y_2}(t,z)$$

The expected utility can be written as follows:

$$\mathbb{E}(u(\mathbf{X})) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u(t,z) f_{X_1X_2}(t,z)\, dtdz$$

$$\mathbb{E}(u(\mathbf{Y})) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u(t,z) f_{Y_1Y_2}(t,z)\, dtdz$$

where $f_{X_1X_2}$ and $f_{Y_1Y_2}$ are the probability density functions of $\mathbf{X}$ and $\mathbf{Y}$, respectively. Only the steps for the integration of $\mathbb{E}(u(\mathbf{X}))$ are shown below, however, the steps for integration of $\mathbb{E}(u(\mathbf{Y}))$ are the same:

$$\mathbb{E}(u(\mathbf{X})) = \int_{-\infty}^{\infty} \Bigg([u(t,z)F_{X_1X_2}(t,z)]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \frac{du}{dt}(t,z) F_{X_1X_2}(t,z)\, dt \Bigg) dz$$

$$= \int_{-\infty}^{\infty} \Bigg(\lim_{z\to\infty} u(t,z)F_{X_1X_2}(t,z) - \lim_{z\to-\infty} u(t,z)F_{X_1X_2}(t,z) - \int_{-\infty}^{\infty} \frac{du}{dt}(t,z) F_{X_1X_2}(t,z)\, dt \Bigg) dz$$

Given $\lim_{z\to-\infty} F_{X_1X_2}(t,z) = 0$ and $\lim_{z\to\infty} F_{X_1X_2}(t,z) = F_{X_1}(t)$:

$$\int_{-\infty}^{\infty} \lim_{z\to\infty} u(t,z)F_{X_1X_2}(t,z)\, dt - \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{du}{dt}(t,z) F_{X_1X_2}(t,z)\, dtdz$$

Integrating the first term gives the following:

$$= \lim_{t\to\infty} u(t,\infty)F_{X_1}(t) - \lim_{t\to-\infty} u(t,\infty)F_{X_1}(-\infty) - \int_{-\infty}^{\infty} \frac{du}{dt}(t,\infty) F_{X_1}(t)\, dt - \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{du}{dt}(t,z) F_{X_1X_2}(t,z)\, dt\, dz$$

Given $F_{X_1}(-\infty) = 0$, $F_{X_1}(\infty) = 1$ and $u(\infty,\infty) = \infty$ or $-\infty$.

$$= - \int_{-\infty}^{\infty} \frac{du}{dt}(t,\infty) F_{X_1}(t)\, dt - \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{du}{dt}(t,z) F_{X_1X_2}(t,z)\, dt\, dz$$

Then, integrating the second term gives the following:

$$= - \int_{-\infty}^{\infty} \frac{du}{dt}(t,\infty) F_{X_1}(t)\, dt - \int_{-\infty}^{\infty} \Bigg(\lim_{t\to\infty} \frac{du}{dz}(\infty,z) F_{X_1X_2}(\infty,z) - \lim_{t\to-\infty} \frac{du}{dz}(-\infty,z) F_{X_1X_2}(-\infty,z) - \int_{-\infty}^{\infty} \frac{d^2u}{dtdz}(t,z) F_{X_1X_2}(t,z) \Bigg) dt\, dz$$

Given $F_{X_1X_2}(-\infty,z) = 0$ and $F_{X_1X_2}(\infty,z) = F_{X_2}(z)$, then:

$$\mathbb{E}(u(\mathbf{X})) = - \int_{-\infty}^{\infty} \frac{du}{dt}(t,\infty) F_{X_1}(t)\, dt - \int_{-\infty}^{\infty} \frac{du}{dz}(\infty,z) F_{X_2}(z)\, dz + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{d^2u}{dtdz}(t,z) F_{X_1X_2}(t,z)\, dt\, dz$$

