Sample-Efficient Multi-Task and Multi-Objective Reinforcement Learning by Combining Multiple Behaviors

Lucas N. Alegre[1], **Ana L. C. Bazzan**[1], **Bruno C. da Silva**[2]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
91.501-970 – Porto Alegre – RS – Brazil

[2]College of Information & Computer Sciences – University of Massachusetts
Amherst – MA – USA

lnalegre@inf.ufrgs.br, bazzan@inf.ufrgs.br, bsilva@cs.umass.edu

Abstract. *One of the main challenges in the field of artificial intelligence, and reinforcement learning (RL) in particular, is the development of generalist and flexible agents capable of solving multiple tasks—each requiring the agent to learn a potentially new, specialized behavior. Tackling this challenge requires agents to learn behaviors that may involve optimizing a single objective, or trading off between multiple conflicting objectives. In this thesis, we study how to design flexible RL agents that can, in a sample-efficient manner, adapt their behavior to solve any given tasks—each of which is defined by multiple (possibly conflicting) objectives. We introduce new* multi-policy *methods that empower RL agents to (i) carefully learn multiple behaviors, each specialized in a particular task; and (ii) combine previously-learned behaviors to efficiently identify solutions to novel tasks, which, importantly, may require the agent to assign different preferences to each of its new objectives. The methods we introduce have strong theoretical guarantees regarding the optimality of the set of behaviors learned by agents and their capability to solve new tasks in a zero-shot manner, even in the presence of function approximation errors. We evaluate the proposed methods in various challenging multi-task and multi-objective RL problems and show that our algorithms outperform various current state-of-the-art methods in domains with both discrete and continuous state and action spaces.*

1. Introduction

In Reinforcement Learning (RL), agents learn to maximize rewards by interacting with their environment through a process of trial and error. While RL techniques have achieved remarkable success in solving complex single-task problems [Silver et al. 2017, Bellemare et al. 2020], most standard RL algorithms are designed to learn a *single* policy tailored to *one* specific task. Real-world problems, by contrast, often require agents to solve *multiple* tasks or to balance multiple conflicting objectives—a setting known as Multi-Objective RL (MORL) [Hayes et al. 2022]. Although the ability to adapt behavior to solve multiple tasks and trade off between conflicting objectives is a hallmark of human intelligence, developing agents capable of efficiently adapting and generalizing to new tasks and objectives remains a key open challenge in the field of AI.

In this thesis, in particular, we investigate the problem of **how to design *flexible* RL agents that can, in a *sample-efficient manner*, *adapt their behavior* to solve *any***

given tasks defined by multiple objectives. We hypothesize that insights from Multi-Task RL (MTRL) and MORL—two subfields often studied independently—can be combined to design novel and principled techniques to address this challenge. In particular, we focus on *multi-policy* methods that empower RL agents to: *(i)* carefully learn a set of policies/behaviors, each specialized in solving a particular task or behaving optimally with respect to one possible preference over objectives; and *(ii)* combine previously-acquired behaviors to identify solutions to arbitrary new tasks efficiently, often in a zero-shot way, without requiring any further learning or interaction with the environment.

In particular, we introduce new methods based on *generalized policy improvement* (GPI) [Barreto et al. 2020]. GPI is a principled technique that enable agents to, when tackling new tasks, rapidly identify new behaviors that are guaranteed to be at least as good (and generally better) as any of the agent's previously acquired behaviors. The contributions of this thesis significantly improve the capabilities of GPI-based multi-policy RL agents and enable them to solve challenging multi-task *and* multi-objective RL problems in a sample-efficient (and often in a *zero-shot*) manner.

