Priority-Objective Reinforcement Learning

Yusuf Al-Husaini, Matthias Rolf
School of Engineering, Computing and Mathematics
Oxford Brookes University, Oxford, UK

***Abstract*—Intelligent agents often have to cope with situations in which their various needs must be prioritised. Efforts have been made, in the fields of cognitive robotics and machine learning, to model need prioritization. Examples of existing frameworks include normative decision theory, the subsumption architecture and reinforcement learning. Reinforcement learning algorithms oriented towards active goal prioritization include the options framework from hierarchical reinforcement learning and the ranking approach as well as the MORE framework from multi-objective reinforcement learning. Previous approaches can be configured to make an agent function optimally in individual environments, but cannot effectively model dynamic and efficient goal selection behaviour in a generalisable framework. Here, we propose an altered version of the MORE framework that includes a threshold constant in order to guide the agent towards making economic decisions in a broad range of 'priority-objective reinforcement learning' (PORL) scenarios. The results of our experiments indicate that pre-existing frameworks such as the standard linear scalarization, the ranking approach and the options framework are unable to induce opportunistic objective optimisation in a diverse set of environments. In particular, they display strong dependency on the exact choice of reward values at design time. However, the modified MORE framework appears to deliver adequate performance in all cases tested. From the results of this study, we conclude that employing MORE along with integrated thresholds, can effectively simulate opportunistic objective prioritization in a wide variety of contexts.**

I. Introduction

Artificial agents that face conflicting needs are required to make choices of which goal it should prioritize based on what it has to spare. An agent that can think both critically and rationally in these pressurised situations is much desired. Models of need prioritization have been employed in the field of human psychology. Self preservation for instance takes president over self esteem in Maslow's classical 'Hierarchy of Needs' [1]. In addition, the concepts of goal conflict and goal hierarchy have been employed in health behaviour models [2]. For an artificial agent to comply with ethical guidelines, the ability to draw effective compromises between conflicting objectives is essential [3]. Many problems in machine learning including those in reinforcement learning require the active prioritization of objectives. An agent is usually expected to focus on the secondary objective after it has satisfied the basic survival needs. Different lines of research in both cognitive robotics and machine learning have attempted to model this type of hierarchical objective behaviour. For example the subsumption architecture is typically employed in cognitive robotics, as well as trial and error approaches [4]. Methods employed in machine learning to tackle these problems primarily revolve around the concept of reinforcement learning [5], [6]. However, these approaches come with drawbacks. For instance, there is, no comprehensive model to describe how an agent should behave when a hierarchical objective is no longer possible to satisfy. In addition, it appears that there are no hierarchical models that provide objective functions with smooth transitions from one objective to the other. We use the term 'Priority-Objective Reinforcement Learning' (PORL) to refer to all frameworks in which an agent switches objectives according to certain conditions in the environment. In this study we explore three different methods of implementing PORL namely, the ranking approach [6], the options framework [7] and a modified version of the MORE framework [8]. We also use the standard linear scalarization of objectives as a baseline. A depiction of a situation that can be addressed with priority-objective reinforcement learning is shown below in figure 1:

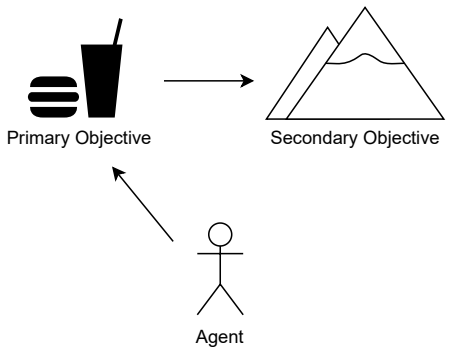


Fig. 1: Priority-Objective Reinforcement Learning Problem

A. Opportunistic Prioritization Of Objectives

When an agent is tasked with an overall goal consisting of primary objectives such as survival needs, the agent should be incentivised to satisfy the primary objectives first before attempting to complete the secondary objective. Moreover, even if a primary objective may be impossible to satisfy, the agent should still attempt to optimise it but in the event of an objective becoming impossible to optimise, there are distinct cases that need to be considered. For example, when an agent is in need of energy but there is no obtainable energy present in the environment, it may still be desirable for the agent to focus on scoring points instead. This process is analogous to intrinsic motivation [9] and enables an agent to be more opportunistic about obtaining rewards altogether. However,

this type of behaviour may not be desirable if the satisfaction of one objective is necessary to prevent the agent from leaving the environment, for example when an autonomous vehicle runs out of fuel and can no longer function. In addition, it is necessary to consider what should happen when certain areas within the environment provide resources and goal satisfaction in less favourable quantities but help to optimise more than one objective at the same time. Should an agent choose areas nearby that better optimise specific objectives or should it aim to optimize objectives in a more or less equal manner despite the hierarchical structure of importance? Furthermore, it is necessary to consider what should happen when the rewards of one or more objectives are constrained, for example when survival resources are scarce. In the field of machine learning, two main lines of research are focused on problems of this nature, namely, hierarchical and multi-objective reinforcement learning.

