Multi-Objective Decision Making for Trustworthy AI

Patrick Mannion*
School of Computer Science
National University of Ireland Galway
patrick.mannion@nuigalway.ie

Fredrik Heintz*
Dept. of Computer Science
Linköping University, Sweden
fredrik.heintz@liu.se

Thommen George Karimpanal*
Applied Artificial Intelligence Institute
Deakin University
thommen.karimpanalgeorge@deakin.edu.au

Peter Vamplew*
School of Engineering, IT and Physical Sciences
Federation University
p.vamplew@federation.edu.au

ABSTRACT

If widespread deployment of AI systems is to be accepted by society in the future, it is crucial that such systems are trustworthy. Trustworthiness for autonomous systems has a number of dimensions including safety, ethics, fairness and explainability. In this paper, we argue that an explicitly multi-objective decision making approach is the correct way to ensure trustworthiness in AI systems, and we address the following fundamental questions in support of our argument: 1) Why is it necessary to treat trustworthiness as a multi-objective problem? 2) How should rewards for a trustworthy agent be specified? 3) How should a trustworthy agent reason over multiple objectives? 4) Where should preferences for a trustworthy agent come from?

KEYWORDS

Multi-objective decision making, trustworthy AI, fairness, ethics, safety, explainability, reinforcement learning, planning

1 INTRODUCTION

A key global challenge is ensuring that AI is beneficial to humanity. The EU has for example decided to focus on human-centered trustworthy AI based on strong collaborations among key stakeholders to maximise the opportunities and minimise the risks. Trustworthiness is a prerequisite for people and societies to develop, deploy, and use AI systems. Since there are many important and incommensurable factors in trustworthiness including transparency, privacy and fairness, they naturally lead to multi-objective decision making (MODeM) problems.

MODeM problems appear in a wide variety of real-world scenarios. For example, when choosing sources for electricity generation, fossil fuels are often cheaper than renewable energy sources such as wind power, however fossil fuels are also generally more damaging to the environment than renewables. Multi-objective decision making approaches explicitly consider the *trade-offs* between conflicting objectives in such problems (e.g. cost vs. environment impact), allowing an appropriate balance between objectives to be achieved in accordance with system designer / user preferences. Algorithmic approaches to solving MODeM problems span many interrelated fields, such as reinforcement learning (RL) [10], planning [10], multi-agent systems [21], game theory [21] and utility theory, to name a few. We argue that to ensure trustworthiness in AI systems, it is necessary to adopt explicitly multi-objective approaches to autonomous decision making.

This paper therefore explores the relation between multi-objective decision making and trustworthy AI. More specifically, we address the following questions:

(1) Why is it necessary to treat trustworthiness as a multi-objective problem?
(2) How should rewards for a trustworthy agent be specified?
(3) How should a trustworthy agent reason over multiple objectives?
(4) Where should preferences for a trustworthy agent come from?

2 TRUSTWORTHINESS AS A MULTI-OBJECTIVE PROBLEM

According to the High-Level Expert Group on AI appointed by the European Commission, Trustworthy AI has three main aspects, which should be met throughout the system's entire life cycle [13]: it should be *lawful*, *ethical*, and *robust*. The four ethical principles are *respect for human autonomy*, *prevention of harm*, *fairness* and *explicability*. Based on these, the High-Level Group defined seven key requirements for Trustworthy AI: (1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination and fairness, (6) environmental and societal well-being and (7) accountability.

These requirements are in essence incommensurable objectives, and making decisions carefully under these requirements is an inherently multi-objective process. This is especially important since in practice there are often tensions between these different objectives, and the objectives of the system designer and/or end user. For example, in some cases one might have to make a trade-off between accuracy and transparency where this choice might not be the same in every particular instance.

To measure an agent's performance and trustworthiness in a multi-objective manner after it has performed a task, one could arrange the values of each quantity that is of interest into a *value vector*, $\mathbf{v}$. For example, in an electricity generation scenario, $\mathbf{v}$ might contain the objectives [`fuel_cost`, `emissions`, `noise`, `safety`]. As well as primary objectives (such as `fuel_cost`), the measured values for the trustworthiness requirements described earlier (such

*All authors contributed equally to the paper

Proc. of the 1st Multi-Objective Decision Making Workshop (MODeM 2021), Hayes, Mannion, Vamplew (eds.), July 14-16, 2021, Online, http://modem2021.cs.nuigalway.ie. 2021.

as `safety`) can also be explicitly accounted for, allowing an agent to optimise for and be evaluated on the desired trustworthiness requirements in addition to its primary goals.