As a key contribution of this thesis, we show that the techniques we introduce enable agents to solve the *optimal policy transfer* problem: constructing a set of policies such that combining them *directly* leads to the optimal policy for *any* novel linear task. In particular, this thesis makes the following main contributions:

- We formalize the connection between multi-task and multi-objective RL and introduce a novel algorithm that solves the previously open problem of optimal policy transfer [Alegre et al. 2022].
- We propose novel prioritization mechanisms for GPI to significantly improve sample efficiency in MORL [Alegre et al. 2023b]. We also introduce an uncertainty-aware extension of GPI that allows it to scale to more challenging high-dimensional problems [Alegre et al. 2026].
- We extend GPI to leverage approximate environment models for better zero-shot transfer, proposing h-GPI, a multi-step extension of GPI that interpolates between model-free policy transfer and model-based planning. Importantly, it allows the agent to become arbitrarily less susceptible to inaccuracies in its learned value function representations [Alegre et al. 2024].
- Finally, we present a comprehensive open-source toolkit for reliable benchmarking and research in MORL, including MO-Gymnasium and MORL-Baselines [Felten et al. 2023]. MO-Gymnasium has been officially adopted as the core multi-objective module of Gymnasium (formerly OpenAI Gym), the most widely used platform for reinforcement learning experimentation. Thousands of researchers worldwide rely on it to design, evaluate, and compare algorithms.

2. Related Work

Two key fields of AI underlie the problems and settings studied in this thesis: multi-task RL (MTRL) and multi-objective RL (MORL). The literature on MTRL is broad, and the definition of *task* varies across different works [Taylor and Stone 2009, Finn et al. 2017, Barreto et al. 2017]. In this thesis, we study settings where tasks differ only by their reward function, and focus on the open problem of how to carefully construct a set of behaviors (*policies*) such that combining them directly leads to the optimal solution for

any novel tasks. We refer to this as the *optimal policy transfer* problem. Although prior methods based on policy transfer have been proposed [Pickett and Barto 2002, Fernández and Veloso 2006, Taylor et al. 2007, Abel et al. 2018], they do not address the problem of constructing a set of policies that allows for *optimal* behaviors (over arbitrary tasks) to be identified directly, without requiring any further learning interactions with the environment.

Similarly, several multi-policy algorithms have been proposed in the MORL literature [Van Moffaert and Nowé 2014, Abels et al. 2019, Yang et al. 2019, Xu et al. 2020]. A key distinction of our work is that we, unlike existing techniques, leverage *transfer learning* to improve the sample efficiency of MORL algorithms. Additionally, previous methods are heuristic and do not have theoretical guarantees of convergence to a CCS (a set containing all Pareto optimal solutions). We, by contrast, show that our methods are guaranteed to converge to an ϵ-CCS, where ϵ is the optimality gap of the algorithm used to learn policies, in a finite number of steps. The contributions based on GPI we introduce in this thesis allow us to design flexible and adaptive agents capable of learning policies in a significantly more sample-efficient manner than previous MORL methods, advancing the state-of-the-art, from an empirical point of view, but also through key novel strong theoretical guarantees on the robustness of the AI agents trained by our methods.

3. Optimal Policy Transfer

In RL, an agent's task is encoded by a *reward function*. When reward functions are expressed as linear combinations of features, and the agent has previously learned a set of policies for different tasks, the framework of *successor features* (SFs) and *generalized policy improvement* (GPI) [Barreto et al. 2017] can be exploited to identify reasonable policies for new tasks in a zero-shot manner. Intuitively, GPI generalizes the policy improvement step by improving a given policy, tasked with solving a particular task, over a *set* of policies, instead of a single one. However, the resulting policy is not guaranteed to be optimal. In [Alegre et al. 2022], we introduce a novel algorithm that addresses this limitation and solves the following *optimal policy transfer* problem: **how to carefully construct a set of policies such that combining them directly leads to the optimal policy for *any* novel linearly-expressible tasks?**

We first show (under mild assumptions) that the transfer learning problem tackled by SFs is equivalent to the problem of learning to optimize multiple objectives in RL. We then introduce a new algorithm, SFOLS (Successor Features Optimistic Linear Support), that learns a set of policies whose SFs form a convex coverage set (CCS) (an optimal set of policies for transfer). In particular, we show that the policies in this set can be combined via GPI to construct optimal behaviors for any new linearly-expressible tasks, without requiring additional training samples.

Theoretical Contributions: This work introduces the *first mathematical framework* characterizing the underlying relation between two seemingly disparate subfields of machine learning: transfer learning (via Successor Features) and Multi-Objective Reinforcement Learning. By leveraging it, we introduce the *first method* capable of solving the **optimal policy transfer problem** for linearly-expressible tasks, which was previously a key open problem in the AI literature.