B. Hierarchical Reinforcement Learning

Hierarchical reinforcement learning (HRL) involves an agent learning to satisfy a series of objectives in a hierarchical manner [5]. In HRL, a policy is typically comprised of sub policies which an agent will aim to optimise prior to learning the optimal policy at the higher level. For tackling problems of this nature, Sutton et al., [7] proposed the options framework. HRL can be considered a type of PORL. However, in this case, the environment encodes the rules of prioritization. In this study we make use of this paradigm for a portion of the experiments we carry out. Alternative HRL approaches include feudal reinforcement learning and the option-critic framework [10], [11]. HRL methods use the principle of abstraction in order to guide an agent through a multi-layered objective problem such as a set of puzzles within a maze. However, they may not be perfectly suited towards problems that require the active switching between two objectives concurrently. These hierarchical methods rely partly on reward shaping [12].

C. Multi-Objective Reinforcement Learning

In multi-objective reinforcement learning (MORL), an agent needs to satisfy more than one objective concurrently [13]. MORL methods can be broadly divided into two types, namely single and multi-policy approaches [14], [15]. Single policy approaches attempt to find a single optimal policy whilst multi-policy methods attempt to find a set of optimal policies for a user to select from. This differs from HRL in which an agent only has one objective to satisfy at any given time. However, MORL methods can be adapted to suit hierarchical objective situations. For example the ranking approach [6] makes use of thresholds and objective ordering to encourage the agent to optimize objectives one at a time. The "Multi-Objective Reward Exponentials" (MORE) framework [8], on the other hand, takes a more dynamic and continuous approach whereby the accumulated past rewards for all the objectives are actively used for dynamically weighting the future values of a given objective. The values of each objective are transformed according to an exponential function which acts as a deficit model, and makes the agent focus on the least achieved objective. The optimal weight dynamics to ensure a balanced need achievement can be derived analytically in this case [8].

D. Contribution and Outline

This paper investigates how the non-linearity of MORE [8] can be utilized to achieve a robust objective prioritization that is not possible with linear approaches. In particular, we demonstrate that this can be achieved by an operation as simple as shifting the reward related to one objective up or down by a threshold constant. While linear reinforcement learning is invariant to such shifts, we demonstrate a profound and meaningful effect within MORE: Subtracting a fixed value from high-priority objective pushes the objective function into the deficit zone, which means MORE will give the objective higher overall priority, but will still give way to secondary objectives once the priority objective is well attained. Our main perspective is that of reward shaping [16] and AI alignment [17], i.e. to give the designer of an agent strong and intuitive command over the agent's eventual behavior in alignment with the design intention.

Results are shown in a 3 by 3 gridworld environment, which has a rather simple and small state space, but allows for a very comprehensive analysis of various challenging need conflicts. In particular, we investigate various combinations of scarcity and abundance of a low- and high-priority objective in order to demonstrate the strongly non-linear semantics of priority. Results are shown to be robust with respect to the exact numerical choice of reward, which is demonstrated by reporting results for a differently scaled reward. Alongside MORE, three distinct existing methods are adopted for PORL and used as baselines, two based on MORL, and one based on HRL.

II. Methodology

We make use of two main approaches to comprehensively tackle priority-objective scenarios. The first of these approaches is hierarchical reinforcement learning from which the options framework is employed. For the second approach, we make use of two multi-objective reinforcement learning frameworks, namely the ranking approach and the MORE approach. In addition, we make use of the standard linear scalarization method as a baseline to compare the frameworks. In the standard MORL approach, the total expected sum of future rewards across all K objectives i.e. the overall value for a given state is calculated on the basis of a weighted sum of estimated values for each objective where the weighting $\mathbf{W} = (W_0, W_1, ...W_{K-1})$ is static and specified by the user:

$$V^{\text{LIN}} = \sum W_k V_k^{\text{LIN}} = \sum W_k E\left[\sum_t \gamma^t r_k(t)\right] . \quad (1)$$

A. The Ranking Approach

The ranking approach [6] consists of a two-layer action selection process. In the multi-objective setting, Q-learning algorithms update vectors of values for each state-action pair.

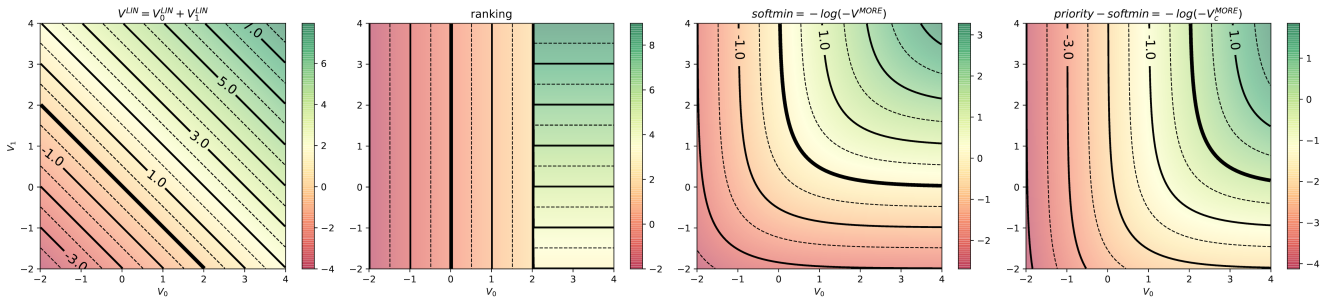


Fig. 2: Utility concepts for the multi-objective problem: (i) standard linear scalarization [8], (ii) ranking, in which the first objective always dominates the second one until a threshold of $c_0 = 2$ is reached, (iii) the non-linear MORE scalarization that corresponds to a softmin function [8], (iv) a MORE scalarization that is shifted by $c_0 = 2$ on the first objective, giving it higher priority, but also continuously weighting in the second objective when the first one becomes satisfied.