Once a multi-objective perspective has been adopted, an important question that must be answered is: how to select the preferred multi-objective outcome? When evaluating different candidate agents to perform a task, the system designer can compare the multi-objective value vectors for each agent; if one agent is superior across all objectives then that agent is the clear winner. However, if e.g. one agent has a low `fuel_cost` but high `noise`, while a second agent has a high `fuel_cost` but low `noise`, their value vectors are *non-dominated* with respect to each other, and an appropriate value vector must be chosen according to the preferences of the system designer or end user. According to the ethical guidelines, it is our individual and collective responsibility as a society to work towards ensuring that all three components help to achieve Trustworthy AI. Having this explicit choice, helps us take the trustworthiness into account.

One option to allow selections between non-dominated vectors is to define a *utility function* u that takes a value vector as input, and returns a scalar value:

$$v_u = u(\mathbf{v}) \tag{1}$$

Given an appropriate utility function, a ranking over possible agents to solve a task could be established, based on the utility of their respective value vectors. More generally, a utility function can be paramaterised by both a value vector $\mathbf{v}$, and another vector $\mathbf{w}$ that specifies the preferences of the system designer. A canonical and widely-used example of a utility function is *linear scalarisation*:

$$v_u = u(\mathbf{v}, \mathbf{w}) = \mathbf{w}^\top \mathbf{v} \tag{2}$$

By explicitly representing the different aspects as a value vector, the problem and the solution is made more understandable by humans which will directly improve the transparency of the approach, the accountability, and the potential for human oversight.

3 REWARD SPECIFICATION

In typical agent-based decision making systems, desired behaviours are achieved by optimising a single clearly defined objective function that is specified by the system designer. However, the design of these objective functions are often arbitrary, and may be counter-intuitive. For example, the performance of agents based on approaches such as reinforcement learning or planning can be evaluated in terms of a designer-specified *reward function*; such agents then aim to learn or plan a *policy* to select actions that maximise the sum of rewards they receive. The general idea is to favourably reward desirable actions and penalise undesirable ones. However, this assumes the foreknowledge of desirable and undesirable actions a priori, which may not be feasible in many practical scenarios. Another fundamental issue when designing reward maximising agents is the lack of principled guidelines for the specification of reward functions. As noted recently by Knox et al. [17], the main textbooks, as well as the wider research community, devote very little attention to the topic of reward design.

In certain types of problems, the design of reward functions to evaluate the performance of an autonomous system is relatively straightforward. For example, for a system conducting high-frequency trading of stocks or other asset classes, an obvious objective is the total profit achieved in dollars; this naturally leads to rewards that are based on the profit/loss for each specific action (trade). When performance can easily be evaluated in quantitative terms (e.g., cost, emissions, travel time etc.), there exist natural specifications for rewards that an agent can optimise for.

However, in many problems it is not straightforward to design reward functions. Consider for example an autonomous driving scenario; how should a collision between an autonomous vehicle and a piece of street furniture be rewarded? Should the reward for such an undesirable outcome be -10, -100, or -1,000,000 even? This is an example of an *arbitrarily defined* reward; the scale of the reward is chosen by the system designer, and is not well-justified. In such settings, the relative reward values assigned to different actions is generally also arbitrary, and the incorrect specification of rewards could have an adverse effect on the policy learned [3]. Worse still, such arbitrarily-defined reward functions often combine multiple different terms representing various desirable/undesirable behaviours into the same reward function, making it nearly impossible for the agent to reason about which aspects of its behaviour map to which desirable/undesirable reward term.

A potential solution for settings where there is no natural quantitative measure of performance could be to move towards event-based learning, where each event that occurs during an agent's interaction with its environment is simply associated with a binary reward (1 if event is `true` and 0 if event is `false`). Such an approach is better suited to situations where the underlying task may lack a natural representation of rewards (as is the case in the earlier example of assigning collision rewards for autonomous vehicles).

