Multi-objective Reinforcement Learning: A Tool for Pluralistic Alignment

Peter Vamplew, Cameron Foale
Federation University Australia
p.vamplew,c.foale@federation.edu.au

Conor F. Hayes
Lawrence Livermore National Laboratory
hayes56@llnl.gov

Richard Dazeley
Deakin University
richard.dazeley@deakin.edu.au

Hadassah Harland
Deakin University
hharland@deakin.edu.au

Abstract

Reinforcement learning (RL) is a valuable tool for the creation of AI systems. However it may be problematic to adequately align RL based on scalar rewards if there are multiple conflicting values or stakeholders to be considered. Over the last decade multi-objective reinforcement learning (MORL) using vector rewards has emerged as an alternative to standard, scalar RL. This paper provides an overview of the role which MORL can play in creating pluralistically-aligned AI.

1 Introduction

Reinforcement learning (RL) has emerged as one of the most powerful tools for creating AI systems capable of autonomous decision-making for sequential tasks [Sutton and Barto, 2018]. Its core mechanism of learning to maximise the expected future return derived from a scalar reward signal can allow it to reach or even exceed human levels of performance. However, this also creates a strong dependency on an accurate definition of the reward signal. If the reward is misspecified or underspecified, the behaviour learned by an RL agent may deviate significantly from what is desired [Taylor, 2016]. Recent studies have shown that current approaches to reward specification may frequently lead to specification errors [Booth et al., 2023, Knox et al., 2023].

Vamplew et al. [2018] argued that framing alignment as a multi-objective problem may assist in overcoming the issues of creating aligned agents using RL. Treating each aspect of the alignment task as a separate objective within a vector reward signal may aid in producing aligned behaviour which is difficult or impossible to achieve using a scalar definition of reward [Vamplew et al., 2022]. Recent years have seen an increasing amount of research activity applying multi-objective reinforcement learning (MORL) techniques to various aspects of alignment.

This paper starts with a brief review of MORL, highlighting its relevance to alignment, before examining the potential of MORL methods for pluralistic alignment, including examples of prior work. These examples are illustrative rather than a comprehensive review, and will focus on alignment of large language models (LLMs) as that has been one of the main areas of application so far.

2 A brief review of MORL

MORL methods assume the environment is a Multi-objective Markov Decision Process (MOMDP), which is a MDP with a vector reward function $\mathbf{R} : S \times A \times S \rightarrow \mathbb{R}^d$ with d objectives. The agent aims to learn a policy π which maximises the return derived from $\mathbf{R}$. However, as both $\mathbf{R}$ and

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

the return are vectors, it is not possible to define a full-ordering over policies. In order to do this, algorithms often assume the existence of a *utility function* $u : \mathbb{R}^d \rightarrow \mathbb{R}$. If u is known in advance and fixed, then the agent can learn a single-policy which is optimal for u [Hayes et al., 2022]. Where u is unknown, subject to change, or difficult to explicitly define, the agent may instead find a set of policies which are optimal under different parameterisations of u. This is known as *multi-policy* MORL. The final decision of which policy to execute can then be selected at run-time. If u is difficult to explicitly define, this policy selection might be carried out directly by the system's stakeholders.

Some MORL systems assume that u is a linear-weighted sum. This is simple to implement, as the MOMDP can be mapped to an equivalent single-objective MDP [Roijers et al., 2013]. However it may fail to correctly capture the intended behaviour, so MORL methods often instead use monotonically-increasing non-linear utility functions. This introduces algorithmic complications, but more accurately represents the stakeholders' true utility. Note that u now can not be applied to the reward received per time-step - instead it is applied to the (possibly discounted) vector returns $\mathbf{v}$ [1].

$$\mathbf{v} = \sum_{t=0}^{T} \gamma^t \mathbf{R}_t \tag{1}$$

3 Applications of MORL to pluralistic alignment

Sorensen et al. [2024b] define three categories of benchmarks for pluralistic alignment, which can be interpreted as specifying desirable characteristics of pluralistic agents. These characteristics are 1) the agent be multi-objective in nature (which will support value-pluralism [Sorensen et al., 2024a]), 2) the agent is steerable to support customisation of trade-offs between objectives, and 3) the agent is able to consider a diverse set of user preferences. In this section we explore how MORL methods can support each of these desirable characteristics, either individually or simultaneously.

3.1 MORL for value pluralistic alignment

This aspect of pluralistic alignment supports consideration of diverse values (e.g. personal freedom, societal harmony, economic benefits, environmental impact). Clearly, MORL naturally supports this aspect as each value can be represented by a separate objective within the reward function $\mathbf{R}$.