Therefore,

$$\mathbb{E}(u(\mathbf{Y})) = - \int_{-\infty}^{\infty} \frac{du}{dt}(t,\infty) F_{Y_1}(t)\, dt - \int_{-\infty}^{\infty} \frac{du}{dz}(\infty,z) F_{Y_2}(z)\, dz + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{d^2u}{dtdz}(t,z) F_{Y_1Y_2}(t,z)\, dt\, dz$$

$$\mathbb{E}(u(\mathbf{X})) - \mathbb{E}(u(\mathbf{Y})) = - \int_{-\infty}^{\infty} \frac{du}{dt}(t,\infty) F_{X_1}(t)\, dt - \int_{-\infty}^{\infty} \frac{du}{dz}(\infty,z) F_{X_2}(z)\, dz + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{d^2u}{dtdz}(t,z) F_{X_1X_2}(t,z)\, dt\, dz + \int_{-\infty}^{\infty} \frac{du}{dt}(t,\infty) F_{Y_1}(t)\, dt + \int_{-\infty}^{\infty} \frac{du}{dz}(\infty,z) F_{Y_2}(z)\, dz - \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{d^2u}{dtdz}(t,z) F_{Y_1Y_2}(t,z)\, dt\, dz$$

For a monotonically increasing utility function $\frac{du}{dt} \geq 0$, $\frac{du}{dz} \geq 0$ and $\frac{d^2u}{dtdz} \leq 0$. Given, the utility function, u, is assumed to be monotonically increasing and for FSD $F_X(t,z) \leq F_{\mathbf{Y}}(t,z)$ which gives the following:

$$\mathbb{E}(u(\mathbf{X})) - \mathbb{E}(u(\mathbf{Y})) \geq 0.$$

Finally,

$$\mathbb{E}(u(\mathbf{X})) \geq \mathbb{E}(u(\mathbf{Y})).$$

□

Using the results from Theorem 4.3, Equation 12 can be updated to include random vectors,

$$P(u(\mathbf{X}) > u(\mathbf{z})) \geq P(u(\mathbf{Y}) > u(\mathbf{z})). \tag{13}$$

Definition 4.4. For random vectors $\mathbf{X}$ and $\mathbf{Y}$, $\mathbf{X}$ is preferred over $\mathbf{Y}$ by all decision makers with a monotonically increasing utility function if, and only if, the following is true:

$$\mathbf{X} \geq_{FSD} \mathbf{Y} \Leftrightarrow$$
$$\forall u : (\forall \mathbf{v} : P(u(\mathbf{X}) > u(\mathbf{v})) \geq P(u(\mathbf{Y}) > u(\mathbf{v}))$$

Using the results from Theorem 4.3 and Definition 4.4, it is possible to extend FSD to MORL. For MORL, under the ESR criterion, the value distribution, $\mathbf{Z}^{\pi}$, is considered to be the full distribution of the returns of a random vector received when executing a policy, π (see Section 3). Value distributions can be used to represent policies. In this case, it is possible to use FSD to obtain a partial ordering over policies. For example, consider two policies, π and π', where each policy has the underlying value distribution $\mathbf{Z}^{\pi}$ and $\mathbf{Z}^{\pi'}$. If $\mathbf{Z}^{\pi} \geq_{FSD} \mathbf{Z}^{\pi'}$ then π will be preferred over π'.

Definition 4.5. Policies π and π' have value distributions $\mathbf{Z}^{\pi}$ and $\mathbf{Z}^{\pi'}$. Policy π is preferred over policy π' by all decision makers with a utility function, u, that is monotonically increasing if, and only if, the following is true:

$$\mathbf{Z}^{\pi} \geq_{FSD} \mathbf{Z}^{\pi'}.$$

Now that a partial ordering over policies has been defined under the ESR criterion for the unknown utility function scenario, it is possible to define a set of optimal policies.

5 SOLUTION SETS FOR ESR

Section 4 defines a partial ordering over policies under the ESR criterion for unknown utility using FSD. In the unknown utility function scenario it is infeasible to learn a single optimal policy [26]. When a user's utility function is unknown, multi-policy MORL algorithms must be used to learn a set of optimal policies. To apply MORL to the ESR criterion in scenarios with unknown utility, a set of optimal policies under the ESR criterion must be defined. In Section 5, FSD is used to define multiple sets of optimal policies for the ESR criterion.

Firstly, a set of optimal policies, known as the undominated set, is defined. The undominated set is defined using FSD, where each policy in the undominated set has an underlying value distribution that is FSD dominant. The undominated set contains at least one optimal policy for all possible monotonically increasing utility functions.