The components of these 'Q-vectors' represent the individual Q-values for each objective. The first stage of the ranking procedure involves capping the elements of the Q-vectors corresponding to a given objective for each candidate action to a threshold value. The next stage involves choosing the maximum of these capped values for the candidate actions. Should two actions have the same value for a given objective, the values of the next objective are used to further assess which action should be chosen.

To be more specific, a function $CQ_{s,a,k}$ is applied to each Q-vector component corresponding to each objective k using a threshold c_k except for the final component i.e. the last objective. The function is defined as follows:

$$CQ_{s,a,k} = \min(Q(s,a,k), c_k) \tag{2}$$

An action a' is then selected in a state s such that $\text{superior}(CQ_{s,a'}, CQ_{s,a}, 0)$ is true for all available actions. The $\text{superior}(CQ_{s,a'}, CQ_{s,a}, k)$ function as described in [6] is defined recursively as follows:

Algorithm 1 Superior Function

```
Input CQ_{s,a}, CQ_{s,a'} ,k:
if CQ_{s,a',k} > CQ_{s,a,k} then
    return true
else if CQ_{s,a',k} = CQ_{s,a,k} then
    if k = K then
        return true
    else
        return superior(CQ_{s,a}, CQ_{s,a'}, k+1)
    end if
else
    return false
end if
```

This preference implies that given two choices, the decision is made based solely on the primary objective, unless both choices have exactly the same value on the primary objective, or the value it reaches is its threshold. In the latter case, the next objective starts to dominate the decision (see Fig. 2).

B. The Options Framework

The Options framework [7] makes use of the same linear scalarization as the baseline. However, in a finite set of states the agent is now also able to select options as well as actions. Options, similar to actions, are choices an agent can take with learnable values associated with them. However, unlike actions, options span several time steps and contain separate policies within themselves. An option is defined as a tuple $\langle I, \pi, \beta \rangle$ where $I \in S$ represents a set of initiation states, π represents the option policy that is used by the agent to select actions once the option is initiated and $\beta \in S$ contains all the states in which the option terminates. The value of an option o in state s under a given policy μ is defined as follows:

$$Q^{\mu}(s,o) = E\{r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + ... | E(o_{\mu}, s, t)\} \tag{3}$$

Where $E(o_{\mu}, s, t)$ denotes the initiation of the policy corresponding to option o in state s at time t followed by the initiation of policy μ when the option has terminated.

Assuming that transition probabilities are not considered, the standard Q-learning method for updating the values of options is as follows:

$$Q(s,o) \leftarrow Q(s,o) + \alpha[r + \gamma^k \max_{o' \in O_{s'}} (Q(s',o')) - Q(s,o)] \tag{4}$$

C. The MORE Algorithm

The Multi-Objective Reward Exponentials (MORE) framework [8] makes use of an exponential function for determining the overall value of a given state with expected sums of future rewards for each objective. Internally, this leads to dynamic weighting of objectives (which can be derived analytically, see [8] for details) that ensures that the momentarily least satisfied objectives are given priority. Given K objectives, the MORE value function on a roll out is defined as follows:

$$V^{\pi}_{\text{MORE}} = -\sum_{k=0}^{K-1} \exp(-V_k^{\text{LIN}}) \tag{5}$$

The rationale for this non-linear scalarization is that it can only be satisfied if all objectives are achieved in a balanced manner,

compared to linear scalarizations in which achievement of different objectives is interchangeable (compare Fig. 2). Priority-objective reinforcement learning can be implemented in this framework by subtracting a constant c_k inside the exponential term to act as a condition for the k-th objective. The modified function is presented as follows:

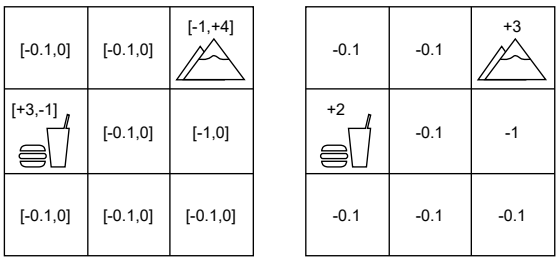
$$V^{\pi}_{\mathrm{MORE}} = -\sum_{k=0}^{K-1} \exp(-(V^{\mathrm{LIN}}_{k} - c_k)) \tag{6}$$

Positive values of c_k introduce a synthetic deficit in the value function, that gives the objective a higher priority since the higher values V^{LIN}_{k} need to be reached to achieve a balanced state after the deduction of c_k (see Fig. 2). c_k is chosen by the designer to express their intention towards the agent's behavior. The values are intuitively related to cumulative reward. For example, $c_0 = 5$ means that the agent needs to accumulate and maintain a reward of $V^{\mathrm{LIN}}_{0} = 5$ on the first objective before other objectives become equally important. The first objective therefore has priority. Once the deficit is equalized, MORE encourages the agent to maintain a balance between objectives, meaning that it will avoid seeking large values for one objective at the expense of the other.