3.1. Empirical Results

We evaluate SFOLS in discrete and continuous-state domains and compare it with state-of-the-art competitors for constructing diverse policy libraries (e.g., WCPI [Zahavy et al. 2021]). Across domains, SFOLS learns a policy set (*behavior basis*) that yields significantly better zero-shot performance while requiring fewer iterations to ensure high-quality coverage over the distribution of tasks.

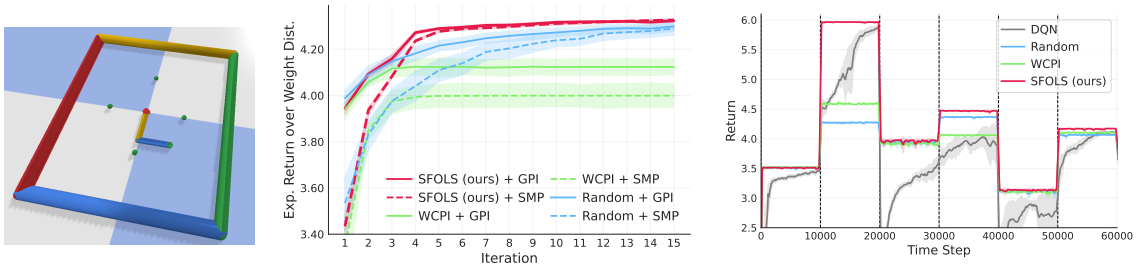

Figure 1. Left: Simulated robotics problem with d=4 reward features/targets. Middle: Expected performance over a distribution of tasks when combining the policy library learned by SFOLS and competing baselines. Right: Performance of SFOLS in a lifelong learning setting.

In Fig. 1, we show that SFOLS consistently achieves better performance (expected return) under a distribution of test tasks while requiring agents to train and identify significantly fewer behaviors (policies). In the right panel, we show that SFOLS can also be used in a *lifelong learning* setting, in which the reward functions change in a non-stationary manner, to quickly adapt its behavior to arbitrary new tasks.

Main Takeaways: SFOLS efficiently learns a compact behavior basis that performs well over the entire space of tasks. It yields robust zero-shot performance when compared to state-of-the-art policy library construction methods, *immediately* adapting the agent's behavior to novel tasks even in non-stationary lifelong learning scenarios.

4. Sample-Efficient MORL via Generalized Policy Improvement

In [Alegre et al. 2023b, Alegre et al. 2026], we introduce a novel MORL algorithm that improves sample efficiency via two novel prioritization techniques. The primary challenge we address is: **how to learn a set of policies containing optimal policies for any user preference in a sample-efficient and robust manner?**

If the agent's reward function in a MORL problem is a linear combination of its objectives, optimal solutions are sets of policies known as *convex coverage sets* (CCS) [Hayes et al. 2022]. Given a CCS, agents can *directly* identify the optimal solution to any novel linear preferences. MORL algorithms that learn a CCS ([Mossalam et al. 2016, Yang et al. 2019]) are sample inefficient ***(i)*** due to the heuristics they use to determine which preferences to train on, at any given moment during the construction of a CCS; and ***(ii)*** because they can only improve a CCS after optimal (or near-optimal) policies are identified—which may require a large number of samples collected through environment interactions.

We address the first challenge via a novel algorithm, GPI-LS, which employs a GPI-based prioritization technique for selecting which preferences to train on. GPI-LS

prioritizes preferences based on a lower bound on performance improvements guaranteed to be achievable via GPI, which accurately and reliably identifies the most relevant preferences to train on when learning a CCS. To address the second issue, we show that our method is an anytime algorithm that monotonically improves the quality of its CCS, even if given *intermediate* (possibly sub-optimal) policies for different preferences. This improves sample efficiency: our method identifies intermediate CCSs with formally bounded maximum utility loss even if there are constraints on the number of times the agent can interact with its environment. GPI-LS is guaranteed to always converge to an optimal solution in a finite number of steps, or an ϵ-optimal solution (for a bounded ϵ) if the agent is limited and can only identify possibly sub-optimal policies.

