Getting Priorities Right: Intrinsic Motivation with Multi-Objective Reinforcement Learning

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

***Abstract*—Intrinsic motivation is a common method to facilitate exploration in reinforcement learning agents. Curiosity is thereby supposed to aid the learning of a primary goal. However, indulging in curiosity may also stand in conflict with more urgent or essential objectives such as self-sustenance. This paper addresses the problem of balancing curiosity, and correctly prioritising other needs in a reinforcement learning context. We demonstrate the use of the multi-objective reinforcement learning framework C-MORE to integrate curiosity, and compare results to a standard linear reinforcement learning integration. Results clearly demonstrate that curiosity can be modelled with the priority-objective reinforcement learning paradigm. In particular, C-MORE is found to explore robustly while maintaining self-sustenance objectives, whereas the linear approach is found to over-explore and take unnecessary risks. The findings demonstrate a significant weakness of the common linear integration method for intrinsic motivation, and the need to acknowledge the potential conflicts between curiosity and other objectives in a multi-objective framework.**

I. Introduction

Artificial curiosity and intrinsic motivation (IM) are commonly employed methods in today's robotics and AI research [1], [2], [3], [4]. Intrinsic motivation facilitates an agent's exploration that is not immediately directed at external rewards, but facilitates the long-term success of the agent by potentially exploring new, beneficial behavioral repertoires that would not be discovered by a greedy exploration method. Successful examples include exploration in sensorimotor spaces for reaching [5] and touch [6], speech [7] and challenging video game environments such as Montezuma's Revenge [8]. Various classification methods have been proposed to capture the enormous variety of different IM methods in use. These focus mostly on the principles in which artificial curiosity is calculated in the first place. For instance, some authors have emphasized the distinction between knowledge-based and competence-based approaches [2]. A common method of knowledge-based IM is to focus on the distribution of states an agent has seen, and actively seek to explore a wider range. Others have emphasized the need for a wider view on artificial creativity [1].

Even though some authors have argued for agents to follow solely intrinsic motivation (e.g. novelty search [9]), the vast majority of studies use IM in conjunction with some (or several) primary objectives. IM in this context is supposed to aid the success of the agent's primary objective. The range of methods to integrate IM into an agent that has other goals is as wide as the range of agent architectures itself. In developmental literature it had classically been argued that distinct phases of exploration might precede more goal-directed activity [10], [11]. However, it is most commonly seen as more efficient, as well as developmentally plausible [12], [13], to closely intertwine exploratory and goal directed activity. This is often seen in the implementation of curiosity in reinforcement learning. A common way to integrate curiosity into a reinforcement learning agent is to express curiosity as a reward and add it to the existing reward that corresponds to the external task [14], [15]. This way, the agent is constantly incentivised to explore alongside its external reward seeking behavior. A drawback of this method is that the agent ends up learning an inseparable mixture of, or compromise between, intrinsic and extrinsic incentives. Calibrating the right compromise between different scalar reward components is a known issue in reinforcement learning, and is usually susceptible to reward hacking behaviors and other safety issues [16]. In some application contexts the agent also knows concrete target states besides the reward alone. In such cases it is common to define a more flexible integration of IM and the external task by giving rewards when the target is either reached, or approached better than before [17], [18]. This avoids the compromise making, but does not generalize into RL setups without any explicitly known target states.

The main challenge in the integration of curiosity into an agent that behaves in the real world is that satisfying curiosity may stand in direct conflict with the agent's primary objective (or objectives). This may be an externally given task which needs to be solved, and should not be neglected at the expense of indulging in curiosity too extensively. Other needs may be much more urgent, and generally of a higher priority, such as the requirements for self-sustenance and safety. 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 goals first before attempting to complete any secondary objectives such as curiosity. However, once a survival need is met, the agent should then go out and explore its environment further in an effort to discover more resources that can better satisfy higher level needs. For example, when an agent has satisfied its energy requirement, it should then be intrinsically motivated to explore the environment in the hope that it may stumble upon greater energy resources. This type of exploratory behaviour is highly desirable because it allows

an agent to make optimal use of the resources present in a given environment. This process is analogous to the intrinsic motivation [3] in biological creatures and enables an agent to be more opportunistic about obtaining rewards altogether. However, this type of behaviour may not be desirable if exploring the environment comes at the cost of the agent's survival, for example when a drone runs out of battery in the pursuit of finding new charging stations. Therefore it is essential that a developing agent prioritises its own self-preservation over and above exploration.