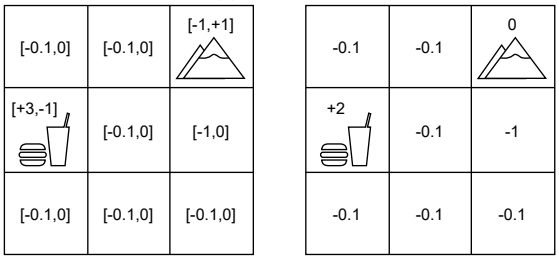
D. Environment

For each framework, 4 distinct cases need to be considered in a two-objective scenario. The first of these cases involves the abundance of rewards for both objectives (see Fig. 3a). This is perhaps the most common scenario an agent will be faced with and the expectation is intuitive. The second case involves the rewards for the secondary objective being constrained whilst the rewards for the first objective remain abundant. The expectation here is that the agent will primarily focus on the primary objective but at the same time the agent will make an effort to satisfy the secondary objective as much as possible. The third case involves the rewards for the first objective being scarce whilst the rewards for the second are left unconstrained. In such a scenario, it is expected that, despite being unable to fully satisfy the primary objective, the agent should be incentivised to obtain rewards for the secondary objective every now and again in an opportunistic manner. However the agent should still ensure that the primary objective is optimised as much as possible. The final scenario involves the rewards for both objectives being constrained. The expectation here is that the agent will focus on the primary objective.

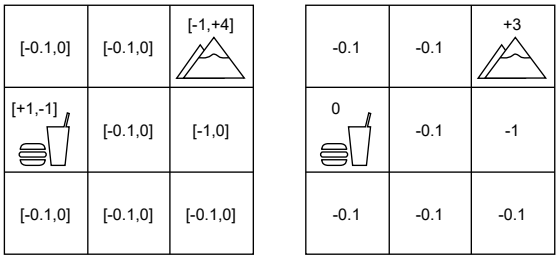
We use the example of a mountain climber as an analogy for these 4 scenarios. Consider a situation in which a mountain climber is faced with the choice of continuing his ascent or stopping at one of many nearby cafes situated on the mountain. A 3 by 3 gridworld multi-objective Markov decision process (MOMDP) can be used as an abstraction for this type of iterative environment where the states are numbered from 0 to 8 starting from the upper left position. The act of stopping at the cafe in state 3 can then be used to represent the enhancement of the primary objective which is to consume an adequate amount



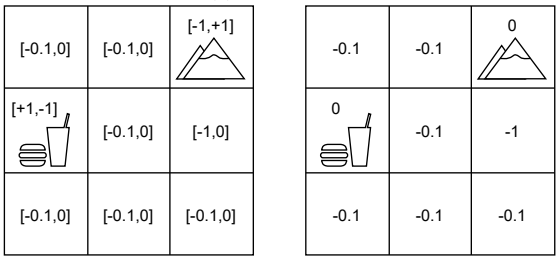
(a) First Case



(b) Second Case



(c) Third Case



(d) Forth Case

Fig. 3: Mountain Climber Example MOMDP Alongside MDP

of food and water in order to survive. The act of entering state 2 i.e. climbing towards the top of the mountain represents the enhancement of the secondary objective which is to reach the summit. Each time the agent enters state 2, it is assumed that the agent has entered the next section of the mountain. For the sake of realism, it is also assumed that when the agent reaches the top of the mountain, it then returns to its starting point and repeats the exercises in order to continue earning rewards for the secondary objective. By assuming that a penalty is incurred whenever the agent enters or stays in a resource-free state, the effect of exhaustion and hunger can effectively be modelled. The MOMDP alongside its equivalent MDP with linear scalarization applied is set up as shown in Fig. 3a.

In order to comprehensively demonstrate that objectives are being prioritized correctly, we carry out a set of experiments

in which a broad range of weights are tested. The algorithms are evaluated in 4 key environments corresponding to the 4 cases described above and both the total rewards and the state dwelling times are recorded.

a) First Case: To begin with, we run all 4 algorithms with the rewards in the environment shown above. The expectation here is that the agent will move between states 2 and 3 in a repetitive manner with a slight preference for staying in state 3. The overall reward accumulated for both objectives should be positive and relatively high. In order to evaluate the options framework effectively, an option is set up with the states 0,1,4,6,7 and 8 being used for initialization and state 3 for termination. These initiation states could represent the states in which the agent begins to feel hunger. The reward values 3 and 4 have been chosen in order to ensure that a fair comparison is made between the options framework and the other algorithms. For the ranking approach and the MORE framework, the MOMDP shown to the left is utilized. The expectation is that the MORE framework will be able to guide the agent so as to optimise both objectives proportionately and ultimately optimize the overall reward. The ranking approach on the other hand, is expected to encourage the agent to head towards and remain in the cafe for the majority of the simulation instead of climbing the mountain at some stage. This is expected to happen primarily because the ranking approach compares the values of each objective in a sequence and as long as the capped value of a primary objective offered by one state, is higher than the capped values of all the other states for that objective, the values of the higher level objectives in the other states are not even considered. In these experiments, we set the threshold for MORE and the ranking approach at a value of 5 to start with, and then again at 10.

b) Second Case: Next we consider the case where the rewards for the secondary objective are scarce but the rewards for the primary objective are abundant. A diagram to depict this environment is shown in Fig. 3b.