Theoretical Contributions: We introduce principled GPI-based prioritization schemes guaranteed to optimally guide the training process and identify an optimal set of policies in a finite number of steps. Furthermore, we introduce the **first model-based MORL method capable of dealing with continuous state spaces** and equip our methods with an uncertainty-aware capabilities to robustly handle approximation errors. Thereby, we introduce a novel technique with bounds on its achievable performance that significantly reduces the required number of samples needed to solve tasks compared to existing approaches.

4.1. Empirical Results

We first evaluate our methods on a simulated robot, where an agent is required to dynamically keep balance and hop forward. Our algorithms consistently identify solutions with higher expected utility than state-of-the-art baselines (Fig. 2). Moreover, the Pareto frontier identified by our method covers a broader range of behaviors that trade-off between the two objectives: maximizing forward speed and jumping height. In the MO-Super Mario domain (Fig. 3), our empirical results highlight that GPI-based prioritization remains effective even in settings with high-dimensional observations (pixel frames).

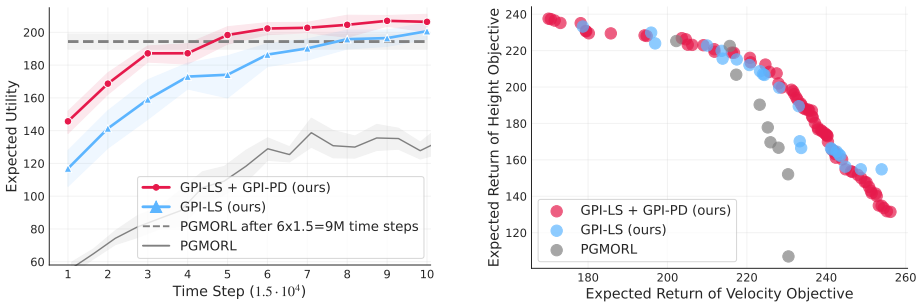

Figure 2. MO-Hopper results. Left: expected performance over a set of preferences over objectives. Right: Pareto frontier produced by each method. Our GPI-based methods achieve higher performance and better coverage of the objective space.

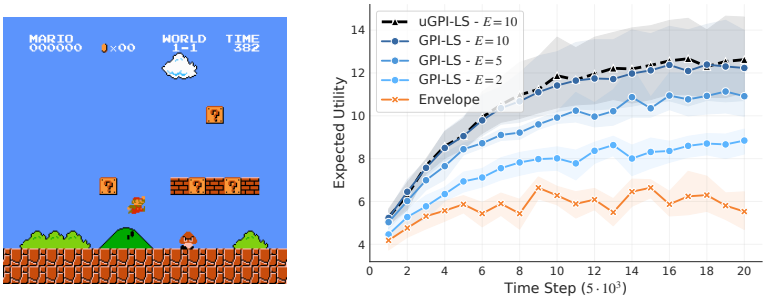

Figure 3. MO-Super Mario results. Left: environment with d=4 conflicting objectives. Right: expected utility over a set of preferences. Our GPI-based approach improves sample efficiency, and its uncertainty-aware action selection capabilities further stabilize learning.

Main Takeaways: Our GPI-based methods significantly improve sample efficiency compared to prior state-of-the-art MORL algorithms, across both discrete and complex continuous control tasks, via a novel active learning mechanism that allows agents to decide how to best alocate their learning time. GPI-LS and GPI-PD consistently identify near-optimal solutions and achieve higher expected performance compared to state-of-the-art competitors. Our method's uncertainty-aware variant (uGPI) robustly stabilizes learning in high-dimensional domains and is supported by novel performance bounds twice as tight as the best bounds known in the literature.