The balancing of needs and orchestration of various behaviors has classically been the role of cognitive architectures. For example, the subsumption architecture [19] allows urgent behaviors to entirely block higher level or less urgent functions. The downside of this design is that synergies between tasks and objectives (if present) cannot be exploited. Models of hormonal systems [20] take a more gradual approach and allow specific needs to be modulated by more urgent ones. However, these systems rely on hand-crafted hormonal dynamics equations that are not based on any common decision theory and therefore very hard to generalize.

Methods in multi-objective reinforcement learning (MORL), on the other hand, are an explicit attempt to establish a consistent decision theory for an agent satisfying several objectives [21], [22], [23], [24]. MORL approaches can broadly be divided into two main types, namely single-policy and multi-policy types [25], [26]. Single-policy approaches attempt to find a unique optimal policy whilst multi-policy approaches attempt to find a group of optimal policies from which a user can select. Within single policy approaches, the standard linear approach computes the weighted sum of the rewards of a given state for each objective. Linear methods have the same disadvantages in terms of compromise and safety as standard reinforcement models, in that the rewards of different objectives can be traded off for each other. While specific safety rewards or even boundaries have been built into the design of such agents [27], general decision rules are hard to articulate in the linear framework. Non-linear methods that prevent trade-offs have however been proposed: The "Multi-Objective Reward Exponentials" (MORE) framework [28] uses a soft-min-like utility function to combine objectives, which results in a balanced achievement of all objectives. This method has also been shown to allow for targeted prioritisation [29] of objectives by the designer where appropriate, which results in behavior that satisfies urgent needs first, but giving in to secondary needs when possible.

Multi-objective reinforcement learning seems uniquely well suited to integrate intrinsic motives into an agent that learns while behaving. It acknowledges the fact that there are distinct objectives, rather than compounding them into one objective in which contributions are interchangeable. Yet, very little work to date has been conducted in this direction. One study [30] presented preliminary results on the integration of IM with linear multi-objective RL with variable weighting, inheriting all the balancing problems from ordinary single-objective reinforcement learning (see also [28]). In another study [31], a more complex, multi-policy learning method for the integration of intrinsic motivation was investigated. This approach delays the actual selection of a solution and allows an external user or supervisor of the system to select a preference after learning. It is therefore not applicable to systems that exhibit lifelong learning.

A. Contribution and Outline

This paper investigates how intrinsic motivation can be integrated into an agent with the non-linear multi-objective reinforcement learning framework C-MORE [29]. We capitalize on C-MORE's ability to dynamically balance the achievement of different needs, while allowing for a gradual prioritisation at the same time. This ability is demonstrated in two environments in which the agent needs to fulfill one or two immediate self-sustenance objectives before exploring further. Experiments show that C-MORE with IM achieves highly robust exploration results, all while maintaining a continuous balance with self-sustenance. In contrast, linear integration of IM in the classical fashion [14], [15] is shown to involve high risk behaviour and a tendency towards abandoning the critical self-sustenance goals. Our study therefore demonstrates a critical weakness of the most common integration method of curiosity into reinforcement learning. This in particular affects RL agents with lifelong learning that learn while behaving as opposed to learning through supervision. The proposed solution tackles the balancing and prioritisation of the objectives directly within the agent's decision making mechanism, rather than relying on operator judgement as in the only two previous MORL studies on IM [30], [31]. Conceptually, this can be seen as a generalization of previous models of hormonal dynamics with homeostatic behavior [20], which relies on handcrafted non-linear hormonal dynamics, into a formal decision theory with a globally consistent utility function.

II. Methodology

We analyse the performances of two different MORL frameworks that have curiosity built into them. The first of which is the standard linear method and the second is the C-MORE framework. We test these frameworks in two separate environments in the case of both 2 and 3 objectives. The environments employed are based on the Multi-Objective Markov Decision Process (MOMDP) model. A MOMDP can be represented by a tuple $\langle S, A, P, \mathbf{r} \rangle$ where S represents the state space, A represents the set of actions an agent can take, and P represents the probability of entering a state s' from a state s when action a is taken. Different from ordinary Markov Decision Processes, the reward $\mathbf{r}$ is a vector, containing the rewards corresponding to K objectives.

