# Environment-Centered Design of Ethical Environments

Doctoral Consortium

Arnau Mayoral-Macau
Artificial Intelligence
Research Institute (IIIA-CSIC)
Bellaterra, Spain
arnau.mayoral@iiia.csic.es

## ABSTRACT

With the rapid introduction of autonomous agents into everyday tasks, concerns about agent alignment to human moral norms are becoming increasingly prominent. The widespread adoption of reinforcement learning (RL) in autonomous decision-making has intensified the challenge of ensuring that these algorithms align agents' behaviour with moral and ethical values. While most common approaches to value alignment focus on the learning algorithms agents use, a recently introduced algorithm called ethical embedding shifts the focus to designing ethical environments rather than modifying agents' learning algorithms. This transition from an agent-centred view to an environment-centred perspective opens new opportunities for developing safe and trustworthy AI agents. This project aims to advance this line of research by exploring environment design as a means to create non-manipulable learning environments, where the environment itself guides agents toward ethical behaviour, regardless of the learning algorithm employed. Furthermore, as a novel contribution to previous work, we integrate this approach with state-of-the-art deep reinforcement learning algorithms, enabling the application of these techniques in realistic environments suitable for training agents that operate in the real world.

## KEYWORDS

Ethical values; Ethical Environment Design; Multi-Agent Reinforcement Learning; Multi-Objective Reinforcement Learning

**ACM Reference Format:**
Arnau Mayoral-Macau. 2025. Environment-Centered Design of Ethical Environments: Doctoral Consortium. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

## 1 RESEARCH STATEMENT

With the growing integration of autonomous agents into everyday tasks [1, 8, 18, 20], the associated risks have become increasingly evident. Consequently, international efforts, such as the Artificial Intelligence (AI) Act [4], aim to ensure that these systems operate in alignment with human values [5, 13, 16].



*Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA.* © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

As reinforcement learning (RL) is becoming increasingly adopted to train autonomous agents for single-agent and multi-agent scenarios, RL literature has expanded its focus to safety and ethical alignment.

While safety AI focuses on guaranteeing that no deployed agent causes harm [6, 7], the field of Machine ethics [12, 22] goes further by also ensuring agents' behaviour includes proactivity in performing good (praiseworthy) actions. The tools to steer agents' behaviour in RL are the rewards, either positive or negative, given to them upon interacting with the environment. Thus, it is common to see how safety and moral value alignment is instilled by means of extrinsic, manually-tuned rewards. These rewards can be framed as penalties and constraints in safety RL or as moral incentives and punishments in the context of machine ethics. Correctly considering these separate objectives will lead to *aligned* policies: policies that achieve the primary goal while always respecting the extra considerations encoded in the extra objective.

As a result, this problem can be addressed as a multi-objective one, involving a trade-off between achieving the primary goal and an additional alignment objective. Consequently, multi-objective RL (MORL) algorithms have been used to learn policies that optimise both of the objectives [2, 17, 19, 23]. However, MORL algorithms often require a prioritisation of the objectives. That leaves alignment to the owner of the learning algorithm, who will have to choose the right prioritisation to achieve aligned policies.

These approaches are inherently agent-centric, as alignment arises from the choice of the learning algorithm, which is determined by the agent's owner at training time. Since alignment should not be left to individual choice, a novel approach has been recently introduced: the ethical embedding (EE) algorithm [10, 11]. This algorithm shifts the responsibility for alignment from the learning algorithm to the environment designer. Using EE, a designer can integrate any alignment function—ethical, safety-related, or any other, encoded as a reward function—alongside the agent's primary objective. Thus, EE creates an embedded reward function, where the alignment desired by the environment designer is imposed.

When learning from the embedded reward function, all agents will find it optimal to be aligned. This is the strength of the approach: with only one objective, any learning algorithm optimising it will produce aligned policies. Through EE, environment designers can create non-manipulable environments in which trained agents behave exactly as intended by the designer. In other words, any third party using the environment to train their agents will consistently obtain aligned policies.

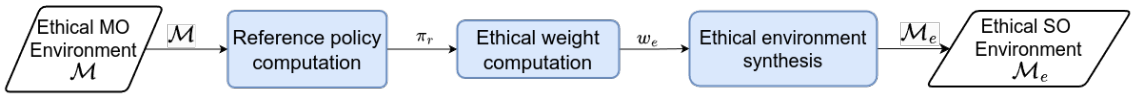


**Figure 1: Approximate Embedding algorithm**

## 2 PRELIMINARY RESULTS

The EE algorithm has been successfully applied to small environments in both single-agent and multi-agent scenarios [10, 11]. Nevertheless, the goal of alignment inherently requires operating in the real world, where interactions with humans emphasize the need for agents aligned with human moral values and emotions. Consequently, it is essential for EE to function effectively in realistic environments. Yet, since EE relies on methods with limited scalability, it struggles to handle complex environments characterised by large state spaces and imperfect information, such as those found in the real world.