The binary reward structure of an event-based approach would obviate the need for such arbitrary reward assignments. Event-based binary rewards were recently [6] shown to be particularly effective in goal-based, real world tasks [30], where it is generally more feasible to assign rewards to specific desirable events, as opposed to every possible state-action pair. Some recent research has also applied event-based binary rewards in multi-objective settings [15, 20]. This approach to specifying rewards could potentially be further developed by decoupling the occurrence of events from how desirable or undesirable they may be. Doing so would allow for a more flexible multi-objective decision making framework, where decisions could be made using a utility function which combines the event-based reward components in accordance with the desirability of the event. This would allow the utility of the multi-objective returns of various actions to be compared. The utility function could in turn be determined separately, through an independent process. Hence, the overall approach here fundamentally demands the specification of what constitutes (a) an event and (b) the desirability of an event. We will discuss (a) here, and (b) later on in Section 5.

A number of approaches could be used to identify distinct event occurrences during learning. With feature-based representations, adaptive variants of k-means clustering [16] have been shown to reliably and autonomously identify distinct regions of the state-space. The key idea is to track the mean and variance of each feature during learning, and to use these quantities to identify significant

deviations from historically observed features, which could be indicate the occurrence of distinct, novel events. Classical concepts such as relative novelty [25] and state importance [4], previously used for sub-goal discovery and apprenticeship learning respectively, could also be employed for the identification of distinct events during learning. The common theme underlying these approaches is the identification of outlying states on the basis of specific state properties. As these properties need not be tied to the objective/reward function, and primarily depend on the state representation, it allows the problem of event identification to be decoupled from that of learning desirable behaviours, making it agnostic to the specific learning algorithms used.

4 TRUSTWORTHY REASONING OVER VECTOR REWARDS

Assuming the environment produces event-based binary reward vectors as described in the previous section, the question then becomes how to select actions based on those vector rewards. Conventional approaches to the creation of utility-maximising agents such as reinforcement learning have generally assumed that the desired behaviour can be encapsulated in the form of a single, scalar measure of utility (the reward signal in single-objective RL) [26]. When faced with a vector reward, these methods need to concatenate the different factors into a single scalar reward term prior to providing the reward to the agent. Most commonly this has been accomplished by simply summing the reward components associated with each factor, either with or without weighting [22]. However this linear scalarisation approach is difficult to tune to find an appropriate trade-off between the different factors, and risks finding an inferior or inappropriate solution [10, 29].

Non-linear scalarisation functions may increase the range of policies which can be found by the agent, and may be a more accurate expression of the desired characteristics of a trustworthy agent. For example, a trustworthy agent must perform in a manner which is within acceptable safety bounds – this can be expressed using threshold lexicographic ordering [7] to ensure that the actions performed keep the probability of an adverse event below an acceptable threshold [8]. This type of scalarisation would fit very naturally with event-driven rewards, where the components correspond to the discounted probability of events.

Other non-linear scalarisations may also be used to capture different aspects of trustworthiness. For example the concept of fairness can be expressed in terms of the Generalised Gini Welfare Index (GGWI) which compares the utility received by different stakeholders and selecting decisions which minimise the extent to which any individual is negatively impacted [24]. In the context of event-based rewards, this may require a two-step approach:

(1) a utility measure is derived for each stakeholder using a scalarisation function which corresponds to their preferences
(2) the GGWI is applied to these individual utility measures to determine the policy which best satisfies our concept of fairness.

However applying a non-linear scalarisation prior to passing the reward to the agent on each time-step may lead to incorrect behaviour from the agent, as the returns are no longer additive [22]. The alternative, and we would argue more appropriate, approach is to adopt an explicitly multi-objective method in which the agent is provided with an unscalarised reward vector. The agent itself then assumes responsibility for learning a suitable policy which takes into consideration all of the objectives. This also allows for the possibility of *multi-policy* learning, in which the agent learns a set of policies which produce different trade-offs between the objectives [1, 28].