A. Standard Linear Method

In the standard linear scalarisation approach, the overall value of a state or state-action pair is the expected sum of future rewards over all K objectives i.e. the total value for a given state is determined on the basis of a weighted sum of calculated values V_k for each objective where the weighting

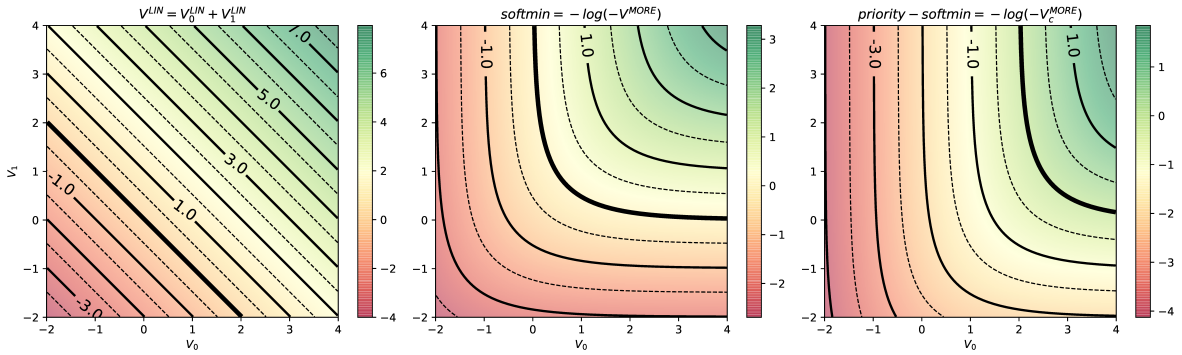


Fig. 1: Utility concepts for the multi-objective problem: (i) standard linear scalarization, (ii) the non-linear MORE scalarization that corresponds to a softmin function [28], (iii) C-MORE, a MORE scalarization that is shifted by $c_0 = 2$ on the first objective, giving it higher priority [29], but also continuously weighting in the second objective when the first one becomes satisfied.

$\mathbf{W} = (W_0, W_1, ..., W_{K-1})$ is constant and predetermined by a user:

$$V_\pi^{\text{LIN}} = \sum_{k=0}^{K-1} W_k V_{\pi,k}^{\text{LIN}} = \sum_{k=0}^{K-1} W_k E_\pi \left[\sum_t \gamma^t r_k(t) \right] . \quad (1)$$

In this study, we use a standard Q-learning implementation to optimize the agent's behavior towards this combined value.

B. The C-MORE Algorithm

The Conditional Multi-Objective Reward Exponentials (C-MORE) framework [29] makes use of a weighted exponential function for calculating the overall value of a given state in accordance with a specified priority vector. Internally, this leads to dynamic and prioritised weighting of objectives (which can be derived analytically, see [28] for details) that ensures that priority is given to the momentarily least satisfied objectives. Given K objectives, the MORE value function is defined as follows:

$$V_\pi^{\text{MORE}} = -\sum_{k=0}^{K-1} \exp(-V_{\pi,k}^{\text{LIN}}) \quad (2)$$

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 achievements of different objectives are interchangeable (compare Fig. 1). 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:

$$V_\pi^{\text{C-MORE}} = -\sum_{k=0}^{K-1} \exp(-(V_{\pi,k}^{\text{LIN}} - c_k)) \quad (3)$$

Positive values of c_k introduce a synthetic deficit in the value function, that gives the objective a higher priority since the higher values $V_{\pi,k}^{\text{LIN}}$ need to be reached to achieve a balanced state after the deduction of c_k (see Fig. 1). 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 $V_0^{\text{LIN}} = 5$ 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 and vice versa. The MORE objective is optimized by the Q-learning-based algorithm which is described in detail in the original MORE paper [28]: in each step, actions are chosen that maximize Eqn. 3, and expected future cumulative rewards $V_{\pi,k}^{\text{LIN}}$ per objective are estimated with standard Q-learning.

C. Intrinsic Motivation

In order to model curiosity, we picked one of the most simple intrinsic motivation approaches, *uncertainty motivation* [2], [32]. The agent keeps track of how often n_s it has visited each state in relation to the amount t of steps so far. Higher rewards are given on states that have been visited rarely. The basic equation for the curiosity state novelty reward signal r_c is

$$r_c = C \left(1 - \frac{n_s}{t}\right), \quad (4)$$

with an application-specific coefficient C [2].