MORL has been applied in a number of contexts to enable a system to balance multiple conflicting values. Examples include performance versus safety tradeoffs [Vamplew et al., 2021, Smith et al., 2023], compliance with moral standards or norms [Rodriguez-Soto et al., 2022, Peschl et al., 2021], or wellbeing, affordability, equity, and environmental sustainability [Chaput et al., 2023].

In the context of finetuning LLMs, Wang, K. et al. (2024) present an approach that involves conditioning model weights on preference weights over the objectives, and demonstrate performance on three objectives reflecting different desirable properties of text summarisation. Wu et al. [2024] use fine-grained human feedback to fine-tune models based on linear combinations of rewards derived from separate reward models for factual accuracy, relevance, and information completeness. Meanwhile Wang, H. et al. (2024) use a context-sensitive mixture-of-experts approach to tune LLM output to trade-off between 19 different objectives which capture values such as honesty, verbosity and safety, while Yang et al. [2024] address the values of harmfulness, helpfulness, and humour.

3.2 MORL for Steerable pluralistic alignment

A key advantage of multi-policy MORL is that it is inherently customisable with respect to stakeholder preferences. Sorensen et al. [2024b] state that for many applications, customisation of the trade-off between objectives at run-time is a desirable characteristic (for example, to match the preferences of the current user of an AI system). Similarly, Chatila et al. [2017] argued for the criticality of being able to adapt to changing values or preferences at a societal level. If the stakeholder using a system changes or if the preferences of an existing stakeholder change, a new policy which is optimal with

[1] For clarity and brevity, we are simplifying some aspects of MORL here. For a more detailed and nuanced discussion, see Hayes et al. [2022].

respect to their preferences can be immediately identified, without the need for re-training which would be required for a single-objective RL agent [Hayes et al., 2022].

Harland et al. [2023] provides an example of these benefits. Their agent learns a Pareto set of policies, and selects a policy to execute. After each action the agent observes the reaction of a human user. If it detects that the human is displeased, the agent apologises, updates its model of the human's preferences, and selects a new policy which is compatible with that model.

Rame et al. [2024] argue that for very large models, it may be impractical to explicitly learn a Pareto set of policies. Instead they start from a pre-trained model, and separately fine-tune a separate copy per objective. They then use linear weight interpolation across these specialised models at run-time to produce a model which is aligned with a particular desired trade-off between the different objectives.

3.3 MORL for Jury-Pluralistic Alignment

Jury-pluralism refers to accounting for the varying preferences amongst a diverse set of stakeholders (system users and/or people or groups impacted by the decisions of the AI). In order to support this in MORL, two things must be true: (1) The set of objectives and corresponding rewards must include all aspects of the problem considered relevant by any stakeholder, and (2) the choice of policy to be executed must reflect the interests of all stakeholders.

This might be achieved by defining a utility-function *u* representing the interests of all parties. In practice this will be difficult, and may simply represent the average preferences of the population, overlooking minority views (a criticism levelled at existing RL from human feedback (RLHF) approaches Sorensen et al. [2024b]).Alternatively, multi-policy MORL could learn a set of Pareto-optimal policies, with the choice of policy to execute made via consultation or voting – this will be time-consuming, and subject to the limitations of voting systems [Dai and Fleisig, 2024].

Sorensen et al's definition of jury-pluralism assumes that each stakeholder is mapped to a scalar reward or utility, and that the agent makes decisions so as to maximise a welfare function applied over this set of stakeholder utilities. This definition can be mapped directly onto an MORL framework by assigning an objective within the MOMDP per stakeholder, and then applying a utility function *u* that finds a suitable tradeoff over all stakeholders.

This framework requires a diverse set SH of identified stakeholders $\{sh_1, ..., sh_n\}$. Each element of the reward $\mathbf{R}$ corresponds to a scalar representation of the values of a specific stakeholder (i.e., the number of objectives $d = n$). The choice of utility function u should appropriately account for the desires of each stakeholder. A relevant line of research here is the body of work on *fair MORL*, where various fair utility functions have been considered. Both Yu et al. [2023] and Michailidis et al. [2023] use the Generalised Gini social welfare function (GGF), which sorts the utilities of stakeholders into ascending order and applies a linear weighting with weights of decreasing value (i.e. placing more emphasis on the lower-valued utilities), as shown in Equation 2. Yu et al. [2023] also propose the use of an extended form, the Generalised GGF[2], which allows certain stakeholders to be prioritised. Meanwhile, Fan et al. [2022] propose using the Nash Social Welfare function (Equation 3), arguing it has the benefit of being invariant to the scale of the stakeholder utilities.

$$GGF_{\mathbf{w}}(\mathbf{v}) = \sum_{i=1}^{d} \mathbf{w}_i \mathbf{v}_i^{\uparrow} \qquad (2) \qquad\qquad NSW(\mathbf{v}) = \left(\prod_{i=1}^{d} \mathbf{v}_i \right)^{\frac{1}{d}} \qquad (3)$$

The concepts of individual preference models and an aggregation function have been applied in the context of LLM finetuning by Park et al. [2024]. They enable the specific form of aggregation function they use (probabilistic opinion pooling) by requiring users to provide feedback in the form of a vector of the probability of selection of each sample LLM output provided in response to a prompt, rather than simply indicating a single preferred response.