The expectation here is that the agent should primarily focus on obtaining rewards for the first objective; however, it should not neglect the second objective. In other words, the agent should spend a adequate amount of time in the cafe i.e. state 3 in order to ensure the primary need is satisfied, although the agent should also continue the ascent by entering state 2 on a regular basis to avoid losing out on the secondary objective. This behaviour is expected primarily because the agent can afford to do so due to the rewards of the first objective being abundant.

c) Third Case: Next we consider the opposite scenario. In this case we analyse what happens when the primary objective is difficult to optimise but the rewards for the secondary objective are abundant. A diagram depicting this test environment is shown in Fig. 3c. Here, the expectation is that the agent will still aim to optimise the primary objective to the level that is required by the condition and then start optimising the higher level one. Overall we expect to see that the reward sums are more or less equal despite the abundance of rewards for the second objective.

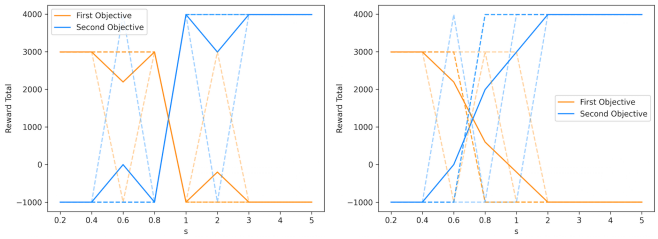
d) Fourth Case: Lastly, we consider the case in which the rewards for both objectives are scarce. The environment we use to test this is shown in Fig. 3d. In this case, the expectation is that an agent will attempt to strictly prioritise the primary objective over and above the higher level one. In addition, given the harsh penalties incurred for each objective in states 2 and 3, the agent is expected to spend a great deal of time in the other states excluding state 5.

III. Results

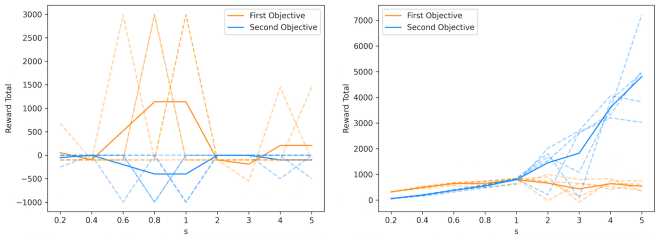
Each algorithm was run over five independent trials of 10000 training steps and 1000 evaluation steps in each environment. The algorithms were also tested with a series of static linear reward weightings in order to investigate the robustness of any observed effects with respect to subtle changes in reward shaping. In order to systematically alter the reward ratio of both needs, the second objective is scaled by a factor s, ranging from 0.2 to 5, and results reported for independent runs over the entire range of s. Any results found consistently across scaling factors can be seen as particularly robust and transferable. MORE is tested with priorities $c_0 = 5$ and $c_0 = 10$, while $c_1 = 0$ is kept constant. This is expected to result in different extents of prioritization of the first, primary objective, which requires a respectively higher cumulative reward for the objectives to achieve balance.

a) First Case: The results for the first case are shown in Fig. 4. We can observe that the linear approach and the options framework are ineffective for all weight combinations tested: they pick one objective entirely while entirely neglecting the other despite their abundance. The ranking approach appears to obtain a positive reward sum for the first objective in most cases but does not achieve a positive result with both objectives. In the case of MORE all weighted objective preferences can deliver positive results. In addition, from the state dwelling times, it appears that in the ranking and MORE frameworks, the agent does move between states implying that the agent is actively choosing which objective to prioritise. MORE gradually capitalizes on the increased attractiveness of the second objective for increasing s, but without ever neglecting the first and primary objective.

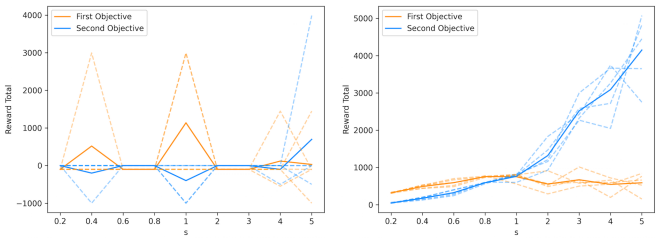
b) Second Case: The performances of each of the algorithms in this environment are shown in Fig. 5. Here we observe once again that the linear approach and the options framework fail to obtain a positive reward for both objectives with all weight combinations. The ranking approach appears to be capable of prioritizing the objectives correctly. However, this method also fails to obtain a positive reward for both objectives. The MORE Framework has met expectations. As we can observe, both objectives have been optimised however the first objective has been distinctively prioritized over the second. From the state time plots it is evident that the MORE framework encourages the agent to spend a more or less equal number of time steps in both state 2 and 3. This demonstrates that the algorithm will not neglect the rewards for the secondary objective when the rewards for the primary objective are abundant.



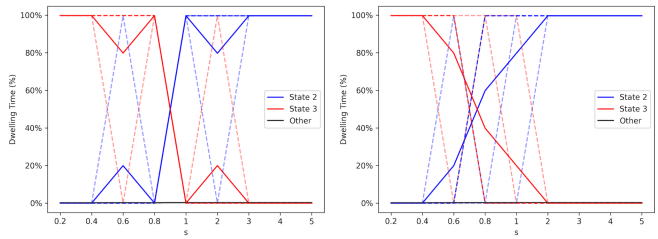
(a) Linear approach and the options framework



