Adaptive Alignment: Dynamic Preference Adjustments via Multi-Objective Reinforcement Learning for Pluralistic AI

Hadassah Harland
Deakin University, Australia
h.harland@research.deakin.edu.au

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

Peter Vamplew
Federation University, Australia
p.vamplew@federation.edu.au

Hashini Senaratne
CSIRO, Australia
hashini.senaratne@data61.csiro.au

Bahareh Nakisa
Deakin University, Australia
bahar.nakisa@deakin.edu.au

Francisco Cruz
University of New South Wales, Australia
f.cruz@unsw.edu.au

Abstract

Emerging research in Pluralistic AI alignment seeks to address how to design and deploy intelligent systems in accordance with diverse human needs and values. We contribute a potential approach for aligning AI with diverse and shifting user preferences through Multi-Objective Reinforcement Learning (MORL), via post-learning policy selection adjustment. This paper introduces the proposed framework, outlines its anticipated advantages and assumptions, and discusses technical details for implementation. We also examine the broader implications of adopting a retroactive alignment approach from a sociotechnical systems perspective.

1 Introduction

Pluralistic alignment has emerged as an area of growing interest within Artificial Intelligence (AI) research [Sorensen et al., 2024a,b]. The term unifies ideas about the diverse, multifaceted, and evolving nature of human values and the challenge this presents to human-aligned AI [Jain et al., 2024, Vamplew et al., 2018]. Given the pluralistic nature of human preferences, human-aligned AI systems must autonomously and independently adapt to fit individual users, use cases, and contexts.

Multi-objective reinforcement learning (MORL) is a powerful AI technique for autonomous sequential decision-making tasks involving multiple, often conflicting, objectives [Hayes et al., 2021]. Multi-policy MORL algorithms can learn several solutions in parallel, each optimised for different objective trade-offs, that can be dynamically selected at runtime. This adaptability and capacity for balancing competing objectives makes MORL a promising platform for pluralistic alignment research.

This paper presents a MORL-based approach to pluralistic AI via *adaptive alignment*, using retroactive policy selection adjustments to continuously realign to user preferences. First, we briefly review AI alignment research, exploring key challenges and highlighting the need for a multi-objective approach. Section 3 presents an adaptive alignment framework, with three stages: *learning*, *selection*, and *execution and review*. We discuss technical considerations for implementation in Section 4. To conclude, we examine implications of retroactive adjustment, emphasising the unavoidable need for post-interaction realignment and the importance of active transparency in human-AI interactions.

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

2 Challenges in AI alignment

Research in AI alignment has grown increasingly critical as AI systems continue gaining ability and prevalence [Taylor et al., 2020]. Ji et al. [2023] describe these efforts according to two main streams; *forward alignment* considers how to design new systems that meet these demands, whereas *backward alignment* looks at regulation, governance, and assurance of existing systems.

Reinforcement learning (RL)-based approaches feature prominently in forward alignment [Ji et al., 2023], leveraging the premise of learning an optimal policy by seeking to maximise an expected cumulative reward [Sutton and Barto, 2018]. In particular, *Reinforcement Learning from Human Feedback* is a popular approach [Ouyang et al., 2022], where the reward function is derived from human preferences. These models can be criticised as resource intensive, both computationally and in requiring manually labelled data [Cao et al., 2024, Casper et al., 2023], leading to the emergence of alternative approaches for automating alignment. For example, the *Constitutional AI* [Bai et al., 2022] and *Reinforcement Learning from AI Feedback* [Lee et al., 2023] algorithms replace the human in the training loop with another AI model to enable self-improvement with less human feedback.

However, these approaches can oversimplify the alignment problem by not accounting for the pluralistic nature of human values [Sorensen et al., 2024b]. The needs and values of different people may vary broadly; even for a single individual across differing contexts, the exact requirements cannot be universally defined [Gabriel, 2020, Mishra, 2023]. By resolving to a single, static solution, these algorithms leave no space to accommodate the natural variability in values and preferences between users and contexts. Furthermore, without the ability to adapt, these solutions may become outdated as preferences change over time.