3.4 MORL for Fully Pluralistic Alignment

The earlier sections address each dimension of pluralistic alignment separately. Jury-pluralistic MORL (Section 3.3) balances the preferences of a diverse set of stakeholders, while value-pluralistic

[2]Yes, that is the Generalised Generalised Gini social welfare function!

(Section 3.1) finds trade-offs across a set of values. Meanwhile, the steerability of multi-policy methods (Section 3.2) applies regardless of the nature of the objectives. Here we consider how MORL might support a steerable agent which is both value-pluralistic and jury-pluralistic.

As in Section 3.3, a fully-pluralistic agent requires an identified set of n stakeholders SH. However rather than representing each stakeholder's desires directly via a scalar reward value, in this framework the rewards represent varied objectives (values) which may be prioritised differently by each stakeholder. Each stakeholder's preferences over those objectives are then represented by a personalised utility function u_i. These individual utility functions are then aggregated by a system-level utility function. For example, the GGF from Equation 2 can be extended to this more general framework as follows:

$$GGF_{\mathbf{w}}(\mathbf{v}) = \sum_{i=1}^{d} \mathbf{w}_i \mathbf{u}_i^{\uparrow}(\mathbf{v}) \tag{4}$$

Multi-policy MORL methods can learn a coverage set of policies representing possible trade-offs between the objectives thereby supporting fairness across the stakeholders, while also allowing for future changes in the preferences of each stakeholder. The appropriate policy for execution can then be determined at run-time, providing a high level of steerability.

While we are not aware of any existing work on fully-pluralistic MORL agents, the PRISM Alignment Project [Kirk et al., 2024] presents an important enabling step towards fully-pluralistic LLMs, by curating a dataset of human feedback suitable for the construction of the multiple reward models required for fully-pluralistic RLHF. This feedback has been gathered from a more diverse set of participants than other feedback datasets, including attempts to provide wider geographic and cultural coverage. Each feedback item is labelled with demographic and other information about the feedback provider. In addition, the feedback is fine-grained, with a rating provided relative to multiple attributes (overall values, fluency, factuality, safety, diversity, creativity, and helpfulness). The dataset also contains measures of the importance which each provider places on each attribute, which would be suitable for creating personalised utility functions u_i.

4 Challenges

A significant issue impeding the development of pluralistic agents using MORL is the lack of suitable human feedback datasets for developing reward models. While the PRISM dataset is a major step forward, its authors acknowledge that it still lacks sufficient diversity (for example, the content is entirely in English, and feedback was gathered from online workers) [Kirk et al., 2024]. Creation of a truly representative dataset will be a major undertaking. Wang, H. et al. (2024) leverage multiple datasets in an attempt to address this issue. However this is complicated as the datasets do not use consistent or compatible preference attributes.

There is a further potential technical issue associated with the approaches proposed in Section 3. The majority of extant MORL algorithms are efficient only for a relatively small number of objectives (most work considers only 2-4 objectives). While this might suffice for some models such as the helpfulness/harmfulness trade-off in LLMs, it may not suffice more generallly. For example, value-pluralistic alignment based on Schwartz's theory of basic values, would require ten objectives [Schwartz, 2012]. Similarly in jury-pluralistic alignment, the set of stakeholders *SH* may be large, and hence the MOMDP will have a correspondingly high number of objectives. Therefore there is a need for further research extending MORL approaches to tasks involving many objectives. This may require fundamentally different methods, as has been found to be the case in optimisation, where multi-objective and many-objective cases often use different algorithms [Fleming et al., 2005]. The weight interpolation approach used by Rame et al. [2024] is an example of the sort of algorithmic innovation which may be required in order to scale up to high-dimensional reward spaces.

Acknowledgments and Disclosure of Funding

This research was supported by Founder's Pledge, the Berkeley Existential Risk Institute, and the Future of Life Institute. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

References

Serena Booth, W Bradley Knox, Julie Shah, Scott Niekum, Peter Stone, and Alessandro Allievi. The perils of trial-and-error reward design: misdesign through overfitting and invalid task specifications. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5920–5929, 2023.

Rémy Chaput, Laetitia Matignon, and Mathieu Guillermin. Learning to identify and settle dilemmas through contextual user preferences. In *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 474–479. IEEE, 2023.

Raja Chatila, Kay Firth-Butterflied, John C Havens, and Konstantinos Karachalios. The IEEE global initiative for ethical considerations in artificial intelligence and autonomous systems. *IEEE Robotics and Automation Magazine*, 24(1):110, 2017.