5. Multi-Step GPI with Approximate Models

The prior contributions focus on constructing and combining a library of behaviors (policies). Importantly, the methods we introduced in the previous sections tackled this problem in a *model-free* setting, i.e., purely from interacting with the environment. A key complementary challenge is: **how to leverage approximate models of the environment to improve zero-shot policy transfer?**

To address this problem, in [Alegre et al. 2023a] we introduce h-GPI, a multi-step extension of GPI that leverages an *approximate* environment model to plan for h steps and then estimate the agent's performance from future states with model-free GPI. At a high level, h-GPI interpolates between standard model-free GPI (when $h = 0$) and fully model-based planning (as h grows). Given an approximate model of the environment $\hat{m}$ and a policy library Π with corresponding action-value estimates, h-GPI selects an action by maximizing the expected return over h model steps and then, at the frontier states reachable after h steps, using the standard GPI value estimate. A key practical design choice is to learn a model that predicts *reward features* (instead of a scalar reward), enabling immediate generalization to any new linear reward function defined by a weight vector encoding the preferences over each feature/objective.

We provide formal guarantees showing that, under bounded approximation errors, the performance lower bound of h-GPI is strictly better than that of standard GPI whenever model errors are negligible. Moreover, the impact of value approximation errors in the policy library is discounted by a factor γ^h, implying that increasing h makes h-GPI arbitrarily less susceptible to inaccuracies in learned SFs/action-values. At the same time, increasing h increases dependence on the approximate model, highlighting a principled

trade-off between model errors and value approximation errors.

Theoretical Contributions: We introduce h-GPI, a framework capable of interpolating between model-free transfer (GPI) and fully model-based planning. We provide formal bounds on how the planning horizon h trades-off approximation errors in the agent's learned model and action-value functions, making the agent arbitrarily less susceptible to inaccuracies in its learned value function representations.

5.1. Empirical Results

We evaluate h-GPI in both tabular and continuous-state domains. Across domains, we observe that incorporating short-horizon planning substantially improves zero-shot transfer performance over standard GPI. In the stochastic FourRoom domain (Fig. 4 (a)), increasing h leads to consistent improvements on unseen test tasks, matching the intuition that deeper lookahead reduces the reliance on potentially inaccurate GPI value estimates.

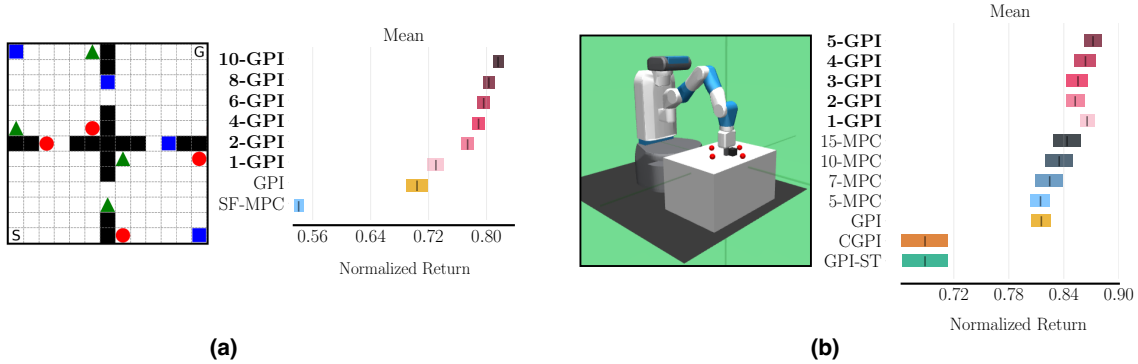

Figure 4. Mean normalized performance over test tasks in the (a) FourRoom and (b) robotic arm (FetchPush) domain. The performance of h-GPI increases significantly as the planning horizon (h) increases.

In a high-dimensional robotics domain (Fig. 4 (b)), h-GPI outperforms standard GPI and state-of-the-art baselines even when allowed to plan for only a few steps. This indicates that small amounts of planning already provide meaningful corrections to purely model-free GPI action selection and reduce sensitivity to value approximation errors.

Main Takeaways: h-GPI consistently outperforms standard GPI and state-of-the-art competing baselines across discrete and continuous-state domains. It requires a significantly smaller planning horizon to achieve consistently superior performance and makes the agent strictly less susceptible to value approximation errors.