Adaptive alignment may help to address this limitation, enabling pluralistic system expression to represent diverse human values and perspectives. Interactive machine learning approaches such as *In-context* [Dong et al., 2022] and *Active* [Taylor et al., 2021] Learning have largely focused on task generalisation. However, RL-based approaches have emerged, enabling adaptive value alignment by representing the task as a Multi-Objective Markov Decision Process (MOMDP) and employing MORL techniques [Harland et al., 2023, Peschl et al., 2022, Rame et al., 2023, Yang et al., 2024].

The need for multi-objective approaches for human-alignment in RL is well established [Casper et al., 2023, Mannion et al., 2021, Vamplew et al., 2018], as *a priori* scalarisation of objectives does not allow for the necessary exploration, visibility, or flexibility of the solution to support alignment [Hayes et al., 2021]. Conversely, representing human values as distinct objectives allows the agent to separately evaluate and balance competing priorities, enabling exploration of a diverse range of potential solutions [Vamplew et al., 2018]. For example, whether to prioritise cleaning or avoid disruptions [Harland et al., 2023, Peschl et al., 2022], or how to balance humour, helpfulness, and harmlessness in a chatbot response [Yang et al., 2024]. Of particular relevance to pluralistic alignment, MORL enables multi-policy learning, such that the specific policy to be executed can be selected *a posteriori* to the learning process [Hayes et al., 2021].

Yet, a major challenge remains; how may a suitable policy be selected given potentially unknown and dynamic user preferences? Hayes et al. [2021] propose the *review and adjust* scenario to address how a MORL system may adapt to dynamic user preferences, describing a process of retroactive updates via manual user selection. We propose an extension to this scenario with an approach that circumvents the manual selection process to dynamically adapt to diverse user preferences.

3 An adaptive alignment framework

In this section, we introduce a framework for pluralistic AI through adaptive alignment in MORL, modelled after the *review and adjust* process (Section 2). The proposed agent adapts to the user's preferences through a *self-review* process, using context and informal signals to minimise the need for direct and specific feedback from the user (Figure 1). Two key features are distinct: an initial default policy is chosen in the selection phase, and the role of reviewer is shifted from the user to the agent.

The basis of this framework is a trained, multi-objective, multi-policy RL algorithm that has learned a set of solutions representing the scope of possible human preferences across multiple different values. The algorithm represents these values as distinct objectives, and each solution describes an optimal policy for a particular set of preferences over these objectives.

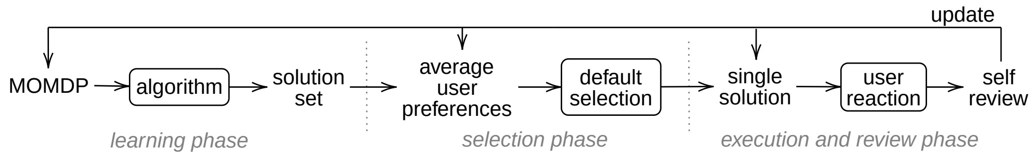

Figure 1: The *self-review* process leverages indirect feedback to adjust to the users' preferences.

The initial policy is selected according to the predicted best fit for the user. For an unknown user, this may either be a universal default selection, or can be initially personalised according to what information is available; for example, the system might prioritise brevity over detail when responding to a voice query. For a familiar user, the choice of policy is informed by previous interactions.

With each execution, the agent observes the user's reaction (e.g., facial expression, nonverbal audio) and performs a self-review. The process draws on information collected about the interaction and relevant contextual factors to identify any misalignment between the current policy and the user's preferences. The agent then selects a new policy to dynamically adjust its behaviour accordingly.

We anticipate some of the advantages to this approach to be the following:

Feedback efficiency and focus: The approach alleviates the need for explicit feedback by using the user's reaction as a signal, which should be less burdensome on the user and minimise the influence of response bias. Furthermore, the brevity of the feedback signal provides a narrow focus, which should inherently reduce the dimensionality of the feedback according to its importance to the user.

Aligning with multiple users: This framework has been designed with multiple users in mind. If the current user changes, the previous user's preferred policy could be stored when the new user's profile is created or loaded, so the system can retain what it has learned about each previous user while operating according to the new user's preferences.

Continuous evolution: The repeating self-review process provides a continuous feedback loop that should enable the system to accommodate new preferences as they arise and so maintain alignment with the users' evolving needs. By also updating the average users' preferences, the system should also be able to improve the initial select for new users and evolve at a broader social level.