Definition 5.1. The undominated set, $U(\Pi)$, is a sub-set of all possible policies for where there exists some utility function, u, where a policy's value distribution is FSD dominant.

$$U(\Pi) = \left\{ \pi \in \Pi \;\middle|\; \exists u, \forall \pi' \in \Pi : \mathbf{Z}^{\pi} \geq_{FSD} \mathbf{Z}^{\pi'} \right\}$$

However, the undominated set may contain excess policies. For example, under FSD, if two dominant policies have value distributions that are equal, then both policies will be in the undominated set. Given both value distributions are equal, a user with a monotonically increasing utility function will not prefer one policy over the other. In this case, both policies have the same expected utility. To reduce the number of policies that must be considered at execution time, for each possible utility function we can keep just one corresponding FSD dominant policy; such a set of policies is called a coverage set (CS).

Definition 5.2. The coverage set, $CS(\Pi)$, is a subset of the undominated set, $U(\Pi)$, where, for every utility function, u, the set contains a policy that has a FSD dominant value distribution,

$$CS(\Pi) \subseteq U(\Pi) \wedge \left(\forall u, \exists \pi \in CS(\Pi), \forall \pi' \in \Pi : \mathbf{Z}^{\pi} \geq_{FSD} \mathbf{Z}^{\pi'} \right)$$

In practice, for scenarios where the utility function is unknown, it is difficult to compute the undominated set or coverage set using FSD because FSD relies on having a user's utility function available to calculate dominance. To address this challenge, expected scalarised returns (ESR) dominance is defined. Multi-policy algorithms can use ESR dominance as a method under the ESR criterion to learn a set of optimal policies.

Definition 5.3. For random vectors $\mathbf{X}$ and $\mathbf{Y}$, $\mathbf{X} >_{ESR} \mathbf{Y}$ for all decision makers with a monotonically increasing utility function if, and only if, the following is true:

$$\mathbf{X} >_{ESR} \mathbf{Y} \Leftrightarrow$$
$$\forall u : (\forall \mathbf{v} : P(u(\mathbf{X}) > u(\mathbf{v})) \geq P(u(\mathbf{Y}) > u(\mathbf{v}))$$
$$\wedge \exists \mathbf{v} : P(u(\mathbf{X}) > u(\mathbf{v})) > P(u(\mathbf{Y}) > u(\mathbf{v}))).$$

ESR dominance (Definition 5.3) extends FSD, however, ESR dominance is a more strict dominance criterion. For FSD, policies that have equal value distributions are considered dominant policies, which is not the case under ESR dominance. Therefore, if a random vector is ESR dominant, the random vector has a greater expected utility than all ESR dominated random vectors. Theorem 5.4 proves that ESR dominance satisfies the ESR criterion when the utility function of the user is unknown for all monotonically increasing utility functions. Theorem 5.4 focuses on random vectors $\mathbf{X}$ and $\mathbf{Y}$ where each random vector has two random variables, such that $\mathbf{X} = [X_1, X_2]$ and $\mathbf{Y} = [Y_1, Y_2]$. $F_{\mathbf{X}}$ and $F_{\mathbf{Y}}$ are the corresponding CDFs and $\mathbf{v} = [t, z]$. However, Theorem 5.4 can easily be extended for random vectors with n random variables ($\mathbf{X} = [X_1, X_2, ..., X_n]$).

Theorem 5.4. *A random vector, $\boldsymbol{X}$, is preferred to a random vector, $\boldsymbol{Y}$, by all decision makers with a monotonically increasing utility function if, and only if, $\boldsymbol{X} \geq_{ESR} \boldsymbol{Y}$:*

$$\boldsymbol{X} >_{ESR} \boldsymbol{Y} \implies \mathbb{E}(u(\boldsymbol{X})) > \mathbb{E}(u(\boldsymbol{Y}))$$

Proof. $\mathbf{X}$ and $\mathbf{Y}$ are random vectors with n random variables. If $\mathbf{X} >_{ESR} \mathbf{Y}$ the following two conditions must be met for all u:

(1) $\forall \mathbf{v} : P(u(\mathbf{X}) > u(\mathbf{v})) \geq P(u(\mathbf{Y}) > u(\mathbf{v}))$
(2) $\exists \mathbf{v} : P(u(\mathbf{X}) > u(\mathbf{v})) > P(u(\mathbf{Y}) > u(\mathbf{v}))$

From Definition 4.4, if $\mathbf{X} \geq_{FSD} \mathbf{Y}$ then the following is true:

$$\forall u : \forall \mathbf{v} : P(u(\mathbf{X}) > u(\mathbf{v})) \geq P(u(\mathbf{Y}) > u(\mathbf{v}))$$

If $\mathbf{X} \geq_{FSD} \mathbf{Y}$, then, from Theorem 4.3, the following is true:

$$\mathbb{E}(u(\mathbf{X})) \geq \mathbb{E}(u(\mathbf{Y}))$$

If condition 1 is satisfied, the expected utility of $\mathbf{X}$ is at least equal to the expected utility of $\mathbf{Y}$, then:

$$\mathbb{E}(u(\mathbf{X})) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} u(t,z) f_{\mathbf{X}}(t,z)\, dt\, dz$$
$$\mathbb{E}(u(\mathbf{Y})) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} u(t,z) f_{\mathbf{Y}}(t,z)\, dt\, dz$$

In order to satisfy condition 2, some limits must exist to give the following,

$$\int_a^b \int_c^d u(t,z) f_{\mathbf{X}}(t,z)\, dt\, dz > \int_a^b \int_c^d u(t,z) f_{\mathbf{Y}}(t,z)\, dt\, dz$$

The minimum requirement to satisfy condition 1 is:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} u(t,z) f_{\mathbf{X}}(t,z)\, dt\, dz = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} u(t,z) f_{\mathbf{Y}}(t,z)\, dt\, dz$$

If condition 1 is satisfied, to satisfy condition 2 some limits must exist:

$$\int_a^b \int_c^d u(t,z) f_{\mathbf{X}}(t,z)\, dt\, dz > \int_a^b \int_c^d u(t,z) f_{\mathbf{Y}}(t,z)\, dt\, dz.$$

Therefore,

$$\int_{-\infty}^{a} \int_{-\infty}^{c} u(t,z) f_{\mathbf{X}}(t,z)\, dt\, dz + \int_a^b \int_c^d u(t,z) f_{\mathbf{X}}(t,z)\, dt\, dz +$$
$$\int_b^{\infty} \int_d^{\infty} u(t,z) f_{\mathbf{X}}(t,z)\, dt\, dz > \int_{-\infty}^{a} \int_{-\infty}^{c} u(t,z) f_{\mathbf{Y}}(t,z)\, dt\, dz +$$
$$\int_a^b \int_c^d u(t,z) f_{\mathbf{Y}}(t,z)\, dt\, dz + \int_b^{\infty} \int_d^{\infty} u(t,z) f_{\mathbf{Y}}(t,z)\, dt\, dz.$$

Finally,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u(t,z) f_{\mathbf{X}}(t,z)\, dt\, dz > \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u(t,z) f_{\mathbf{Y}}(t,z)\, dt\, dz$$

if $\mathbf{X} >_{ESR} \mathbf{Y}$, then,

$$\mathbb{E}(u(\mathbf{X})) > \mathbb{E}(u(\mathbf{Y})).$$

□

In the ESR dominance criterion defined in Definition 5.3, the utility of different vectors is compared. However, it is not possible to calculate the utility of a vector when the utility function is unknown. In this case, Pareto dominance [21] can be used instead to determine the relative utility of the vectors being compared.

Definition 5.5. **A** Pareto dominates ($\succ_p$) **B** if the following is true:

$$\mathbf{A} >_p \mathbf{B} \Leftrightarrow (\forall i : \mathbf{A}_i \geq \mathbf{B}_i) \wedge (\exists i : \mathbf{A}_i > \mathbf{B}_i). \tag{14}$$

For monotonically increasing utility functions, if the value of an element of the vector increases, then the scalar utility of the vector also increases. Therefore, using Definition 5.5, if vector **A** Pareto dominates vector **B**, for a monotonically increasing utility function, **A** has a higher utility than **B**. To make ESR comparisons between value distributions, Pareto dominance can be used.