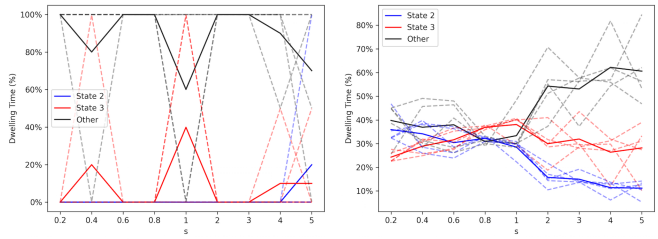
(b) Ranking approach and the MORE framework with $c_0 = 5$



(c) Ranking approach and the MORE framework with $c_0 = 10$

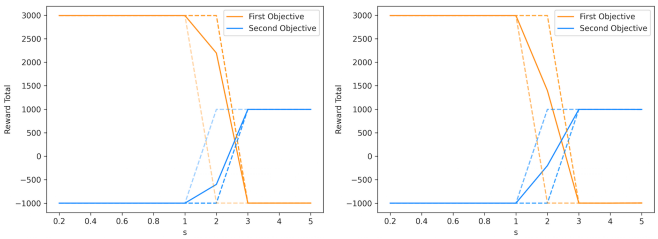


(d) State dwelling times for the linear and options approach

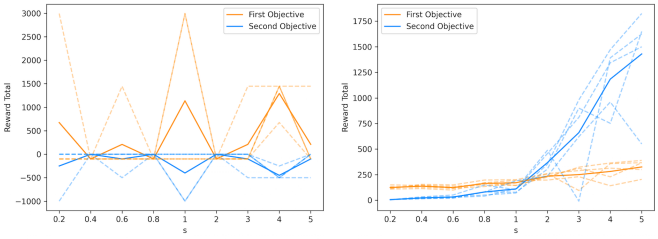


(e) State dwelling times for ranking and MORE with $c_0 = 10$

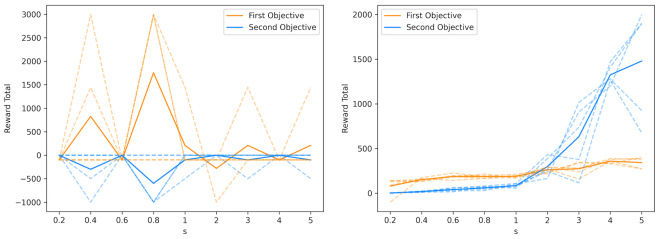
Fig. 4: Case 1 results



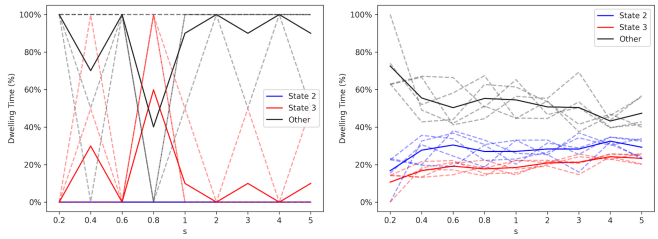
(a) Linear approach and the options framework



(b) Ranking approach and the MORE framework with $c_0 = 5$



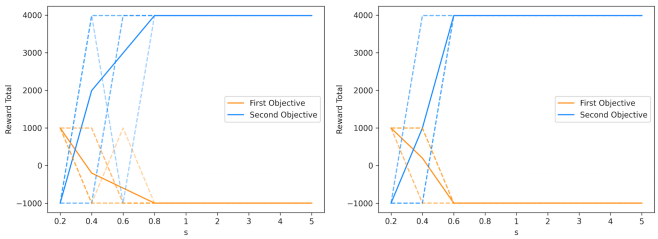
(c) Ranking approach and the MORE framework with $c_0 = 10$



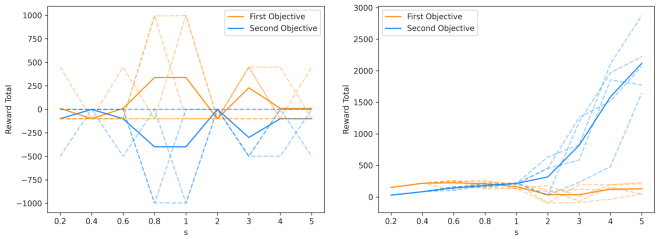
(d) State dwelling times for ranking and MORE with $c_0 = 10$

Fig. 5: Case 2 results

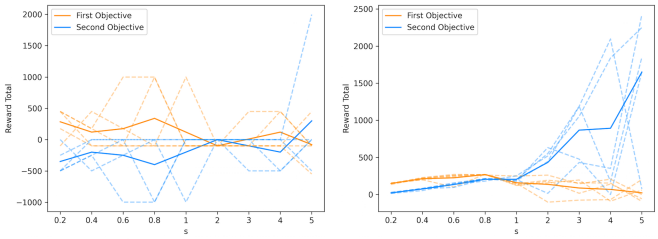
c) Third Case: The results for this case are shown in Fig. 6. Linear and options approach provide negative scores for at least one objective. This behaviour is also present in the ranking approach for all weight combinations. The MORE framework appears to provide the closest result to what is expected and further, up until this point, most of the reward sums for both objectives obtained by the algorithm are positive. From the state dwelling times it is also evident that the linear approach is incapable of dynamically switching states as needs change over time implying that there is no active prioritisation taking place. On the other hand, the MORE framework appears to move between states quite frequently and in this case it appears to spend a larger part of the simulation in state 3 despite the rewards being more abundant in state 2. This behaviour is in contrast with the previous case where the rewards of the primary objective