Jessica Dai and Eve Fleisig. Mapping social choice theory to RLHF. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*, 2024.

Zimeng Fan, Nianli Peng, Muhang Tian, and Brandon Fain. Welfare and fairness in multi-objective reinforcement learning. *arXiv preprint arXiv:2212.01382*, 2022.

Peter J Fleming, Robin C Purshouse, and Robert J Lygoe. Many-objective optimization: An engineering design perspective. In *International conference on evolutionary multi-criterion optimization*, pages 14–32. Springer, 2005.

Hadassah Harland, Richard Dazeley, Bahareh Nakisa, Francisco Cruz, and Peter Vamplew. AI apology: interactive multi-objective reinforcement learning for human-aligned AI. *Neural Computing and Applications*, 35(23):16917–16930, 2023.

Conor F Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M Zintgraf, Richard Dazeley, Fredrik Heintz, et al. A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents and Multi-Agent Systems*, 36(1):26, 2022.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, et al. The PRISM alignment project: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *arXiv preprint arXiv:2404.16019*, 2024.

W Bradley Knox, Alessandro Allievi, Holger Banzhaf, Felix Schmitt, and Peter Stone. Reward (mis) design for autonomous driving. *Artificial Intelligence*, 316:103829, 2023.

Dimitris Michailidis, Willem Röpke, Sennay Ghebreab, Diederik M Roijers, and Fernando P Santos. Fairness in transport network design-a multi-objective reinforcement learning approach. In *2023 Adaptive and Learning Agents Workshop at AAMAS*, 2023.

Chanwoo Park, Mingyang Liu, Dingwen Kong, Kaiqing Zhang, and Asuman E Ozdaglar. RLHF from heterogeneous feedback via personalization and preference aggregation. In *ICML 2024 Workshop: Aligning Reinforcement Learning Experimentalists and Theorists*, 2024.

Markus Peschl, Arkady Zgonnikov, Frans A Oliehoek, and Luciano C Siebert. MORAL: Aligning AI with human norms through multi-objective reinforced active learning. *arXiv preprint arXiv:2201.00012*, 2021.

Alexandre Rame, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, and Matthieu Cord. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *Advances in Neural Information Processing Systems*, 36, 2024.

Manel Rodriguez-Soto, Marc Serramia, Maite Lopez-Sanchez, and Juan Antonio Rodriguez-Aguilar. Instilling moral value alignment by means of multi-objective reinforcement learning. *Ethics and Information Technology*, 24(1):9, 2022.

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

Shalom H Schwartz. An overview of the Schwartz theory of basic values. *Online readings in Psychology and Culture*, 2(1):11, 2012.

Benjamin J Smith, Robert Klassert, and Roland Pihlakas. Using soft maximin for risk averse multi-objective decision-making. *Autonomous Agents and Multi-Agent Systems*, 37(1):11, 2023.

Taylor Sorensen, Liwei Jiang, Jena D Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, et al. Value Kaleidoscope: Engaging AI with pluralistic human values, rights, and duties. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19937–19947, 2024a.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*, 2024b.

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

Jessica Taylor. Quantilizers: A safer alternative to maximizers for limited optimization. In *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

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

Peter Vamplew, Cameron Foale, Richard Dazeley, and Adam Bignold. Potential-based multiobjective reinforcement learning approaches to low-impact agents for AI safety. *Engineering Applications of Artificial Intelligence*, 100:104186, 2021.

Peter Vamplew, Benjamin J Smith, Johan Källström, Gabriel Ramos, Roxana Rădulescu, Diederik M Roijers, Conor F Hayes, Fredrik Heintz, Patrick Mannion, Pieter JK Libin, et al. Scalar reward is not enough: A response to Silver, Singh, Precup and Sutton (2021). *Autonomous Agents and Multi-Agent Systems*, 36(2):41, 2022.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. *arXiv preprint arXiv:2406.12845*, 2024a.

Kaiwen Wang, Rahul Kidambi, Ryan Sullivan, Alekh Agarwal, Christoph Dann, Andrea Michi, Marco Gelmi, Yunxuan Li, Raghav Gupta, Avinava Dubey, et al. Conditioned language policy: A general framework for steerable multi-objective finetuning. *arXiv preprint arXiv:2407.15762*, 2024b.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36, 2024.

Rui Yang, Xiaoman Pan, Feng Luo, Shuang Qiu, Han Zhong, Dong Yu, and Jianshu Chen. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment. *arXiv preprint arXiv:2402.10207*, 2024.

Guanbao Yu, Umer Siddique, and Paul Weng. Fair deep reinforcement learning with generalized Gini welfare functions. In *International Conference on Autonomous Agents and Multiagent Systems*, pages 3–29. Springer, 2023.