Definition 5.6. For random vectors **X** and **Y**, $\mathbf{X} >_{ESR} \mathbf{Y}$ for all monotonically increasing utility functions if, and only if, the following is true:

$$\mathbf{X} >_{ESR} \mathbf{Y} \Leftrightarrow$$

$$\forall \mathbf{v} : P(\mathbf{X} >_P \mathbf{v}) \geq P(\mathbf{Y} >_P \mathbf{v}) \wedge \exists \mathbf{v} : P(\mathbf{X} >_P \mathbf{v}) > P(\mathbf{Y} >_P \mathbf{v}).$$

Therefore, as per Definition 5.7, ESR dominance can be used to give a partial ordering over policies.

Definition 5.7. For value distributions $\mathbf{Z}^{\pi}$ and $\mathbf{Z}^{\pi'}$ for policies π and π', π is preferred over π' by all decision makers with a monotonically increasing utility function if, and only if, the following is true:

$$\mathbf{Z}^{\pi} >_{ESR} \mathbf{Z}^{\pi'}$$

Using ESR dominance, it is possible to define a set of optimal policies, known as the ESR set.

Definition 5.8. The ESR set, $ESR(\Pi)$, is a sub-set of all policies where each policy in the ESR set is ESR dominant,

$$ESR(\Pi) = \{\pi \in \Pi \mid \nexists \pi' \in \Pi : \mathbf{Z}^{\pi'} >_{ESR} \mathbf{Z}^{\pi}\}. \tag{15}$$

The ESR set is a set of non-dominated policies, where each policy in the ESR set is ESR dominant. The ESR set can be considered a coverage set, given no excess policies exist in the ESR set. It is viable for a multi-policy MORL method to use ESR dominance to construct the ESR set, given Pareto dominance is used to determine ESR dominance when the utility function of a user is unknown.

6 RELATED WORK

The various orders of stochastic dominance have been used extensively as a method to determine the optimal decision when making decisions under uncertainty in economics [7], finance [1, 4], game theory [9], and various other real-world scenarios [5]. However, stochastic dominance has largely been overlooked in systems that learn. Cook and Jarret [8] use various orders of stochastic dominance and Pareto dominance with genetic algorithms to compute optimal solution sets for an aerospace design problem with multiple objectives when constrained by a computational budget. Martin et al. [16] use second-order stochastic dominance (SSD) with a single-objective distributional RL algorithm [6]. Martin et al. [16] use SSD to determine the optimal action to take at decision time, and this approach is shown to learn good policies during experimentation.

7 CONCLUSION & FUTURE WORK

The ESR criterion has largely been ignored by the MORL community, with the exception of the work of Roijers et al. [25, 26] and Hayes et al. [11, 12]. While these works present single-policy algorithms that are suitable to learn policies under the ESR criterion, a formal definition of the necessary requirements to satisfy the ESR criterion had not previously been defined. In Section 3, we outline, through examples and definitions, the necessary requirements to satisfy the ESR criterion. The formal definitions outlined in Section 3 ensure that an optimal policy can be learned when the utility function of the user is known under the ESR criterion. However, in the real world, a user's preferences over objectives (or utility function) may be unknown at the time of learning.

Prior to this paper, a suitable solution set for the unknown utility function scenario under the ESR criterion had not been defined. This long-standing research gap has restricted the applicability of MORL in real-world scenarios under the ESR criterion. In Section 4 and Section 5 we define the necessary solution sets required for multi-policy algorithms to learn a set of optimal policies under the ESR criterion when the utility function of a user is unknown. This work aims to answer some of the existing research questions regarding the ESR criterion. Moreover, we aim to highlight the importance of the ESR criterion when applying MORL to real-world scenarios. In order to successfully apply MORL to the real world, we must implement new single-policy and multi-policy algorithms that can learn solutions for non-linear utility functions in various scenarios.

A promising starting point for future work would be to learn a set of optimal solutions under the ESR criterion in a multi-objective multi-armed bandit setting. Learning an optimal set of policies in a bandit setting is a natural starting point for any new multi-policy algorithm and would require implementing the new dominance criteria outlined in this paper.