In this research project, we plan to create an approximate embedding algorithm, that follows the lines of the original EE, but further investigates making it work in realistic environments. The recent advances in deep reinforcement learning (DRL) for single-agent and multi-agent scenarios, along with the environment-centred perspective, open up new opportunities to advance the state-of-the-art regarding trustworthy AI and value alignment.

We now describe current work and results achieved throughout the first year this research project.

***MORL formalisation.*** The initial multi-objective problem of alignment can be easily formalised as a multi-objective Markov Decision Process (MOMDP) [14] for single-agent environments. In a MOMDP, $m$ different objectives are encoded as $m$ different reward functions $R_1, \cdots, R_m$. Thus, for each action of the agent, the environment returns a reward vector $\vec{r} = (r_1, \cdots, r_m)$. Following our alignment problem, we can define MOMDP with two objectives for the agent: a primary goal within the environment $R_0$, and a complementary objective $R_a$ that assesses the alignment of the agent's actions to a set of considerations.

***Ethical Embedding [10, 11].*** The original EE approach, henceforth *optimal embedding*, utilises RL algorithms with convergence properties, such as Value Iteration or Q-Learning, to determine an alignment weight $w_a$ that integrates both objectives into a single one through a linear scalarisation: $R = R_0 + w_a \cdot R_a$. This weight, $w_a$, is specifically calculated to ensure that agents adhere to the alignment function $R_a$ while optimising the primary objective $R_0$. Since $w_a$ controls how great the alignment incentives are, it is assumed that for a large enough $w_a$, the alignment objective will be totally prioritised. Consequently, policies trained to maximise $R$ will inherently be aligned policies.

***Approximate Embedding.*** The *approximate* embedding (AE) represents the main contribution of the research proposal. This algorithm reformulates optimal embedding to operate effectively in realistic environments with large state spaces and partial observability. Due to scalability limitations, these changes render methods like Value Iteration and Q-Learning impractical. Instead, the focus shifts to DRL, which leverages the generalisation power of deep learning and has demonstrated strong performance in complex, realistic environments [3, 15, 21].

Our research is now focused on the automated design of ethically aligned multi-agent environments, building on the work of Rodriguez-Soto et al. [11]. Their study applies the optimal embedding to the Ethical Gathering Game (EGG), where multiple agents must gather resources from a grid map to survive. This environment, inspired by the gathering environment of [9], was modified by Rodriguez-Soto et al. to shift the focus from resource depletion to creating an unequal setting where efficient agents assist inefficient agents in achieving survival. However, due to the limited scalability of the optimal embedding, they had to reduce the original environment to a simple $3 \times 4$ grid map with two agents.

Some experiments were carried out to demonstrate how the approximate embedding could achieve the same result as the optimal embedding in the reduced EGG. Then, after the first year of this research project, we have some promising results on the EGG with five agents, partial observability, and the original map size of [9].

As for now, our new algorithm called *approximate embedding* (AE) computes the alignment weight $w_a$ needed to build single-objective environments aligned to an ethical moral value encoded in $R_a$. The AE algorithm consists of three main steps, depicted in Figure 1. First, a reference policy $\pi_r$ is computed directly in the MO environment using an agent-centric algorithm. This policy $\pi_r$ corresponds to a policy where all agents totally prioritise (non-linearly) the alignment objective over the individual objective. Second, we search for an *approximately minimal* alignment weight $w_a$ capable of creating a scalarised environment that incentivises agents to learn a policy as ethical as the reference policy $\pi_r$. Finally, once the search finishes and we have the final $w_a$, the ethically-aligned environment $\mathcal{M}_e$ is returned as the final environment with a scalarised reward function of the form $R_0 + w_a \cdot R_a$.

## 3 CONCLUSIONS AND FUTURE WORK

To summarise, our research project aims to further investigate environment-centred algorithms for value-alignment and trustworthy AI, focusing on making these algorithms feasible for realistic environments and applicable to autonomous agents operating in the real world. Additional research directions we plan to investigate in the short term include optimising the algorithm's searching step with MORL techniques, such as Optimistic Linear Support, which can help find the necessary alignment weight faster and decrease the computational burden. As long-term goals, we aim to extend the approach to designing environments that align with multiple alignment functions and deploying aligned agents effectively in the field of robotics.