A range of approaches to multi-objective decision-making have been proposed and evaluated in the literature, particularly in the area of multi-objective reinforcement learning (MORL). Any of these may potentially be applicable to the creation of trustworthy agents. For a summary of these algorithms we recommend [10, 14] and [22].

One issue that must be considered when using non-linear utility functions to make decisions is the choice of optimisation criterion: scalarised expected returns (SER) or expected scalarised returns (ESR) [22]. For linear utility functions both criteria are equivalent [10], however for non-linear utility functions the ESR and SER criteria can lead to significantly different behaviours [11, 23].

In settings where the safety of a decision is paramount (e.g. selecting medical treatments), ESR is the correct optimisation criterion, as the utility of the user would be derived from a single execution of the agent's policy. On the other hand, the SER criterion may be most appropriate for situations where fairness is important, as in this case the utility is derived from the expected value of objectives over multiple policy iterations, e.g. when a government makes a series of decisions on how to allocate funding to different sectors of society. Therefore, when taking a multi-objective approach to trustworthy decision making, choosing the correct optimisation criterion is crucial to ensuring that the desired behaviour is achieved.

Another point to take into consideration for ESR settings where safety is important, or where there is a high degree of uncertainty about the outcomes of actions, is that it may be beneficial to move beyond simple expected value approaches to reasoning about rewards, and instead consider probability distributions over possible reward values when making trustworthy decisions. Research into distributional multi-objective decision making is still at an early stage [9, 11], but could potentially allow agents to more easily avoid negative outcomes, or to make decisions that do not exceed a certain level of risk of negative outcomes that is deemed acceptable by the system designer.

A serious shortcoming of agents that reason using expected value approaches can be demonstrated with a simple example. Consider two actions a_1 and a_2. Each time a_1 is selected, a reward of 1.0 is returned. Each time action a_2 is selected, there is a 50% chance of getting a reward of 0.0 and 50% chance getting a reward of 2.0. Both actions have the same expected value of 1.0, however there is much greater uncertainty about the outcome of a_2. If safety and consistency of outcomes is desirable, it is clear that a trustworthy agent should always prefer action a_1. However, agents that select actions based on expected rewards will not be able to make this distinction during action selection, as the information about the probabilities of rewards occurring is lost when storing expected reward values only.

Reasoning about agent behaviours in the manner that we have outlined has a huge advantage over traditional single-objective approaches in terms of explicability, which includes explainability

and transparency. If interrogated about why a specific decision was made, a single-objective agent will simply report back that the action taken had a higher expected reward than all other actions, which is not very informative; traditionally designed rewards are often meaningless as we outlined in Section 3, so the relative expected reward values for actions contain very little useful information about why a certain action is selected. A MODeM agent using the ideas that we have outlined would have a much greater degree of explicability; the agent could report the expected (or distributional) reward values for each objective of interest, and then demonstrate that the action leading to the most preferred compromise over objectives was chosen according to the user's utility function. Such an approach gives a clear link between the decisions made by an agent during deployment, and the preferences specified by the system designer over the range of possible agent behaviours [5].

Of course in this section we have yet to account for how the details of utility functions are derived – for example, how are thresholds determined for a safety-aware agent, and where do the individual stakeholder utility functions come from in our fairness example? The following section will address this issue.

5 PREFERENCES OVER OBJECTIVES FOR TRUSTWORTHY AGENTS

The problem of learning trustworthy preferences for different events depends on several subjective factors, which one may only be able to capture via one of more forms of human input. Indirect ways to infer human values typically use expert demonstrations and imitation learning [12]. However, such approaches are tedious and burdensome from a human perspective, and it imposes a requirement of human expertise for the task under consideration.

Preference elicitation may be a more direct and promising method to develop utility functions for trustworthy agents in the future. For example, Zintgraf et al. [31] successfully used pairwise comparison queries over outcomes and Gaussian processes to model the utility of domain experts for a traffic regulation problem with 11 objectives. Such methods could readily be adapted to model user utility with regard to trustworthy behaviours, such as fairness and safety.

Large-scale studies such as the MIT Moral Machine study [2] also demonstrate preference elicitation through the use of voting systems as a tool for ethical decision making for autonomous vehicles. Here, human participants were presented with, and made to choose between pairs of scenarios that differed from each other in terms of one or more morally-sensitive features. Based on the tens of millions of collected responses, the study constructed approximate ethical models to capture the collective preferences over all voters. Although such systems help build prior models of human preferences, they rely on carefully designed surveys targeted at solving specific problems, which may or may not translate well when applied to real world scenarios.