We found this expression to be rather limiting in the environments tested, as most rewards will be numerically very close to C, while some large outliers may be closer to zero. We therefore applied a further transformation that provides a more stable numeric range, while also expressing more nuance between different rarely visited states:

$$r_c = C_2 \arctan\left(C_1 \cdot \left(1 - \frac{n_s}{t}\right) + C_0\right) \quad (5)$$

The use of the $\arctan$ function serves to cap the numeric range of the rewards against extreme numeric lows and highs. This was found necessary in order to let any of the tested algorithms explore successfully. We chose $C_2 = 5$ in order to scale the IM rewards to the same general scale as the other rewards. We further chose $C_1 = 6N$ (where N is the number of states) and $C_0 = 15$ for all cases. These values were manually chosen

ahead of the learning experiments with the aim of placing the most common values on the center of the $\arctan$'s non-linearity.

D. Environment

For each framework, two separate multi-objective gridworld environments are used to assess the performances of the agent. We employ two primary objectives which are meant to constitute an agents overall health. The first of these objectives (extrinsic objective 1) is the energy level which is a quantity that decreases whenever the agent moves. This objective is modelled using a motion penalty of -0.1. The second (extrinsic objective 2) is an agent's physical health. This is a value that can increase or decrease by fixed amounts. Both primary objectives can be recharged on designated states, but can also take extra damage on other designated perilous states. The expectation is that the agent will balance and maintain these two objectives as well as explore the environment in an effort to find novelty states and ultimately satisfy its intrinsic motivation.

The first environment is shown in Fig. 2a and consists of a 7 by 7 gridworld scenario in which the agent starts in an isolated corner of the map with only a handful of resources within its reach, and then makes its way out into the surrounding areas in order to obtain further rewards. The design choice of using obstacles was motivated in part by the challenges that come with objectively analysing curious exploration. The reward and penalty states have been chosen in key locations in order to accurately assess the agent's rational behaviour.

The second environment is shown in Fig. 2b and is a much larger 7 by 20 gridworld MOMDP that positions resources of varying levels of benefit at different distances from the isolated corner where the agent initially spawns. The sporadic positioning of various reward and penalty states across the environment allows for an accurate assessment of the exploratory behaviour and intelligence of both agents using both frameworks.

In order to comprehensively demonstrate that the level of exploration is rationed accordingly, we carry out two separate sets of experiments. In one set of experiments, the rewards of the two main objectives are added together. This corresponds to the overall health of the agent which can then be assessed with a curiosity need (an intrinsic objective) alongside it. The purpose of this condition is to simplify the scenario for the linear approach so that it only has to deal with IM and one other objective, rather than immediately tackling IM and two other objectives. In the second set, the rewards of all the objectives are left separate and therefore correspond to energy and physical health alongside curiosity respectively. The algorithms are tested against each other in all 4 cases and the performances are analysed therein.

a) First Case - First Environment with 2 Objectives: To begin with, both frameworks are run in the first environment with a single objective alongside the curiosity motive built into it. The expectation here is that the agent will work its way out of the upper corner of the environment and work its way down in order to maximise its reward. On the majority of occasions the framework should find all 3 attractive states but it should not do this at the expense of taking significant penalties to the main objective.

b) Second Case - First Environment with 3 Objectives: In the next experiment, the two frameworks are run in the same environment but this time with 2 main objectives alongside IM. This setup is meant to represent a more complex and realistic scenario in which an agent has to balance multiple primary competing needs such as battery power and physical health, whilst maintaining an appropriate level of curiosity and exploration. The expectation here is not too dissimilar to that of the first case. However, a greater level of focus on seeking primary rewards is expected here.