4 Techniques for adaptive alignment

The framework described in the previous section relies on two key assumptions: 1) a suitable model can be developed to accurately detect and attribute misalignment in the system based on the information available to the agent at execution, and 2) given the output from this model, the system can perform an update by selecting an alternative policy that is better aligned with the user's preferences. In this section, we discuss specific techniques to address these assumptions.

4.1 Interpreting user reactions as feedback

The first assumption requires an interpretation model that enables the user's reaction to act as a feedback signal. That is, we want a model M to transform a reaction signal ζ into an update $\vec{\Delta\Xi}$ to the user's preferences $\vec{\Xi}$. This model should incorporate information about the interaction (θ) so that the signal can be interpreted in context ($M \to M(\theta)$). Constituent factors should include the outcome of the execution [MacGlashan et al., 2017], the usual distribution of user preferences for the given use case, and the history of any prior interactions with this user.

One approach could be to define $\vec{\Delta}$ explicitly using a loss measure derived from individually *idealised* reward values R_i^{ideal} for each objective i. We assume the reaction signal $\zeta \in \mathcal{N}(\mu, \sigma^2)$ to be a normal scalar, but other forms are possible [Jeon et al., 2020]. Equation 1 provides an example, given $\hat{\zeta} \in \mathcal{N}(0,1)$ transformed via Bayesian estimation, scaling factor α_i, and activation threshold τ_i.

$$\Delta_i = \alpha_i \hat{\zeta}(R_i^{observed} - R_i^{ideal}) - \tau_i \quad \forall i \tag{1}$$

An alternative approach could be to employ a RL algorithm by representing the task as a contextual bandit problem [Bouneffouf et al., 2020]. The model would use the context θ and a reward signal derived from the transformed reaction $\hat{\zeta}$. Similar models have previously been used for simulating cognitive empathy in human-robot interaction [Bagheri et al., 2021].

4.2 Solution updates via post-learning policy selection adjustment

The second assumption requires a process for selecting a policy $\pi' \in \Pi$ that best aligns with the user's preferences Ξ from the set of known Pareto-optimal policies Π [Hayes et al., 2021]. Possible approaches depend on how Π is represented. Pareto-based methods, such as Pareto Q-Learning [Moffaert and Nowé, 2014], store learned policies as vector returns. The format eases direct policy comparison at the cost of high computational complexity to reproduce a policy from its expected return [Felten, 2024]. Conversely, approximate methods, such as conditioned networks [Abels et al., 2019], may learn a parametric policy $\pi(\phi)$, or use interpolation to compute a mixture policy using a weighted combination of learned policies [Rame et al., 2023, Yang et al., 2024].

If each policy in Π can be mapped directly to a return vector, it is possible to calculate an ordering over the policies using a utility function u derived from Ξ. The definition of u could be as simple as applying $\vec{\Xi}$ directly as weights for linear scalarisation [Hayes et al., 2021], or may incorporate non-linear features such as thresholds and lexicographical ordering [Harland et al., 2023].

If a direct ordering over the policies is not feasible, it might be more suitable to employ a steering-based approach [Vamplew et al., 2017]. Instead of selecting a completely new policy, steering enables stepwise updates by moving along the Pareto front to the next closest policy or mixture of policies in the direction of the update $\vec{\Delta\Xi}$. This approach benefits from a smaller search space and progressive updates that may appear more stable to the user, but may be slower to implement large-scale changes.

The agent adapts its behaviour by executing the updated policy selection π'. The update itself is strictly not a learning process, as the underlying policies are fixed. However, this update process might also help inform aspects of earlier phases (Figure 1): contributing additional data towards the average user preferences to continuously adapt the default selection, and providing an indication of possible objectives not captured in the MOMDP.

5 Implications of a retroactive approach

The adaptive alignment framework we proposed in this paper follows a retroactive approach to pluralistic AI, with some accompanying implications. We consider these implications through the sociotechnical systems perspective; in matters related to human users, AI algorithms are inseparable from the sociotechnical systems within which they are embedded [Kudina and van de Poel, 2024].

Technical challenges for safety: As noted by Ji et al. [2023], algorithms that learn through human feedback may be particularly susceptible to risks of reward hacking and scalable oversight [Amodei et al., 2016]. This could be further aggravated by a self-supervisory method such as we have described that may allow potential issues to be obscured. To minimise this risk, it may be beneficial to incorporate backward alignment features such as explanations to provide transparency on how update decisions are made.