6. Conclusion and Future Work

This thesis investigated the research problem of designing *flexible* reinforcement learning (RL) agents that can, in a *sample-efficient manner*, *adapt their behavior* to solve *any* given tasks—each defined via a separate reward function and possibly multiple conflicting objectives. In particular, this thesis set out to demonstrate that, based on novel formal and mathematical connections between multi-objective RL (MORL) and multi-task RL (MTRL), the above-mentioned problem can be tackled via novel and principled *multi-policy* methods that empower RL agents to *(i) carefully learn a set of policies*, each of

which specialized in solving a particular task under different preferences over objectives; and *(ii) adaptively combine such behaviors* to rapidly (or instantaneously) identify new behaviors that solve any given unseen tasks in a zero-shot manner.

This work represents a major advancement in the field of AI and RL. We introduce not only methods with strong theoretical guarantees, many of which significantly improves upon the existing performance bounds, but also empirically outperform all state-of-the-art competitors in challenging domains. This type of robustness and assured performance, even in the face of changing tasks and priorities between objectives, is a key step towards enabling the use of AI methods for optimal decision making in critical applications where strict performance guarantees are paramount.

Theoretical Impact: We solved a key open problem of *optimal policy transfer*, significantly advancing the state-of-the-art in multi-objective reinforcement learning. The contributions of this thesis have already had a significant impact in the multi-task and multi-objective RL communities. Immediately after the thesis defense, we introduced a *hierarchical* method to combine behaviors that extends the expressiveness and zero-shot transfer capabilities of the methods proposed in this thesis, enabling also for the optimal zero-shot solution to *non-linear* tasks [Alegre et al. 2025a].
Practical and Open-Source Impact: We developed **MO-Gymnasium**, now the global standard for MORL research as part of the **Farama Foundation**, with thousands of users and contributors. Moreover, the open-source contributions have been widely adopted by the RL community, with over 80 citations in the literature and over 850 stars on GitHub.
Real-World Applications: As a result of an internship at Disney Research Zürich, in [Alegre et al. 2025b], a novel MORL controller leveraging the methods proposed in this thesis was introduced. It was deployed in real life and used to control a real-world robot to perform diverse and complex motions, such as fighting and dancing.
Scientific Recognition: This thesis has led to eight publications in top-tier venues (NeurIPS, ICML, AAMAS, TMLR) that have influenced many subsequent works.

Future Work. As future work, we plan to employ our methods to enhance agent-based large language models (LLMs) [Liu et al. 2023], which strongly rely on existing RL techniques. Aligning LLMs with human values has been shown to require balancing conflicting objectives, such as succinctness and completeness [Wang et al. 2024]. Our multi-policy provides a principled way for LLMs to dynamically adapt to arbitrary user preferences, enabling more flexible, reliable, and efficient multi-objective alignment with human values and goals.

References

Abel, D., Jinnai, Y., Guo, S. Y., Konidaris, G., and Littman, M. (2018). Policy and value transfer in lifelong reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80.

Abels, A., Roijers, D. M., Lenaerts, T., Nowé, A., and Steckelmacher, D. (2019). Dynamic weights in multi-objective deep reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97.

Alegre, L. N., Bazzan, A. L. C., Barreto, A., and Silva, B. C. d. (2025a). Constructing an Optimal Behavior Basis for the Option Keyboard. In *Advances in Neural Information Processing Systems 38*.

Alegre, L. N., Bazzan, A. L. C., and da Silva, B. C. (2022). Optimistic linear support and successor features as a basis for optimal policy transfer. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162.

Alegre, L. N., Bazzan, A. L. C., Nowé, A., and da Silva, B. C. (2023a). Multi-step generalized policy improvement by leveraging approximate models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, volume 36.

Alegre, L. N., Bazzan, A. L. C., Roijers, D. M., Nowé, A., and da Silva, B. C. (2023b). Sample-efficient multi-objective learning via generalized policy improvement prioritization. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems*.

Alegre, L. N., Bazzan, A. L. C., Roijers, D. M., Nowé, A., and da Silva, B. C. (2024). Generalized policy improvement for efficient and robust multi-objective reinforcement learning. *Autonomous Agents and Multiagent Systems (JAAMAS)*.