Frazier et al. [19] recently proposed a more organic approach for capturing human preferences, in which they utilised naturally occurring stories in comic strips as a training source, and extracted human societal norms encoded in them. The authors used natural language descriptors found in the comic strips to classify how well they align with the main character's behaviour. From the point of view of scalability, such an approach is superior, as it leverages existing data sources to extract societal preferences.

Considering that societal values and morals are context dependent, and tend to evolve over time (sometimes rapidly, as in the case of an emergency such as the COVID-19 pandemic), it is essential that learned preference models are flexible, and able to react to these changes. This type of flexibility was identified as an essential requirement of human-aligned AI by The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems [27, p25]. Such flexibility may also be required to remain consistent with contemporary legislative requirements. One way to achieve this could be through multi-policy methods mentioned in Section 4, in which multiple possible policies that vary in their trade-offs for different objectives are learned, and stored as a behaviourally diverse policy set. Whenever the requirements change (e.g., due to a change in legislation), an appropriate policy could be selected through efficient search mechanisms without having to train the agents from scratch [18]. Such frameworks could present robust and adaptive ways of handling a diverse range of ethical requirements as dictated by contemporary societal rules and norms.

6 CONCLUSION AND FUTURE WORK

In this paper, we have outlined how we believe the problem of creating trustworthy autonomous systems can be successfully tackled using MODeM principles. We hope that this work will serve as an inspiration to other researchers, and will help to drive the adoption of multi-objective approaches across a wide variety of applications. To conclude, we present a list of interesting future research directions that could be explored when applying MODeM principles to design trustworthy AI:

- The development of principled guidelines for reward design.
- Methods to classify and automatically identify events of interest via binary event reward encoding.
- Comprehensive studies on the implications of optimisation criteria (ESR vs. SER) for the trustworthiness of AI systems.
- New distributional reward algorithms to enable deployment of trustworthy AI in risk-aware and safety-critical settings.
- Development of MODeM approaches to explicability, along with benchmarking of such approaches against the current state-of-the-art.
- Further development of methods that allow system designers to specify utility functions that enable trustworthy behaviour.

REFERENCES

[1] Axel Abels, Diederik Roijers, Tom Lenaerts, Ann Nowé, and Denis Steckelmacher. 2019. Dynamic weights in multi-objective deep reinforcement learning. In *International Conference on Machine Learning*. PMLR, 11–20.

[2] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The moral machine experiment. *Nature* 563, 7729 (2018), 59–64.

[3] Nick Bostrom. 2014. *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press.

[4] Jeffery Allen Clouse. 1996. *On integrating apprentice learning and reinforcement learning.* University of Massachusetts Amherst.

[5] Francisco Cruz, Richard Dazeley, and Peter Vamplew. 2019. Memory-based explainable reinforcement learning. In *Australasian Joint Conference on Artificial Intelligence*. Springer, 66–77.

[6] Justin Fu, Avi Singh, Dibya Ghosh, Larry Yang, and Sergey Levine. 2018. Variational inverse control with events: A general framework for data-driven reward definition. *arXiv preprint arXiv:1805.11686* (2018).

[7] Zoltán Gábor, Zsolt Kalmár, and Csaba Szepesvári. 1998. Multi-criteria reinforcement learning.. In *ICML*, Vol. 98. Citeseer, 197–205.

[8] Peter Geibel and Fritz Wysotzki. 2005. Risk-sensitive reinforcement learning applied to control under constraints. *Journal of Artificial Intelligence Research* 24 (2005), 81–108.