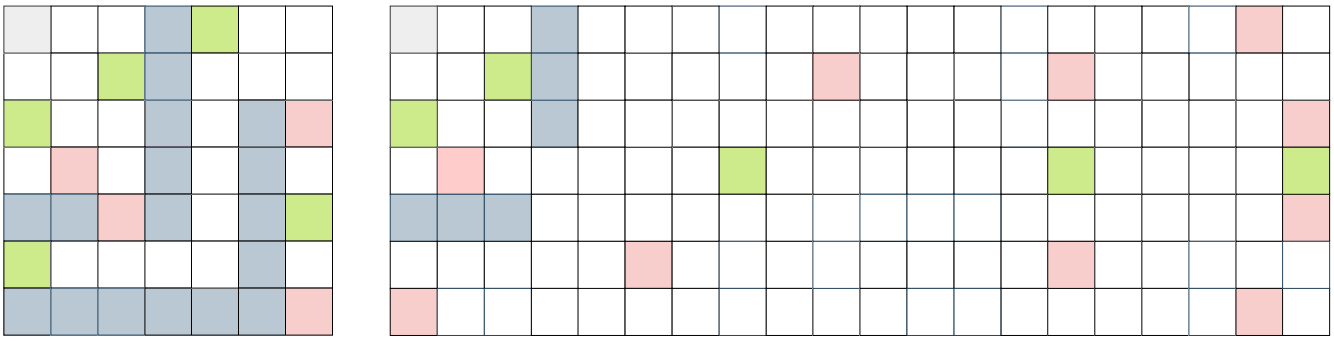
c) Third Case - Second Environment with 2 Objectives: In the following experiment, the frameworks are tested in the second environment with just two objectives. The expectation here is that the agent will make its way out of the enclosed area in the upper left corner and start exploring the large open area from left to right in a progressive fashion. In doing so the agent is expected to find all 3 attractive states on the majority of occasions. However, the continual exploration of an environment in which such a large number of penalty states is present is not always desirable. Therefore the agent is expected to withdraw from the area to the far right of the map when the third attractive state has been discovered.

d) Fourth Case - Second Environment with 3 Objectives: In the final experiment, the two algorithms are tested in the second environment with 3 objectives. This setup is meant to represent a realistic open field environment in which the agent needs to explore an area whilst making sure it protects itself from physical damage as well as fuel shortages. Here again, the expectation is that the agent will work its way out from the corner and courageously explore the area from left to right but in such a way that it avoids danger and stays intact whilst doing so.

III. Results

We ran each framework for 100000 time steps for 10 independent trials. For the first environment we used a reset time of 100 and for the second 250. We measured the number of rounds in which each of the attractive states were discovered alongside the rewards attained for each objective. We also set the priority constants in the C-MORE framework to $C = (5, 5, 0)$, corresponding to high-priority for the first two objectives, and lower priority for the IM objective. In the case of a single primary objective, we used the priority vector $C = (5, 0)$. The central evaluation metric is the total reward for each objective. The total reward for the IM component was capped at 10000 in all visualizations to allow for an easier comparison with other totals.

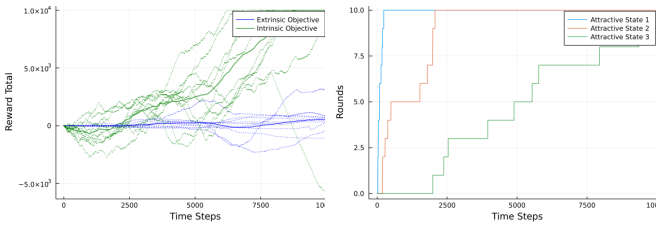
a) First Environment - Cases 1 and 2: The results for the first environment are shown in Fig. 3. The figure on the left shows the reward totals over time for each objective and all ten independent runs. The plots on the right show how many of the ten runs have found each of the three attractive target states (outside the initial 3x3 room) over time. We can



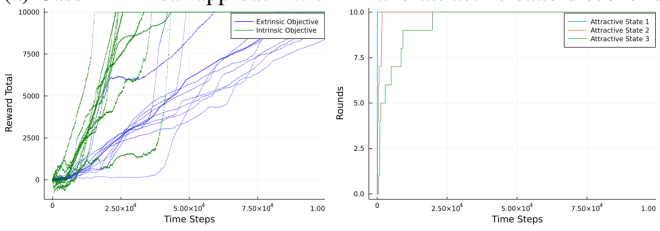
(a) First Environment

(b) Second Environment

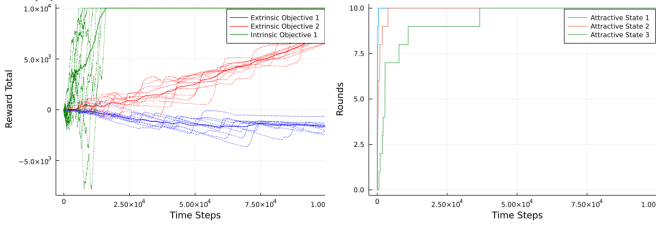
Fig. 2: The two test environments are shown with their two environmental rewards components for each state. The initial state in each environment is in the top left corner. States with positive rewards are shown in green, and negative rewards in red. States that denote obsticles are shaded in grey. The magnitude of rewards is larger for states more distant from the initial state.