Alegre, L. N., Roijers, D. M., Nowé, A., Bazzan, A. L. C., and da Silva, B. C. (2026). Generalized policy improvement for efficient and robust multi-objective reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 40(1).

Alegre, L. N., Serifi, A., Grandia, R., Müller, D., Knoop, E., and Bächer, M. (2025b). AMOR: Adaptive Character Control through Multi-Objective Reinforcement Learning. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, SIGGRAPH Conference Papers '25.

Barreto, A., Dabney, W., Munos, R., Hunt, J. J., Schaul, T., van Hasselt, H. P., and Silver, D. (2017). Successor features for transfer in reinforcement learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, volume 30.

Barreto, A., Hou, S., Borsa, D., Silver, D., and Precup, D. (2020). Fast reinforcement learning with generalized policy updates. *Proceedings of the National Academy of Sciences*, 117(48).

Bellemare, M. G., Candido, S., Castro, P. S., Gong, J., Machado, M. C., Moitra, S., Ponda, S. S., and Wang, Z. (2020). Autonomous navigation of stratospheric balloons using reinforcement learning. *Nature*, 588(7836).

Felten, F., Alegre, L. N., Nowé, A., Bazzan, A. L. C., Talbi, E.-G., Danoy, G., and da Silva, B. C. (2023). A toolkit for reliable benchmarking and research in multi-objective reinforcement learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, volume 36.

Fernández, F. and Veloso, M. (2006). Probabilistic policy reuse in a reinforcement learning agent. In *Proceedings of the 5th International Joint Conference on Autonomous Agents and Multiagent Systems*.

Finn, C., Abbeel, P., and Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70.

Hayes, C. F., Rădulescu, R., Bargiacchi, E., Källström, J., Macfarlane, M., Reymond, M., Verstraeten, T., Zintgraf, L. M., Dazeley, R., Heintz, F., Howley, E., Irissappane, A. A., Mannion, P., Nowé, A., Ramos, G., Restelli, M., Vamplew, P., and Roijers, D. M. (2022). A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents and Multi-Agent Systems*, 36(1).

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., Wu, Z., Zhao, L., Zhu, D., Li, X., Qiang, N., Shen, D., Liu, T., and Ge, B. (2023). Summary of ChatGPT-related research and perspective towards the future of large language models. *Meta-Radiology*, 1(2).

Mossalam, H., Assael, Y. M., Roijers, D. M., and Whiteson, S. (2016). Multi-objective deep reinforcement learning. *arXiv preprint arXiv:1610.02707*.

Pickett, M. and Barto, A. G. (2002). Policyblocks: An algorithm for creating useful macro-actions in reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., and Hassabis, D. (2017). Mastering the game of go without human knowledge. *Nature*, 550(7676).

Taylor, M. E. and Stone, P. (2009). Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10(56).

Taylor, M. E., Whiteson, S., and Stone, P. (2007). Transfer via inter-task mappings in policy search reinforcement learning. In *Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems*. IFAAMAS.

Van Moffaert, K. and Nowé, A. (2014). Multi-objective reinforcement learning using sets of Pareto dominating policies. *Journal of Machine Learning Research*, 15(1).

Wang, K., Kidambi, R., Sullivan, R., Agarwal, A., Dann, C., Michi, A., Gelmi, M., Li, Y., Gupta, R., Dubey, K. A., Rame, A., Ferret, J., Cideron, G., Hou, L., Yu, H., Ahmed, A., Mehta, A., Hussenot, L., Bachem, O., and Leurent, E. (2024). Conditional language policy: A general framework for steerable multi-objective finetuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Xu, J., Tian, Y., Ma, P., Rus, D., Sueda, S., and Matusik, W. (2020). Prediction-guided multi-objective reinforcement learning for continuous robot control. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119.

Yang, R., Sun, X., and Narasimhan, K. (2019). A generalized algorithm for multi-objective reinforcement learning and policy adaptation. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*.

Zahavy, T., Barreto, A., Mankowitz, D. J., Hou, S., O'Donoghue, B., Kemaev, I., and Singh, S. (2021). Discovering a set of policies for the worst case reward. In *Proceedings of the 9th International Conference on Learning Representations*.