[9] Conor Francis Hayes, Mathieu Reymond, Diederik Marijn Roijers, Enda Howley, and Patrick Mannion. 2021. Risk Aware and Multi-Objective Decision Making with Distributional Monte Carlo Tree Search. In *Proceedings of the Adaptive and Learning Agents Workshop (at AAMAS 2021)*. https://ala2021.vub.ac.be/papers/ALA2021_paper_33.pdf

[10] Conor Francis 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 Marijn Roijers. 2021. A Practical Guide to Multi-Objective Reinforcement Learning and Planning. *arXiv preprint arXiv:2103.09568* (2021). https://arxiv.org/abs/2103.09568

[11] Conor Francis Hayes, Timothy Verstraeten, Diederik Marijn Roijers, Enda Howley, and Patrick Mannion. 2021. Expected Scalarised Returns Dominance: A New Solution Concept for Multi-Objective Decision Making. *arXiv preprint arXiv:2106.01048* (May 2021). https://arxiv.org/abs/2106.01048

[12] Todd Hester, Matej Vecerik, Olivier Pietquin, Marc Lanctot, Tom Schaul, Bilal Piot, Dan Horgan, John Quan, Andrew Sendonaris, Ian Osband, et al. 2018. Deep q-learning from demonstrations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[13] High-Level Expert Group on AI. 2019. *Ethics guidelines for trustworthy AI.* Report. European Commission, Brussels. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

[14] Yaochu Jin. 2006. *Multi-objective machine learning*. Vol. 16. Springer Science & Business Media.

[15] Johan Källström and Fredrik Heintz. 2019. Tunable Dynamics in Agent-Based Simulation using Multi-Objective Reinforcement Learning. In *Proceedings of the Adaptive and Learning Agents Workshop (at AAMAS 2019)*.

[16] Thommen George Karimpanal and Erik Wilhelm. 2017. Identification and off-policy learning of multiple objectives using adaptive clustering. *Neurocomputing* 263 (2017), 39–47.

[17] W Bradley Knox, Alessandro Allievi, Holger Banzhaf, Felix Schmitt, and Peter Stone. 2021. Reward (Mis) design for Autonomous Driving. *arXiv preprint arXiv:2104.13906* (2021).

[18] Ayaka Kume, Eiichi Matsumoto, Kuniyuki Takahashi, Wilson Ko, and Jethro Tan. 2017. Map-based multi-policy reinforcement learning: enhancing adaptability of robots by deep reinforcement learning. *arXiv preprint arXiv:1710.06117* (2017).

[19] Md Sultan Al Nahian, Spencer Frazier, Mark Riedl, and Brent Harrison. 2020. Learning norms from stories: A prior for value aligned agents. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 124–130.

[20] David O'Callaghan and Patrick Mannion. 2021. Tunable Behaviours in Sequential Social Dilemmas using Multi-Objective Reinforcement Learning. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021)*. http://www.ifaamas.org/Proceedings/aamas2021/pdfs/p1610.pdf

[21] 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, 1 (2020), 10.

[22] 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.

[23] 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

[24] Umer Siddique, Paul Weng, and Matthieu Zimmer. 2020. Learning Fair Policies in Multi-Objective (Deep) Reinforcement Learning with Average and Discounted Rewards. In *International Conference on Machine Learning*. PMLR, 8905–8915.

[25] Özgür Şimşek and Andrew G Barto. 2004. Using relative novelty to identify useful temporal abstractions in reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*. 95.

[26] Rich Sutton. [n.d.]. *The reward hypothesis.* http://incompleteideas.net/rlai.cs.ualberta.ca/RLAI/rewardhypothesis.html

[27] The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems. 2016. Ethically Aligned Design: A Vision for Prioritizing Wellbeing With Artificial Intelligence and Autonomous Systems.

[28] Peter Vamplew, Richard Dazeley, Adam Berry, Rustam Issabekov, and Evan Dekker. 2011. Empirical evaluation methods for multiobjective reinforcement learning algorithms. *Machine learning* 84, 1 (2011), 51–80.

[29] 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.

[30] Henry Zhu, Justin Yu, Abhishek Gupta, Dhruv Shah, Kristian Hartikainen, Avi Singh, Vikash Kumar, and Sergey Levine. 2019. The Ingredients of Real World Robotic Reinforcement Learning. In *International Conference on Learning Representations*.

[31] Luisa M Zintgraf, Diederik M Roijers, Sjoerd Linders, Catholijn M Jonker, and Ann Nowé. 2018. Ordered preference elicitation strategies for supporting multi-objective decision making. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1477–1485.