Inevitability of misalignment: Prior to the first interaction with a given user, it is not possible to know that user's preferences perfectly. This fact goes beyond any question of feasibility; even if you were to assume the most advanced superintelligence conceivable, it is not philosophically impossible to predict a user's preferences with absolute certainty. Thus, there will always be a need for AI systems to perform retroactive corrections to realign with the evolving needs of the user. The framework proposed herein is an example of one such system, but it does not need to be used in isolation. Rather, this approach should be combined with suitable and effective predictive alignment techniques to minimise the use of adaptive alignment to only where it is necessary.

The need for explanations and repair: The retroactive approach relies on information from the user to identify discrepancies between the current settings and the user's true preferences. This necessitates that misaligned behaviour has already occurred, and corrective action alone may be insufficient to address any harms incurred. Furthermore, users may adapt their own behaviour throughout an interaction, as they develop an understanding of how the system behaves. The system may continue to adapt, but any subsequent solution will be calibrated according to the context of the previous interaction, one step behind. Thus, a retroactive alignment approach may warrant incorporating both reparations and explanations to support user needs at the sociotechnical scale.

Acknowledgments and Disclosure of Funding

This research was supported by an Australian Government Research Training Program (RTP) and a Commonwealth Scientific and Industrial Research Organisation (CSIRO) Top-Up Scholarship.

References

Axel Abels, Diederik M Roijers, Tom Lenaerts, Ann Nowé, and Denis Steckelmacher. Dynamic Weights in Multi-Objective Deep Reinforcement Learning. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete Problems in AI Safety. 6 2016. URL `http://arxiv.org/abs/1606.06565`.

Elahe Bagheri, Oliver Roesler, Hoang Long Cao, and Bram Vanderborght. A Reinforcement Learning Based Cognitive Empathy Framework for Social Robots. *International Journal of Social Robotics*, 13(5):1079–1093, 8 2021. ISSN 18754805. doi: 10.1007/s12369-020-00683-4.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI Feedback. 12 2022. URL `http://arxiv.org/abs/2212.08073`.

Djallel Bouneffouf, Irina Rish, and Charu Aggarwal. Survey on Applications of Multi-Armed and Contextual Bandits. In *2020 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8. IEEE, 7 2020. ISBN 978-1-7281-6929-3. doi: 10.1109/CEC48606.2020.9185782.

Boxi Cao, Keming Lu, Xinyu Lu, Jiawei Chen, Mengjie Ren, Hao Xiang, Peilin Liu, Yaojie Lu, Ben He, Xianpei Han, Le Sun, Hongyu Lin, and Bowen Yu. Towards Scalable Automated Alignment of LLMs: A Survey. 6 2024. URL `http://arxiv.org/abs/2406.01252`.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Cornell Tech, Jérémy Scheurer, Apollo Research, Javier Rando, Eth Zurich, Rachel Freedman, Berkeley Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Micah Carroll, Andi Peng, Phillip Christoffersen, Stewart Slocum, Mit Csail, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Erdem Bıyık, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL `https://openreview.net/forum?id=bx24KpJ4Eb`.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, and Zhifang Sui. A Survey on In-context Learning. 2022.

Florian Felten. *Multi-Objective Reinforcement Learning*. PhD thesis, University of Luxembourg, 7 2024.

Iason Gabriel. Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30(3):411–437, 9 2020. ISSN 15728641. doi: 10.1007/s11023-020-09539-2.

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, 8 2023. ISSN 0941-0643. doi: 10.1007/s00521-023-08586-x. URL `https://link.springer.com/10.1007/s00521-023-08586-x`.

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. A Practical Guide to Multi-Objective Reinforcement Learning and Planning. *Autonomous Agents and Multi-Agent Systems*, 36(1), 3 2021. doi: 10.1007/s10458-022-09552-y. URL `http://arxiv.org/abs/2103.09568http://dx.doi.org/10.1007/s10458-022-09552-y`.

Shomik Jain, Vinith Suriyakumar, Kathleen Creel, and Ashia Wilson. Algorithmic Pluralism: A Structural Approach To Equal Opportunity. In *2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2024*, pages 197–206. Association for Computing Machinery, Inc, 6 2024. ISBN 9798400704505. doi: 10.1145/3630106.3658899.