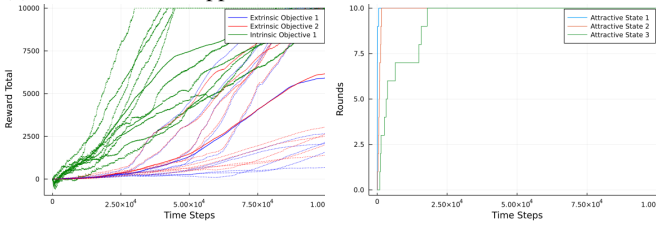
(a) Case 1 - Linear approach with IM and attractive state discoveries



(b) Case 1 - C-MORE with IM and attractive state discoveries



(c) Case 2 - Linear approach with IM and attractive state discoveries



(d) Case 2 - C-MORE with IM and attractive state discoveries

Fig. 3: Case 1 and 2 results

	Case 1		Case 2	
	Linear	C-MORE	Linear	C-MORE
EO-1	43.77 %	1.07 %	96.54%	1.74%
EO-2			6.82 %	1.38%
IM	22.96 %	3.38 %	3.24 %	2.00%

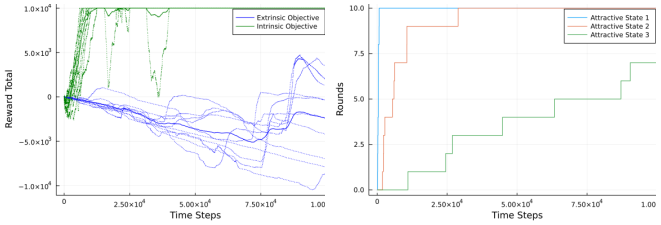
TABLE I: Percentage of time-steps during which each objective (extrinsic EO-1 and EO-2 and intrinsic IM) had a significantly negative reward total in the first environment.

	Case 3		Case 4	
	Linear	C-MORE	Linear	C-MORE
EO-1	91.37%	2.15 %	96.75 %	9.89 %
EO-2			57.52 %	2.17 %
IM	2.40 %	2.98 %	2.54 %	2.73 %

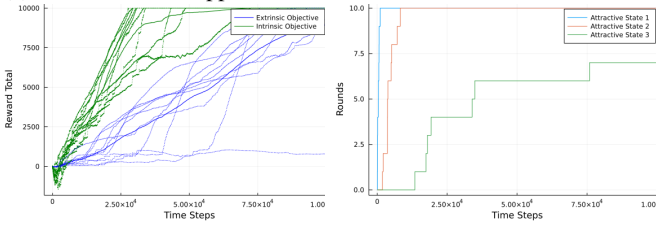
TABLE II: Percentage of time-steps during which each objective (extrinsic EO-1 and EO-2 and intrinsic IM) had a significantly negative reward total in the second environment.

observe that the linear approach with IM in both cases does not appear to meet expectations. The linear approach is able to locate the attractive states consistently, but it does not reliably optimise all the objectives. The linear framework in this case appears to optimize predominately the curiosity objective at the expense of the self-sustenance rewards. The C-MORE framework displays a faster and more consistent discovery of all attractive target states, but also reliably maintains a positive reward for all objectives. Table I shows in what percentage of time steps the reward totals were negative for each approach and objective. The linear method displays extensive periods of negative achievement, while C-MORE only shows occasional dips of the total which are largely caused by initial exploration.

