A Brief Guide to Multi-Objective Reinforcement Learning and Planning

JAAMAS track

Conor F. Hayes*
University of Galway
c.hayes13@universityofgalway.ie

Roxana Rădulescu*
Vrije Universiteit Brussel
roxana.radulescu@vub.be

Eugenio Bargiacchi
Vrije Universiteit Brussel

Johan Källström
Linköping University

Matthew Macfarlane
University of Amsterdam

Mathieu Reymond
Vrije Universiteit Brussel

Timothy Verstraeten
Vrije Universiteit Brussel

Luisa M. Zintgraf
University of Oxford

Richard Dazeley
Deakin University

Fredrik Heintz
Linköping University

Enda Howley
University of Galway

Athirai A. Irissappane**
Amazon

Patrick Mannion
University of Galway

Ann Nowé
Vrije Universiteit Brussel

Gabriel Ramos
University of Vale do Rio dos Sinos

Marcello Restelli
Politecnico di Milano

Peter Vamplew
Federation University

Diederik M. Roijers
City of Amsterdam

ABSTRACT

Real-world sequential decision-making tasks are usually complex, and require trade-offs between multiple – often conflicting – objectives. However, the majority of research in reinforcement learning (RL) and decision-theoretic planning assumes a single objective, or that multiple objectives can be handled via a predefined weighted sum over the objectives. Such approaches may oversimplify the underlying problem, and produce suboptimal results. This extended abstract outlines the limitations of using a semi-blind iterative process to solve multi-objective decision making problems. Our extended paper [4], serves as a guide for the application of explicitly multi-objective methods to difficult problems.

KEYWORDS

Multi-Objective, Reinforcement Learning, Planning

ACM Reference Format:
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. 2023. A Brief Guide to Multi-Objective Reinforcement Learning and Planning: JAAMAS track. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

* Equal contribution.
** The work was done prior to joining Amazon.

1 INTRODUCTION

Real-world sequential decision-making tasks can have multiple, often conflicting, objectives [1, 3, 5, 8, 21]. Reinforcement learning (RL) and decision-theoretic planning have been used extensively to solve such problems by maximising a scalar reward signal [16]. In settings with multiple objectives, RL approaches assume the objectives can be combined into a single scalar reward using a predefined weighted sum. An iterative process is used to tune the weights for each objective. During learning the algorithm is tuned, turned on, then the reward function is re-engineered until the behaviour is satisfactory. However, such an approach may produce suboptimal results in practical settings [2, 18, 20].

Here, we argue an iterative process is problematic for a number of reasons: (a) it is a semi-blind manual process, (b) it prevents people who should take the decisions from making well-informed trade-offs, (c) it damages the explainability of the decision-making process, (d) it cannot handle different types of preferences that human decision makers might actually have, and finally (e) preferences between the objectives may change over time and a single objective agent will have to be retrained when this happens.

Motivating Example. Planning a journey involves a number of objectives (such as minimising travel time and cost whereas maximising comfort and reliability [6, 7, 11, 12]), together with *sequential* decisions that need to be made along the trip. For instance, if your trip relies on multiple transportation modes, you may need to promptly switch to another mode when facing delays or malfunctions. Moreover, given the competitive nature of traffic, your objectives are usually affected by other users, which increases the uncertainties associated with your decision. If you cannot articulate your preferences explicitly in a single formula, or if this formula

is non-linear, then you have a genuine multi-objective problem, which requires a multi-objective approach.

2 A MULTI-OBJECTIVE APPROACH

As highlighted above, single objective RL and planning methods utilise an iterative process to linearly combine objectives when solving multi-objective problems. Such an approach has several limitations, and as a result an explicitly multi-objective approach should be followed. To motivate a multi-objective approach, we briefly discuss the aforementioned limitations.

First, let us discuss reason (a). If we engineer a scalar reward function through an iterative process until we reach an acceptable behaviour, we try out multiple reward functions, each of which is a scalarisation of the actual objectives. However, we do not systematically inspect all possible reward functions. In other words, we may meet our minimal threshold for acceptable behaviours, but we only observe a subset of all possible scalarisations. Therefore, although an acceptable solution may be found, it can be arbitrarily far away from optimal utility.

Next, let us discuss reason (b), given the reward function needs to be engineered a priori, there is uncertainty as to the effects a reward function may have on the policy. Additionally, the decision power is put where it does not belong: with the AI engineers, since they are the ones tasked with adjusting the associated weights for the reward function and, thus, effectively making assumptions about the preferences of the actual decision makers. In practical settings this is not a responsibility that can be left to AI engineers. By taking an explicitly multi-objective approach it is possible to remove such responsibilities from the AI engineer. Multi-objective algorithms can be used to compute all possibly optimal policies [13, 14, 19, 23], where the computed policies can be inspected by a system expert before making a decision.