Hong Jun Jeon, Smitha Milli, and Anca Dragan. Reward-rational (implicit) choice: A unifying formalism for reward learning. 2020.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Kwan Yee Ng, Juntao Dai, Xuehai Pan, Aidan O'Gara, Yingshan Lei, Hua Xu, Brian Tse, Jie Fu, Stephen McAleer, Yaodong Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo, and Wen Gao. AI Alignment: A Comprehensive Survey. 10 2023. URL `http://arxiv.org/abs/2310.19852`.

Olya Kudina and Ibo van de Poel. A sociotechnical system perspective on AI. *Minds and Machines*, 34(3):21, 6 2024. ISSN 1572-8641. doi: 10.1007/s11023-024-09680-2.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. 9 2023. URL `http://arxiv.org/abs/2309.00267`.

James MacGlashan, Mark K Ho, Robert Loftin, Bei Peng, Guan Wang, David L Roberts, Matthew E Taylor, and Michael L Littman. Interactive Learning from Policy-Dependent Human Feedback. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2285–2294. PMLR, 9 2017. URL `https://proceedings.mlr.press/v70/macglashan17a.html`.

Patrick Mannion, Fredrik Heintz, Thommen George Karimpanal, and Peter Vamplew. Multi-Objective Decision Making for Trustworthy AI. In *Proceedings of the Multi-Objective Decision Making (MODeM) Workshop*, 2021.

Abhilash Mishra. AI Alignment and Social Choice: Fundamental Limitations and Policy Implications. *SSRN Electronic Journal*, 2023. doi: 10.2139/ssrn.4605509.

Kristof Van Moffaert and Ann Nowé. Multi-Objective Reinforcement Learning using Sets of Pareto Dominating Policies. *Journal of Machine Learning Research*, 15:3663–3692, 2014.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang Sandhini Agarwal Katarina Slama Alex Ray John Schulman Jacob Hilton Fraser Kelton Luke Miller Maddie Simens Amanda Askell, Peter Welinder Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Markus Peschl, Arkady Zgonnikov, Frans A. Oliehoek, and Luciano C. Siebert. MORAL: Aligning AI with Human Norms through Multi-Objective Reinforced Active Learning. In *AAMAS 2022: 21st International Conference on Autonomous Agents and Multiagent Systems (Virtual)*, pages 1038–1046. International Foundation for Autonomous Agents and Multiagent Systems, 12 2022. URL `http://arxiv.org/abs/2201.00012`.

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. In A Oh, T Naumann, A Globerson, K Saenko, M Hardt, and S Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 71095–71134. Curran Associates, Inc., 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/e12a3b98b67e8395f639fde4c2b03168-Paper-Conference.pdf`.

Taylor Sorensen, Liwei Jiang, Jena Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, Maarten Sap, John Tasioulas, and Yejin Choi. Value Kaleidoscope: Engaging AI with Pluralistic Human Values, Rights, and Duties. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 19937–19947, 9 2024a. doi: 10.1609/aaai.v38i18.29970. URL `http://arxiv.org/abs/2309.00779http://dx.doi.org/10.1609/aaai.v38i18.29970`.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. Position: A Roadmap to Pluralistic Alignment. In *Forty-first International Conference on Machine Learning*, 2024b.

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

Annalisa T. Taylor, Thomas A. Berrueta, and Todd D. Murphey. Active learning in robotics: A review of control principles. *Mechatronics*, 77, 8 2021. ISSN 09574158. doi: 10.1016/j.mechatronics.2021.102576.

Jessica Taylor, Eliezer Yudkowsky, Patrick LaVictoire, and Andrew Critch. Alignment for Advanced Machine Learning Systems. In *Ethics of Artificial Intelligence*, pages 342–382. Oxford University Press, 9 2020. doi: 10.1093/oso/9780190905033.003.0013. URL `https://academic.oup.com/book/33540/chapter/287906349`.

Peter Vamplew, Rustam Issabekov, Richard Dazeley, Cameron Foale, Adam Berry, Tim Moore, and Douglas Creighton. Steering approaches to Pareto-optimal multiobjective reinforcement learning. *Neurocomputing*, 263:26–38, 11 2017. ISSN 09252312. doi: 10.1016/j.neucom.2016.08.152.

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), 2018. ISSN 15728439. doi: 10.1007/s10676-017-9440-6.

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. 2 2024. URL `http://arxiv.org/abs/2402.10207`.