Cooperative Multi-Agent Alignment via Boolean Task Algebras and Team Morality Chains

Doctoral Consortium

Simon Rosen
University of the Witwatersrand
Johannesburg, South Africa
simon.rosen@wits.ac.za

ABSTRACT

Cooperative multi-agent reinforcement learning (MARL) offers a principled route to deploying teams of autonomous agents, but standard scalar-reward optimisation can produce misaligned behaviour in safety-critical settings. This PhD research studies *cooperative multi-agent alignment*, decomposed into *intention alignment* (faithful execution of specified tasks) and *value alignment* (strict adherence to priority-ordered normative constraints). For intention alignment, I develop agent-level compositional task specification via a cooperative extension of Boolean Task Algebras, paired with goal-oriented learning to support zero-shot generalisation across tasks. For value alignment, I build on MoralityGym and morality chains and develop a lexicographical reinforcement learning approach based on principled scalarisation to enforce team-level moral priorities. I outline how these components integrate by treating task reward as the lowest-priority objective within a Team Morality Chain, yielding cooperative policies that execute intended tasks subject to non-negotiable safety constraints.

KEYWORDS

Cooperative MARL, AI Alignment, Task Specifications, Lexicographic RL, Machine Ethics, Safe MARL

ACM Reference Format:
Simon Rosen. 2026. Cooperative Multi-Agent Alignment via Boolean Task Algebras and Team Morality Chains: Doctoral Consortium. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/SCKG1056

1 INTRODUCTION

Cooperative Multi-Agent Reinforcement Learning (MARL) offers a robust framework for training teams to coordinate under coupled dynamics [2]. As these systems are deployed in high-stakes domains such as warehouse automation and disaster responses, reliance on scalar reward maximisation can lead to undesirable and unsafe behaviours. In particular, agents optimising a dense scalar signal may gamble with safety constraints or exploit specification ambiguities to maximise returns - failure modes that are amplified by multi-agent interaction effects [5, 13]. This motivates the problem of *Cooperative Multi-Agent Alignment*: ensuring that a team's joint behaviour adheres to human values while effectively executing intended tasks.

In this PhD, I address this problem by decomposing alignment into two distinct yet coupled challenges: (1) *Intention Alignment*, ensuring agents achieve the specific tasks assigned to them; and (2) *Value Alignment*, ensuring the team respects normative constraints regardless of the task. I argue that these components admit different mathematical treatments. Intention specification is naturally compositional and goal-oriented, while value constraints are often hierarchical and must be treated as non-negotiable requirements on joint behaviour. Accordingly, I adopt an explicit, top-down specification approach, which supports clear articulation and verification of strict normative hierarchies, rather than attempting to infer them from preference data.

Consider a disaster-response team: robots must coordinate to extinguish fires or evacuate casualties (*Intention*), but must simultaneously adhere to strict norms such as "never harm a human" or "minimise collateral damage" (*Value*). Crucially, these values should not be treated as costs to be traded against mission reward; instead, they impose constraints on the joint policy that must take precedence over task completion.

This extended abstract outlines a framework for specifying and enforcing such behaviours. My prior work establishes foundations for both pillars: I developed a method for zero-shot cooperative task generalisation [9], and introduced the *MoralityGym* benchmark [10] to formalise hierarchical value constraints via morality chains. Specifically, the task-generalisation method enables cooperative agents to transfer goal-oriented knowledge across task instances without additional training [9], while MoralityGym provides environments in which agents must satisfy higher-priority moral norms before optimising lower-priority objectives [10]. Current work focuses on (i) extending Boolean Task Algebras to specify agent intentions in heterogeneous teams and (ii) developing a principled lexicographical RL scalarisation technique to enforce team-level norms. The final phase of the PhD will integrate these components, using the Boolean algebra to construct the task-reward signal that appears as the lowest-priority objective within a Team Morality Chain.