ACKNOWLEDGEMENTS

Conor F. Hayes is funded by the National University of Ireland Galway Hardiman Scholarship. This research was supported by funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" program.

REFERENCES

[1] Mukhtar M. Ali. 1975. Stochastic dominance and portfolio analysis. *Journal of Financial Economics* 2, 2 (1975), 205–229. https://doi.org/10.1016/0304-405X(75)90005-7

[2] A. B. Atkinson and F. Bourguignon. 1982. The Comparison of Multi-Dimensioned Distributions of Economic Status. *The Review of Economic Studies* 49, 2 (04 1982), 183–201. https://doi.org/10.2307/2297269 arXiv:https://academic.oup.com/restud/article-pdf/49/2/183/4720580/49-2-183.pdf

[3] Vijay S. Bawa. 1975. Optimal rules for ordering uncertain prospects. *Journal of Financial Economics* 2, 1 (1975), 95 – 121. https://doi.org/10.1016/0304-405X(75)90025-2

[4] Vijay S. Bawa. 1978. Safety-First, Stochastic Dominance, and Optimal Portfolio Choice. *The Journal of Financial and Quantitative Analysis* 13, 2 (1978), 255–271. http://www.jstor.org/stable/2330386

[5] Vijay S. Bawa. 1982. Research Bibliography-Stochastic Dominance: A Research Bibliography. *Manage. Sci.* 28, 6 (June 1982), 698–712. https://doi.org/10.1287/mnsc.28.6.698

[6] Marc G Bellemare, Will Dabney, and Rémi Munos. 2017. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*. PMLR, Sydney, 449–458.

[7] E. Choi and Stanley Johnson. 1988. Stochastic Dominance and Uncertain Price Prospects. *Center for Agricultural and Rural Development (CARD) at Iowa State University, Center for Agricultural and Rural Development (CARD) Publications* 55 (01 1988). https://doi.org/10.2307/1059583

[8] Laurence Cook and Jerome Jarrett. 2018. Using Stochastic Dominance in Multi-Objective Optimizers for Aerospace Design Under Uncertainty. https://doi.org/10.2514/6.2018-0665

[9] Peter C Fishburn. 1978. Non-cooperative stochastic dominance games. *International Journal of Game Theory* 7, 1 (1978), 51–61.

[10] Josef Hadar and William R. Russell. 1969. Rules for Ordering Uncertain Prospects. *The American Economic Review* 59, 1 (1969), 25–34. http://www.jstor.org/stable/1811090

[11] Conor F Hayes, Mathieu Reymond, Diederik M Roijers, Enda Howley, and Patrick Mannion. 2021. Risk-Aware and Multi-Objective Decision Making with Distributional Monte Carlo Tree Search. *arXiv preprint arXiv:2102.00966* (2021). https://arxiv.org/abs/2102.00966

[12] Conor F Hayes, Mathieu Reymond, Diederik M Roijers, Enda Howley, and Patrick Mannion. 2021 In Press. Distributional Monte Carlo Tree Search for Risk-Aware and Multi-Objective Reinforcement Learning. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, Vol. 2021. IFAAMAS.

[13] Conor F. Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M. Zintgraf, Richard Dazeley, Fredrik Heintz, Enda Howley, Athirai A. Irissappane, Patrick Mannion, Ann Nowé, Gabriel Ramos, Marcello Restelli, Peter Vamplew, and Diederik M. Roijers. 2021. A Practical Guide to Multi-Objective Reinforcement Learning and Planning. arXiv:2103.09568 [cs.AI]

[14] David Levhari, Jacob Paroush, and Bezalel Peleg. 1975. Efficiency Analysis for Multivariate Distributions. *The Review of Economic Studies* 42, 1 (1975), 87–91. http://www.jstor.org/stable/2296822

[15] Haim Levy. 1992. Stochastic Dominance and Expected Utility: Survey and Analysis. *Management Science* 38, 4 (1992), 555–593. http://www.jstor.org/stable/2632436

[16] John Martin, Michal Lyskawinski, Xiaohu Li, and Brendan Englot. 2020. Stochastically Dominant Distributional Reinforcement Learning. In *International Conference on Machine Learning*. PMLR, 6745–6754.