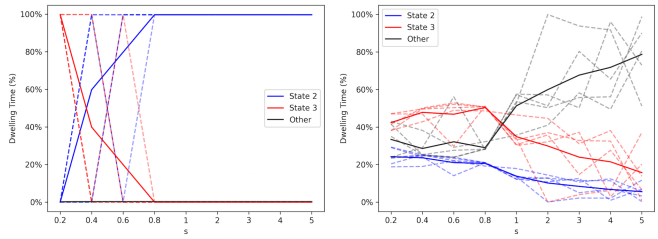
(a) Linear approach and the options framework



(b) Ranking approach and the MORE framework with $c_0 = 5$

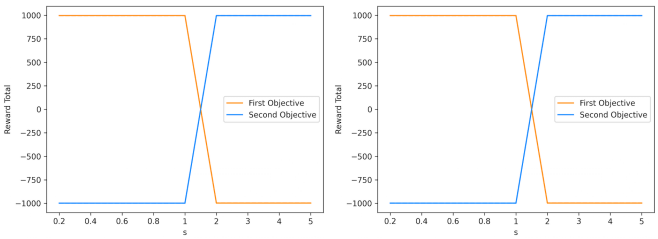


(c) Ranking approach and the MORE framework with $c_0 = 10$

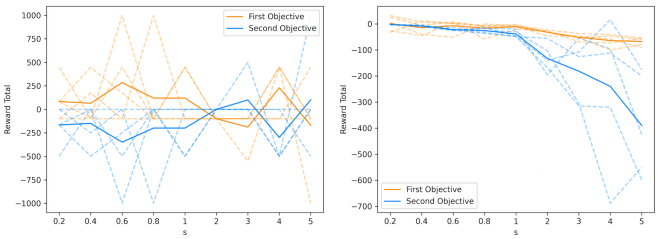


(d) State dwelling times for the linear approach and the MORE framework with $c_0 = 10$

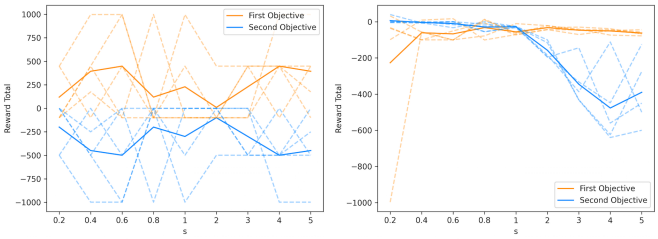
Fig. 6: Case 3 results



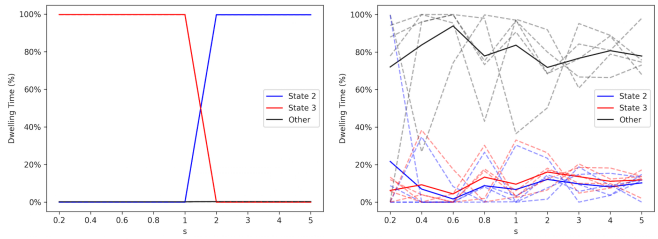
(a) Linear approach and the options framework



(b) Ranking approach and the MORE framework with $c_0 = 5$



(c) Ranking approach and the MORE framework with $c_0 = 10$



(d) State dwelling times for the options approach and the MORE framework with $c_0 = 10$

Fig. 7: Case 4 results

were abundant. This asymmetrical pattern of behaviour clearly demonstrates that the primary objective is being prioritised over the secondary one.

d) Fourth Case: The results for this case are shown in Fig. 7. This environment does not allow an agent to obtain positive rewards on both objectives simultaneously. However, the approach that provides the most promising reward sums overall is MORE. In addition, relative to the priorities, the MORE framework still optimises both objectives in a manner that is expected in this type of environment. From the state dwelling times it appears that in this case where both rewards have been constrained, the ranking framework can somewhat simulate state prioritisation. However, in comparison to the MORE framework, it still fails to correctly prioritise objectives across the board. In addition, unlike in the second case where the rewards for the first objective were abundant, this time the MORE framework appears to neglect the rewards for the second objective because doing so would hamper the rewards of the already constrained primary objective. Furthermore, it appears that the MORE framework has figured out that it is more economical to remain outside of the states 2 and 3 for the majority of the simulation due to the relatively strong penalties incurred in these states on one of the objectives. Rather, the agent spends time on the much less penalized 'other' states and occasionally tops up the primary objective in state 3.

IV. Discussion

From the results shown above, it appears that the algorithms have performed as expected. In the first case, we observe that both the linear and the options framework spend the majority

of time in either state 2 or state 3 which is undesirable. The static weighting s can in each environment be tuned to make the agent focus on the primary objective (by scaling down the rewards for the secondary one with $s<1$), but which will come at the complete expense of the secondary objective. Worse than that imbalance is that the agent will rather suddenly switch from optimizing only one objective to the other at some threshold value of s that is different for each of the four tested environments. Hence, from a design and reward shaping perspective there is no choice that generalizes well across environments.

The MORE framework appears to be striking a more appropriate balance between the rewards offered by these states i.e. the enhancements of the survival and the climbing objectives. In particular, it displays sensible behavior across all static reward scaling factors s, and changes its decisions in predictable and gradual ways with changing s, unlike the abrupt and hard to predict changes observed for the linear approaches. We have demonstrated that this holds for different choices of c_0 at design time, with the larger value 10 achieving a stricter prioritization than 5.