2 ENSURING INTENTION ALIGNMENT VIA AGENT-LEVEL BOOLEAN TASK SPECIFICATIONS

A central bottleneck in intention alignment is task specification. The standard approach encodes tasks as reward functions, often supported by reward shaping [4]. Small misspecifications can induce

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/SCKG1056

unintended optimisation and "specification gaming" behaviours [12]. Formal alternatives such as temporal logic specifications and automata-based reward representations can improve interpretability and verification, but they often introduce additional structure that complicates learning and scaling in multi-agent settings [6].

My approach specifies tasks compositionally using a Boolean task algebra. In the single-agent setting, a task is defined via desirable and undesirable terminal outcomes (goal sets), and new tasks are constructed using Boolean operators (AND/OR/NOT) over base tasks [8]. This provides an interpretable and modular specification language, where complex instructions are built from simple primitives rather than tuned via dense reward engineering.

Contribution (in progress): I extend this idea to cooperative MARL by defining agent-level task algebras, where each agent's task is expressed as a Boolean composition of terminal goal conditions. The goal is to preserve interpretability ("what each agent is trying to achieve") while supporting heterogeneous teams. The learning component builds on my prior work on goal-oriented multi-task cooperative MARL [9], which learns goal-conditioned cooperative representations that support zero-shot inference across new tasks sampled from a compositional task space. This aligns with goal-conditioned RL foundations such as universal value function approximators (UVFAs) [11]. In particular, this method demonstrates that cooperative policies can generalise across task instances by reusing learned goal-oriented structure [9].

Modelling assumption (this extended abstract): I adopt centralized training, decentralized execution (CTDE). CTDE is standard in cooperative MARL because it supports efficient training while producing policies that run without centralized control at deployment [2]. Extending my prior goal-oriented framework to CTDE, while retaining task-algebra compositionality, is part of ongoing work.

3 VALUE ALIGNMENT VIA TEAM-LEVEL MORALITY CHAINS AND LEXICOGRAPHICAL OPTIMISATION

Even with correct task specification, cooperative agents may pursue goals in ways that violate safety and human values. Safe RL commonly represents constraints via constrained Markov decision processes and often solves them using Lagrangian/primal–dual optimisation and related methods [1, 3]. However, such methods typically encode constraints through scalar penalties or thresholds, and do not directly represent priority-ordered moral norms that should not be traded off for performance.

My value-alignment work builds on morality chains: ordered hierarchies of moral norms that define a lexicographic objective [10, 12]. **Published contribution:** Morality chains and MoralityGym provide a benchmark suite of environments in which an agent must satisfy high-priority norms before optimising lower-priority objectives, enabling systematic specification and evaluation of hierarchical moral alignment. In particular, MoralityGym demonstrates that satisfying strict moral hierarchies can be challenging for standard reinforcement learning methods, even when task objectives are straightforward [10].

Planned extension: I generalise morality chains to Team Morality Chains, in which each norm constrains the joint behaviour of a heterogeneous cooperative team. Team performance is evaluated via social welfare (the sum of agent returns) [2], while the morality chain imposes constraints over the joint policy independently of the agents' individual tasks. This extension motivates a multi-agent extension of MoralityGym to evaluate the interaction between coordination pressure and moral constraints.

Contribution (in progress): I develop a lexicographical RL approach based on principled scalarisation. The aim is to derive reward-weight bounds that are sufficient to induce lexicographically optimal policies over deterministic policy classes, drawing on discrete lexicographic optimisation and MORL foundations [7]. This provides a route to enforcing strict priority ordering while remaining compatible with standard reinforcement learning objectives. I then plan to learn suitable weights via an approach analogous in spirit to Lagrangian methods, but targeted at strict lexicographic satisfaction rather than thresholded trade-offs.

4 UNIFYING INTENTION AND VALUE ALIGNMENT

The unified framework treats tasks as the low-level optimisation target and morality chains as the high-level constraint structure. Agent-level Boolean task specifications define the task-reward components ("what each agent should accomplish"), while a Team Morality Chain defines a lexicographic ordering over normative objectives that constrain the joint policy ("what must not be traded off"). Formally, the combined objective is to satisfy the morality chain in priority order and, subject to those constraints, maximise cooperative task completion.