Another issue with scalar reward functions is the lack of (a posteriori) explainability (c). Not taking an explicitly multi-objective approach can rob us of essential information that we might need to evaluate or understand our agents. Consider the case in which a robot collided with and destroyed a vase and we would like to input an alternative decision, such as swerving away from the vase. An agent with a single all encompassing objective that has learnt a scalar value function will then, for example, tell us there was a 3.451 reduction in value for this other policy, which provides little insight. If instead, the agent could have told us that in the objective of damage to property the probability of damaging the vase would have dropped to practically 0, but the probability of running into the family dog increased by 0.5% (a different objective), this would give us insight into what went wrong. We might also think that a 0.5% increase in the likelihood of bumping into the dog would have been acceptable – especially if this would not have been an actual danger to it, but rather an inconvenience – if the robot could have definitely avoided destroying the vase, signaling an error in the utility function. We may also disagree for different reasons: we may think that the agent has overestimated the risk of colliding with the dog, i.e., an error in the value-estimate for that objective.

Furthermore (d), not all human preferences can be handled by scalar additive reward functions [13]. In certain settings, a user's preferences ought to be modelled with a non-linear utility function. For non-linear utility functions, an a priori scalarisation becomes mathematically impossible within many reinforcement learning frameworks, as scalarisation would break the additivity of the reward function. For some domains, this might still be acceptable, as the resulting loss of optimality may not have a major impact. However, in important domains where ethical or moral issues become apparent, single-objective approaches require explicitly combining these factors together with other objectives (such as economic outcomes) in a way that may be unacceptable to many people [22]. Similarly, designing single-objective rewards may be difficult or impossible for scenarios where we wish to ensure fair or equitable outcomes for multiple participants [15, 17].

Finally (e), humans are known to change their minds from time to time. Therefore, preferences between trade-offs in the different objectives may well change over time. An explicitly multi-objective system can train agents to be able to handle such preference changes, thereby preempting the need to discover a new policy whenever such changes occur. This increases the applicability of multi-objective decision-making agents, as agents do not need to be taken out of operation to be updated and they can simply switch policy to match the new user preferences.

3 THE UTILITY-BASED APPROACH

Early work in multi-objective sequential decision-making largely adopted an axiomatic approach in which the optimal solution set is assumed to be the Pareto front (see [4] for all definitions). An advantage of this approach is that it leads to a solution set which will contain an optimal policy for any possible monotonically increasing utility function, and axiomatic methods can derive these solutions without any need to explicitly consider the details of those potential utility functions. However, this set is typically large, and may be prohibitively expensive to retrieve.

In practical applications, a lot more might be known about the utility function of the user, due to domain knowledge. Using an axiomatic approach would make it difficult to exploit this knowledge, and a lot of time and effort might be spent on computing a solution set which contains some members with very low utility for the user. The utility-based approach aims to derive the optimal solution set from the available knowledge about the utility function of the user, and which types of policies are allowed. This knowledge allows constraints to be placed on the solution set, reducing its size and thereby improving learning efficiency and making it easier for users or systems to select their preferred policy [13]. Considering the user utility first is key to the successful application of any AI in decision problems. In multi-objective problems, it is especially important, as the properties of the user's utility may drastically alter the desired solution, what methods are available, and even—in some cases [9, 10]—whether stable solutions even exist.

4 CONCLUSION

Considering the reasons outlined above, a multi-objective approach to decision making is necessary in many practical settings. In this work, we briefly outline why taking an explicitly multi-objective utility-based approach to planning and learning may be essential to deploying AI-based solution for real-world sequential decision problems.

ACKNOWLEDGMENTS

Conor F. Hayes is funded by the University of Galway Hardiman Scholarship. This research was supported by funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" program. Roxana Rădulescu is supported by the Research Foundation Flanders (FWO postdoctoral fellowship 1286223N). Johan Källström and Fredrik Heintz were partially supported by the Swedish Governmental Agency for Innovation Systems (grant NFFP7/2017-04885), and the Wallenberg Artificial Intelligence, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. Luisa Zintgraf was supported by the 2017 Microsoft Research PhD Scholarship Program, and the 2020 Microsoft Research EMEA PhD Award.

REFERENCES

[1] Daniel Bryce, William Cushing, and Subbarao Kambhampati. 2007. Probabilistic planning is multi-objective. *Arizona State University, Tech. Rep. ASU-CSE-07-006* (2007).