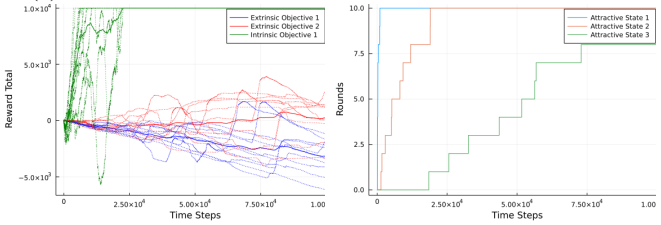
b) Second Environment - Cases 3 and 4: The performances of each of the algorithms in this environment are shown in Fig. 4. Here again, we can observe that, the linear approach with IM in both cases does not appear to meet expectations. While the curiosity rewards accumulate very quickly, the primary objectives are negative in many instances. The linear approach finds the first two attractive states reliably, but struggles to discover the third one.



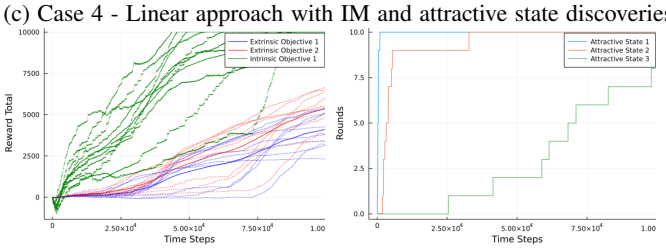
(a) Case 3 - Linear approach with IM and attractive state discoveries



(b) Case 3 - C-MORE with IM and attractive state discoveries



(c) Case 4 - Linear approach with IM and attractive state discoveries



(d) Case 4 - C-MORE with IM and attractive state discoveries

Fig. 4: Case 3 and 4 results

The C-MORE framework scores very similarly in terms of exploration and attractive state discovery. However, it consistently manages to maintain positive reward totals for both primary objectives at the same time. Therefore, once again, the C-MORE framework has met expectations.

IV. Discussion

From the results shown above, it appears that when IM is incorporated into linear reinforcement learning, the level of active exploration of a developing agent increases and the frequency of resource discoveries appears to climb. However, what has become apparent from this research is the risk and potential unreliability of the standard linear approach with IM. In all four cases, we observe how the linear RL agent goes primarily for exploration over and above self-preservation as the rewards for the primary objectives dip and stay below zero on a significant proportion of time steps. On the other hand, incorporating IM into the C-MORE framework appears to produce stable results in simple cases and can produce reliable performances in more realistic scenarios involving multiple primary objectives. In all 4 cases, the C-MORE framework has managed to locate all the attractive states in a significant proportion of runs whilst maintaining positive rewards for both primary objectives almost throughout the entire simulation. Therefore, when IM is incorporated within a need balancing framework such as C-MORE it can produce reliable results in a range of environments. These results have so far been confirmed in the four demonstrated cases, and for one particular (but very common) choice of IM implementation. In addition, only a limited range of priority constants were tested. While further studies should continue this investigation with different IM implementations and environments, the results found here were robust and so far seem to generalize.

V. Conclusion

In this study we incorporated a state novelty based intrinsic motivation system into two multi-objective reinforcement learning frameworks. We then tested these frameworks in two different environments involving both a single primary objective and two primary objectives alongside a curiosity reward. From the results we conclude that the C-MORE framework enhanced with IM appears to deliver better performance overall than the linear approach. The relatively limited complexity of the tested environments employed in the study was sufficient to distinguish the performances of linear RL and C-MORE. The findings of this study confirm clearly that when intrinsic motivation is combined with a reward balancing framework such as C-MORE it can produce desirable curiosity behaviour in an agent that is tasked with exploring an unfamiliar environment. C-MORE's ability to balance, and also to dynamically prioritise objectives is crucial to this positive experimental outcome. The findings therefore underpin the need to treat intrinsic motivation as a truly separate objective through a multi-objective lens, rather than compounding curiosity and other needs into a single objective.

References

[1] J. Schmidhuber, "Formal theory of creativity, fun, and intrinsic motivation (1990–2010)," *IEEE transactions on autonomous mental development*, vol. 2, no. 3, pp. 230–247, 2010.

[2] P.-Y. Oudeyer and F. Kaplan, "What is intrinsic motivation? a typology of computational approaches," *Frontiers in neurorobotics*, vol. 1, p. 6, 2009.

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

[4] K. E. Merrick and M. L. Maher, *Motivated reinforcement learning: curious characters for multiuser games*. Springer Science & Business Media, 2009.

[5] A. Baranes and P.-Y. Oudeyer, "R-IAC: Robust intrinsically motivated exploration and active learning," *IEEE Transactions on Autonomous Mental Development*, vol. 1, no. 3, pp. 155–169, 2009.