The ranking approach, on the other hand, appears to be encouraging the agent to remain in state 3, which is not reflective of how an agent should perceive higher level rewards. We analyse these results from the perspective of solving problems in which a balanced resolution of objectives is needed. For example, a living agent cannot survive with just oxygen, it also requires the active consumption of food and water. However, oxygen still remains the number one priority. The design assumption here is that optimising an objective in any case is better than optimizing none. What is evident from the experiments carried out in this study is that, according to the reward metrics of each framework, MORE appears to produce the desired behaviour whilst, the ranking, options and linear approach fail to do so. In the second case, the options framework and the linear approach appear to produce inadequate results whilst the ranking and MORE frameworks appear to demonstrate promising ones. However, the MORE framework appears to deliver more desirable results across the preference space. In the third case, virtually all frameworks except MORE fail to meet expectations and in the final case, it seems that apart from MORE, only the ranking approach has managed to induce the correct behaviour.

V. Conclusion

In this study we created an environment to reflect four key objective prioritization scenarios, implemented four PORL frameworks, and tested these frameworks in the environment. Considering the results we conclude that the modified MORE framework appears to have delivered the best performance overall. There are, of course, limitations to this study in that the reduced number of frameworks implemented, the reduced complexity of the environments and the short time frames in which these simulations were run are all factors that would need further analysis. However, it appears that the complexity of these environments employed in the study was sufficient to distinguish the performances of the frameworks that have been tested. The findings of this study may prove useful in attempting to unify various theories in cognitive robotics such as the cognitive explanatory gap (CEG) [18] and the cognitive development model proposed by Lones et al., [4].

References

[1] A. H. Maslow, "A theory of human motivation." *Psychological Review*, vol. 50, no. 4, pp. 370–396, 1943. [Online]. Available: https://doi.org/10.1037/h0054346

[2] S. Maes and W. Gebhardt, "Self-regulation and health behavior: The health behavior goal model," in *Handbook of self-regulation*, M. Boekaerts, P. R. Pintrich, and M. Zeidner, Eds. Academic Press, 2000, pp. 343–368.

[3] S. Lo Piano, "Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward," *Humanities and Social Sciences Communications*, vol. 7, no. 1, p. 9, Jun 2020. [Online]. Available: https://doi.org/10.1057/s41599-020-0501-9

[4] J. Lones, M. Lewis, and L. Canamero, "From sensorimotor experiences to cognitive development: Investigating the influence of experiential diversity on the development of an epigenetic robot," *Frontiers in Robotics and AI*, vol. 3, p. 44, 2016. [Online]. Available: https://www.frontiersin.org/article/10.3389/frobt.2016.00044

[5] S. Li, R. Wang, M. Tang, and C. Zhang, "Hierarchical reinforcement learning with advantage-based auxiliary rewards," 2019.

[6] P. Vamplew, R. Dazeley, A. Berry, R. Issabekov, and E. Dekker, "Empirical evaluation methods for multiobjective reinforcement learning algorithms," *Machine Learning*, vol. 84, no. 1, pp. 51–80, 2011. [Online]. Available: https://doi.org/10.1007/s10994-010-5232-5

[7] R. S. Sutton, D. Precup, and S. Singh, "Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning," *Artificial intelligence*, vol. 112, no. 1-2, pp. 181–211, 1999.

[8] M. Rolf, "The need for more: Need systems as non-linear multi-objective reinforcement learning," in *2020 Joint IEEE 10th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 2020, pp. 1–8.

[9] A. Aubret, L. Matignon, and S. Hassas, "A survey on intrinsic motivation in reinforcement learning," *arXiv preprint arXiv:1908.06976*, 2019.

[10] A. S. Vezhnevets, S. Osindero, T. Schaul, N. Heess, M. Jaderberg, D. Silver, and K. Kavukcuoglu, "Feudal networks for hierarchical reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2017, pp. 3540–3549.

[11] P.-L. Bacon, J. Harb, and D. Precup, "The option-critic architecture," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[12] E. Wiewiora, "Reward shaping," in *Encyclopedia of Machine Learning*, C. Sammut and G. I. Webb, Eds. Boston, MA: Springer US, 2010, pp. 863–865. [Online]. Available: https://doi.org/10.1007/978-0-387-30164-8_731

[13] K. Van Moffaert, M. M. Drugan, and A. Nowé, "Scalarized multi-objective reinforcement learning: Novel design techniques," in *IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, 2013, pp. 191–199.

[14] Z. Gabor, Z. Kalmar, and C. Szepesvari, "Multi-criteria reinforcement learning," in *International Conference on Machine Learning (ICML-98)*, Madison, WI, 1998. [Online]. Available: http://citeseer.ist.psu.edu/gabor98multicriteria.html

[15] L. Barrett and S. Narayanan, "Learning all optimal policies with multiple criteria," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 41–47.

[16] A. Y. Ng, D. Harada, and S. Russell, "Policy invariance under reward transformations: Theory and application to reward shaping," in *Icml*, vol. 99, 1999, pp. 278–287.

[17] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in ai safety," *arXiv preprint arXiv:1606.06565*, 2016.

[18] J. A. Reggia, G. E. Katz, and G. P. Davis, "Humanoid cognitive robots that learn by imitating: Implications for consciousness studies," *Frontiers in Robotics and AI*, vol. 5, p. 1, 2018. [Online]. Available: https://www.frontiersin.org/article/10.3389/frobt.2018.00001