[17] Andreu Mas-Colell, Michael Dennis Whinston, Jerry R Green, et al. 1995. *Microeconomic theory*. Vol. 1. Oxford university press New York.

[18] Kristof Van Moffaert and Ann Nowé. 2014. Multi-Objective Reinforcement Learning using Sets of Pareto Dominating Policies. *Journal of Machine Learning Research* 15, 107 (2014), 3663–3692. http://jmlr.org/papers/v15/vanmoffaert14a.html

[19] H. Nakayama, T. Tanino, and Y. Sawaragi. 1981. Stochastic Dominance for Decision Problems with Multiple Attributes and/or Multiple Decision-Makers. *IFAC Proceedings Volumes* 14, 2 (1981), 1397 – 1402. https://doi.org/10.1016/S1474-6670(17)63673-5 8th IFAC World Congress on Control Science and Technology for the Progress of Society, Kyoto, Japan, 24-28 August 1981.

[20] David O'Callaghan and Patrick Mannion. 2021. Exploring the Impact of Tunable Agents in Sequential Social Dilemmas. *arXiv preprint: arXiv:2101.11967* (2021). https://arxiv.org/abs/2101.11967

[21] Vilfredo Pareto. 1896. *Manuel d'Economie Politique*. Vol. 1. Giard, Paris.

[22] Roxana Rădulescu, Patrick Mannion, Diederik M. Roijers, and Ann Nowé. 2020. Multi-objective multi-agent decision making: a utility-based analysis and survey. *Autonomous Agents and Multi-Agent Systems* 34, 10 (2020).

[23] Roxana Rădulescu, Patrick Mannion, Yijie Zhang, Diederik M Roijers, and Ann Nowé. 2020. A utility-based analysis of equilibria in multi-objective normal-form games. *The Knowledge Engineering Review* 35 (2020).

[24] Scott F. Richard. 1975. Multivariate Risk Aversion, Utility Independence and Separable Utility Functions. *Management Science* 22, 1 (1975), 12–21. http://www.jstor.org/stable/2629784

[25] Diederik M Roijers, Denis Steckelmacher, and Ann Nowé. 2018. Multi-objective Reinforcement Learning for the Expected Utility of the Return. In *Proceedings of the Adaptive and Learning Agents workshop at FAIM 2018*.

[26] Diederik M Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. 2013. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research* 48 (2013), 67–113.

[27] Diederik M. Roijers, Shimon Whiteson, and Frans A. Oliehoek. 2014. Linear Support for Multi-Objective Coordination Graphs. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems* (Paris, France) *(AAMAS '14)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1297–1304.

[28] Marco Scarsini. 1988. Dominance Conditions for Multivariate Utility Functions. *Management Science* 34, 4 (1988), 454–460. http://www.jstor.org/stable/2631934

[29] Songsak Sriboonchitta, Wing-Keung Wong, s Dhompongsa, and Hung Nguyen. 2009. *Stochastic Dominance and Applications to Finance, Risk and Economics*. https://doi.org/10.1201/9781420082678

[30] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA.

[31] Peter Vamplew, Richard Dazeley, Adam Berry, Rustam Issabekov, and Evan Dekker. 2011. Empirical evaluation methods for multiobjective reinforcement learning algorithms. *Machine Learning* 84 (07 2011), 51–80. https://doi.org/10.1007/s10994-010-5232-5

[32] Peter Vamplew, John Yearwood, Richard Dazeley, and Adam Berry. 2008. On the Limitations of Scalarisation for Multi-objective Reinforcement Learning of Pareto Fronts. In *AI 2008: Advances in Artificial Intelligence*, Wayne Wobcke and Mengjie Zhang (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 372–378.

[33] Weijia Wang and Michèle Sebag. 2012. Multi-objective Monte-Carlo Tree Search *(Proceedings of Machine Learning Research, Vol. 25)*, Steven C. H. Hoi and Wray Buntine (Eds.). PMLR, Singapore, 507–522.

[34] Elmar Wolfstetter. 1999. *Topics in Microeconomics: Industrial Organization, Auctions, and Incentives*. Cambridge University Press. https://doi.org/10.1017/CBO9780511625787