[2] Tim Brys, Kristof Van Moffaert, Kevin Van Vaerenbergh, and Ann Nowé. 2013. On the behaviour of scalarization methods for the engagement of a wet clutch. In *2013 12th International Conference on Machine Learning and Applications*, Vol. 1. IEEE, 258–263.

[3] A Castelletti, Francesca Pianosi, and Marcello Restelli. 2013. A multiobjective reinforcement learning approach to water resources systems operation: Pareto frontier approximation in a single run. *Water Resources Research* 49, 6 (2013), 3476–3486.

[4] 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. 2022. A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents and Multi-Agent Systems* 36, 1 (2022), 1–59.

[5] Ammar Jalalimanesh, Hamidreza Shahabi Haghighi, Abbas Ahmadi, Hossein Hejazian, and Madjid Soltani. 2017. Multi-objective optimization of radiotherapy: distributed Q-learning and agent-based simulation. *Journal of Experimental & Theoretical artificial intelligence* 29, 5 (2017), 1071–1086.

[6] Patrick Mannion, Jim Duggan, and Enda Howley. 2016. An Experimental Review of Reinforcement Learning Algorithms for Adaptive Traffic Signal Control. In *Autonomic Road Transport Support Systems*. Springer, Cham, 47–66. https://doi.org/10.1007/978-3-319-25808-9_4

[7] Juan de Dios Ortúzar and Luis G. Willumsen. 2011. *Modelling transport* (4 ed.). John Wiley & Sons, Chichester, UK.

[8] Francesca Pianosi, Andrea Castelletti, and Marcello Restelli. 2013. Tree-based fitted Q-iteration for multi-objective Markov decision processes in water resource management. *Journal of Hydroinformatics* 15, 2 (2013), 258–270.

[9] 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).

[10] 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), e32. https://doi.org/10.1017/S0269888920000351

[11] Gabriel de O. Ramos, Bruno C. da Silva, Roxana Rădulescu, Ana L. C. Bazzan, and Ann Nowé. 2020. Toll-based reinforcement learning for efficient equilibria in route choice. *The Knowledge Engineering Review* 35 (2020), e8. https://doi.org/10.1017/S0269888920000119

[12] Gabriel de O. Ramos, Roxana Rădulescu, Ann Nowé, and Anderson R. Tavares. 2020. Toll-Based Learning for Minimising Congestion under Heterogeneous Preferences. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020)*, B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, and G. Sukthankar (Eds.). IFAAMAS, Auckland, New Zealand, 1098–1106.

[13] 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.

[14] Diederik M. Roijers, Shimon Whiteson, and Frans A. Oliehoek. 2015. Computing Convex Coverage Sets for Faster Multi-Objective Coordination. *Journal of Artificial Intelligence Research* 52 (2015), 399–443.

[15] Umer Siddique, Paul Weng, and Matthieu Zimmer. 2020. Learning Fair Policies in Multiobjective (Deep) Reinforcement Learning with Average and Discounted Rewards. In *International Conference on Machine Learning*.

[16] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

[17] Peter Vamplew, Richard Dazeley, Cameron Foale, Sally Firmin, and Jane Mummery. 2018. Human-aligned artificial intelligence is a multiobjective problem. *Ethics and Information Technology* 20, 1 (2018), 27–40.

[18] Peter Vamplew, John Yearwood, Richard Dazeley, and Adam Berry. 2008. On the limitations of scalarisation for multi-objective reinforcement learning of Pareto fronts. In *Australasian Joint Conference on Artificial Intelligence*. Springer, 372–378.

[19] Kristof Van Moffaert and Ann Nowé. 2014. Multi-objective reinforcement learning using sets of pareto dominating policies. *The Journal of Machine Learning Research* 15, 1 (2014), 3483–3512.

[20] Kevin Van Vaerenbergh, Abdel Rodríguez, Matteo Gagliolo, Peter Vrancx, Ann Nowé, Julian Stoev, Stijn Goossens, Gregory Pinte, and Wim Symens. 2012. Improving wet clutch engagement with reinforcement learning. In *The 2012 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[21] Timothy Verstraeten, Ann Nowé, Jonathan Keller, Yi Guo, Shuangwen Sheng, and Jan Helsen. 2019. Fleetwide data-enabled reliability improvement of wind turbines. *Renewable and Sustainable Energy Reviews* 109 (2019), 428–437. https://doi.org/10.1016/j.rser.2019.03.019

[22] Wendell Wallach and Colin Allen. 2008. *Moral machines: Teaching robots right from wrong*. Oxford University Press.

[23] DJ White. 1982. Multi-objective infinite-horizon discounted Markov decision processes. *Journal of mathematical analysis and applications* 89, 2 (1982), 639–647.