This composition matches the structure of my scalarisation-based lexicographical RL approach: the lexicographic solver produces a sequence (or adaptation) of scalar reward signals, where the lowest-priority term corresponds to task performance. In the unified setting, that lowest-priority objective is instantiated by the goal-conditioned task reward induced by Boolean-composed agent tasks and cooperative goal inference. This yields a clear division of roles: Boolean task algebra provides an explicit and inspectable intention specification, while lexicographic optimisation enforces team-level moral priorities over the resulting cooperative behaviour.

5 CONCLUSION AND FUTURE WORK

This PhD targets cooperative multi-agent alignment by developing (i) interpretable intention specification using Boolean task algebras and (ii) team-level value specification and enforcement using morality chains solved through lexicographical RL. Published results establish the two foundations: a goal-oriented cooperative task generalisation method [9] and the MoralityGym benchmark introducing morality chains [10]. Current work develops the Boolean-algebra CTDE MARL extension and the lexicographical scalarisation method. Remaining steps are to extend morality chains and lexicographical RL to cooperative teams, develop a multi-agent MoralityGym-style evaluation environment, and integrate the full intention and value alignment pipeline into a single framework for aligned cooperative MAS.

REFERENCES

[1] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. 2017. Constrained policy optimization. In *International conference on machine learning*. PMLR, 22–31.

[2] Stefano V Albrecht, Filippos Christianos, and Lukas Schäfer. 2024. *Multi-agent reinforcement learning: Foundations and modern approaches*. MIT Press.

[3] Eitan Altman. 1999. *Constrained Markov decision processes*. Routledge.

[4] Takumi Aotani, Taisuke Kobayashi, and Kenji Sugimoto. 2021. Bottom-up multi-agent reinforcement learning by reward shaping for cooperative-competitive tasks. *Applied Intelligence* 51, 7 (2021), 4434–4452.

[5] Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, Wolfram Barfuss, Jakob Foerster, Tomáš Gavenčiak, et al. 2025. Multi-agent risks from advanced ai. *arXiv preprint arXiv:2502.14143* (2025).

[6] Borja G León and Francesco Belardinelli. 2020. Extended markov games to learn multiple tasks in multi-agent reinforcement learning. *arXiv preprint arXiv:2002.06000* (2020).

[7] Kaisa Miettinen. 1999. *Nonlinear multiobjective optimization*. Vol. 12. Springer Science & Business Media.

[8] Geraud Nangue Tasse, Steven James, and Benjamin Rosman. 2020. A boolean task algebra for reinforcement learning. *Advances in Neural Information Processing Systems* 33 (2020), 9497–9507.

[9] Simon Rosen, Abdel Mfougouon Njupoun, Geraud Nangue Tasse, Steven James, and Benjamin Rosman. [n.d.]. Optimal Task Generalisation in Multi-Agent Reinforcement Learning. In *Coordination and Cooperation for Multi-Agent Reinforcement Learning Methods Workshop*.

[10] Simon Rosen, Siddarth Singh, Ebenezer Gelo, Helen Sarah Robertson, Ibrahim Suder, Victoria Williams, Benjamin Rosman, Geraud Nangue Tasse, and Steven James. 2026. MoralityGym: A Benchmark for Evaluating Hierarchical Moral Alignment in Sequential Decision-Making Agents. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. Accepted for publication.

[11] Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. 2015. Universal value function approximators. In *International conference on machine learning*. PMLR, 1312–1320.

[12] Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. 2022. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems* 35 (2022), 9460–9471.

[13] Peter R Wurman, Raffaello D'Andrea, and Mick Mountz. 2008. Coordinating hundreds of cooperative, autonomous vehicles in warehouses. *AI magazine* 29, 1 (2008), 9–9.