[6] F. Gama, M. Shcherban, M. Rolf, and M. Hoffmann, "Goal-directed tactile exploration for body model learning through self-touch on a humanoid robot," *IEEE Transactions on Cognitive and Developmental Systems*, 2021.

[7] C. Moulin-Frier, S. M. Nguyen, and P.-Y. Oudeyer, "Self-organization of early vocal development in infants and machines: the role of intrinsic motivation," *Frontiers in psychology*, vol. 4, p. 1006, 2014.

[8] M. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos, "Unifying count-based exploration and intrinsic motivation," *Advances in neural information processing systems*, vol. 29, 2016.

[9] J. Lehman and K. O. Stanley, "Abandoning objectives: Evolution through the search for novelty alone," *Evolutionary computation*, vol. 19, no. 2, pp. 189–223, 2011.

[10] J. Piaget, *The Origin of Intelligence in the Child*. Routledge and Kegan Paul, 1953.

[11] D. Bullock, S. Grossberg, and F. H. Guenther, "A self-organizing neural model of motor equivalent reaching and tool use by a multijoint arm," *Cognitive Neuroscience*, vol. 5, no. 4, pp. 408–435, 1993.

[12] B. I. Bertenthal, "Origins and early development of perception, action, and representation," *Annual Reviews Psychology*, vol. 47, pp. 431–459, 1996.

[13] M. Rolf, J. J. Steil, and M. Gienger, "Goal babbling permits direct learning of inverse kinematics," *IEEE Trans. Autonomous Mental Development*, vol. 2, no. 3, 2010.

[14] J. Schmidhuber, "Adaptive confidence and adaptive curiosity," in *Institut fur Informatik, Technische Universitat Munchen, Arcisstr. 21, 800 Munchen 2*. Citeseer, 1991.

[15] N. Chentanez, A. Barto, and S. Singh, "Intrinsically motivated reinforcement learning," *Advances in neural information processing systems*, vol. 17, 2004.

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

[17] C. Colas, P. Fournier, M. Chetouani, O. Sigaud, and P.-Y. Oudeyer, "Curious: intrinsically motivated modular multi-goal reinforcement learning," in *International conference on machine learning*. PMLR, 2019, pp. 1331–1340.

[18] S. Hart and R. Grupen, "Intrinsically motivated affordance discovery and modeling," in *Intrinsically Motivated Learning in Natural and Artificial Systems*. Springer, 2013, pp. 279–300.

[19] R. Brooks, "A robust layered control system for a mobile robot," *IEEE journal on robotics and automation*, vol. 2, no. 1, pp. 14–23, 1986.

[20] 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

[21] 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

[22] D. M. Roijers, P. Vamplew, S. Whiteson, and R. Dazeley, "A survey of multi-objective sequential decision-making," *Journal of Artificial Intelligence Research*, vol. 48, pp. 67–113, 2013.

[23] 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.

[24] C. F. Hayes, T. Verstraeten, D. M. Roijers, E. Howley, and P. Mannion, "Expected scalarised returns dominance: a new solution concept for multi-objective decision making," *Neural Computing and Applications*, pp. 1–21, 2022.

[25] Z. Gabor, Z. Kalmar, and C. Szepesvari, "Multi-criteria reinforcement learning," in *International Conference on Machine Learning (ICML-98)*, Madison, WI, 1998.

[26] 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.

[27] J. Garcıa and F. Fernández, "A comprehensive survey on safe reinforcement learning," *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 1437–1480, 2015.

[28] 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.

[29] Y. Al-Husaini and M. Rolf, "Priority-objective reinforcement learning," in *2021 IEEE International Conference on Development and Learning (ICDL)*, 2021, pp. 1–8.

[30] P. Morere and F. Ramos, "Intrinsic Exploration as Multi-Objective RL," *arXiv preprint arXiv:2004.02380*, 2020.

[31] S. Abdelfattah, K. Kasmarik, and J. Hu, "Evolving robust policy coverage sets in multi-objective markov decision processes through intrinsically motivated self-play," *Frontiers in Neurorobotics*, p. 65, 2018.

[32] X. Huang and J. Weng, "Motivational system for human-robot interaction," in *International Workshop on Computer Vision in Human-Computer Interaction*. Springer, 2004, pp. 17–27.