## REFERENCES


[1] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109* (2023).

[2] Krishnendu Chatterjee, Joost-Pieter Katoen, Stefanie Mohr, Maximilian Weininger, and Tobias Winkler. 2023. Stochastic games with lexicographic objectives. *Formal Methods in System Design* (March 2023). https://doi.org/10.1007/s10703-023-00411-4

[3] Christian Schroeder de Witt, Tarun Gupta, Denys Makoviichuk, Viktor Makoviychuk, Philip H. S. Torr, Mingfei Sun, and Shimon Whiteson. 2020. Is Independent Learning All You Need in the StarCraft Multi-Agent Challenge?

[4] European Comission. 2021. Artificial Intelligence Act. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021PC0206. Accessed: 2024-01-22.

[5] Iason Gabriel. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines* 30 (09 2020), 411–437. https://doi.org/10.1007/s11023-020-09539-2

[6] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. 2021. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916* (2021).

[7] José Hernández-Orallo, Fernando Martínez-Plumed, Shahar Avin, and Sean O. Heigeartaigh. 2019. Surveying Safety-relevant AI characteristics. In *AAAI workshop on artificial intelligence safety (SafeAI 2019)*. CEUR Workshop Proceedings, 1–9. https://riunet.upv.es/handle/10251/146561

[8] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 7873 (2021), 583–589.

[9] Joel Z. Leibo, Vinícius Flores Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. 2017. Multi-agent Reinforcement Learning in Sequential Social Dilemmas. *CoRR* abs/1702.03037 (2017). http://arxiv.org/abs/1702.03037

[10] Manel Rodriguez-Soto, Maite Lopez-Sanchez, and Juan A Rodriguez-Aguilar. 2021. Multi-Objective Reinforcement Learning for Designing Ethical Environments.. In *IJCAI*. 545–551.

[11] Manel Rodriguez-Soto, Maite Lopez-Sanchez, and Juan A Rodriguez-Aguilar. 2023. Multi-objective reinforcement learning for designing ethical multi-agent environments. *Neural Computing and Applications* (2023), 1–26.

[12] Francesca Rossi and Nicholas Mattei. 2019. Building Ethically Bounded AI. *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (07 2019), 9785–9789. https://doi.org/10.1609/aaai.v33i01.33019785

[13] Stuart Russell, Daniel Dewey, and Max Tegmark. 2015. Research priorities for robust and beneficial artificial intelligence. *Ai Magazine* 36, 4 (2015), 105–114.

[14] 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 (April 2020). https://doi.org/10.1007/s10458-019-09433-x

[15] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms.

[16] Nate Soares and Benja Fallenstein. 2014. Aligning superintelligence with human interests: A technical research agenda. *Machine Intelligence Research Institute (MIRI) technical report* 8 (2014).

[17] Alperen Tercan and Vinayak S. Prabhu. 2024. Thresholded Lexicographic Ordered Multiobjective Reinforcement Learning. http://arxiv.org/abs/2408.13493 arXiv:2408.13493 [cs].

[18] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575, 7782 (2019), 350–354.

[19] Kyle Wray, Shlomo Zilberstein, and Abdel-Illah Mouaddib. 2015. Multi-objective MDPs with conditional lexicographic reward preferences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 29. https://ojs.aaai.org/index.php/AAAI/article/view/9647 Issue: 1.

[20] Peter R Wurman, Samuel Barrett, Kenta Kawamoto, James MacGlashan, Kaushik Subramanian, Thomas J Walsh, Roberto Capobianco, Alisa Devlic, Franziska Eckert, Florian Fuchs, et al. 2022. Outracing champion Gran Turismo drivers with deep reinforcement learning. *Nature* 602, 7896 (2022), 223–228.

[21] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. 2022. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in Neural Information Processing Systems* 35 (2022), 24611–24624.

[22] Han Yu, Zhiqi Shen, Chunyan Miao, Cyril Leung, Victor R. Lesser, and Qiang Yang. 2018. Building Ethics into Artificial Intelligence. In *IJCAI*. 5527–5533.

[23] Ziyan Zhao, Siyi Li, Shixin Liu, MengChu Zhou, Xingyang Li, and Xiaochun Yang. 2024. Lexicographic Dual-Objective Path Finding in Multi-Agent Systems. *IEEE Transactions on Automation Science and Engineering* (2024), 1–11. https://doi.org/10.1109/TASE.2024.3440169 Conference Name: IEEE Transactions on Automation Science and Engineering.