Enhancing AI Moral Value Representation through Argumentation Frameworks with Gradual Semantics

Master's Thesis

Edgar Demeude[1]

Supervised by:

Rémy Chaput[2], Guillaume Muller[3], Maxime Morge[1], and Bruno Yun[1]

[1] Université Lyon 1, LIRIS, UMR 5205, Villeurbanne, France
[2] CPE Lyon, CNRS, INSA Lyon, Université Lyon 1, Université Lumière Lyon 2, École Centrale de Lyon, LIRIS, UMR 5205, Villeurbanne, France
[3] Ecole Nationale Supérieure des Mines, MA/G2I, F-42023, Saint-Étienne, France

June 2026

Abstract. Aligning the behavior of artificial agents with human values is a fundamental societal challenge for autonomous systems, particularly within multi-agent environments. Traditional approaches rely on opaque scalar rewards that are easily subverted by reward hacking, while rule-based symbolic methods are often too rigid to handle complex trade-offs. To bridge this gap, this report introduces a novel neuro-symbolic architecture that integrates Multi-Agent Reinforcement Learning (MARL) with a Weighted Bipolar Argumentation Framework (WBAF). By modeling ethical dilemmas as structured debates where values and environmental facts attack or support one another, our framework computes continuous acceptability degrees to guide agent decisions. Evaluated on an "Ethical Gardeners" simulation requiring agents to balance profit, ecology, and social equality, our approach achieves comparable or superior multi-objective performance relative to standard MARL baselines. Crucially, it delivers these competitive results while providing behavioral transparency and an interpretable reward design.

Keywords: Machine ethics; Argumentation; Reinforcement learning; Hybrid neuro-symbolic learning; Multi-agents systems; Moral alignment

Table of Contents

1 Introduction

The integration of autonomous artificial intelligence agents into society has accelerated the need for systems that not only perform efficiently but also adhere to human moral guidelines. This requirement, studied within the domain of Computational Machine Ethics, seeks to solve the value alignment problem [33]: the challenge of ensuring that an AI's operational objectives do not inadvertently lead to unethical or harmful societal outcomes. In standard Reinforcement Learning (RL) paradigms, agents learn optimal behaviors by maximizing scalar reward signals. However, in complex environments, agents often exploit flaws in their reward functions: they maximize their explicit scores while failing to respect underlying human norms. This problem is known as reward hacking [8].

To mitigate these risks, recent literature has explored hybrid neuro-symbolic architectures that combine the adaptability of RL with the logical transparency of symbolic reasoning, such as Argumentation Theory. Yet, existing argumentation-based RL frameworks predominantly rely on classical, extension-based semantics. In complex normative scenarios, these binary models evaluate actions as either entirely accepted or completely rejected, failing to capture the continuous, nuanced nature of real-world moral dilemmas where ethical compliance exists on a spectrum.

Problem statement and objectives The core problem addressed in this report is the lack of explicit, continuous, and auditable mechanisms to represent human moral values within reinforcement learning loops. While standard reward functions are mathematically opaque and prone to specification hacking, existing symbolic alternatives rely on rigid, binary constraints. Consequently, autonomous agents lack the structural capacity to manipulate, integrate, and respect the degrees of compliance inherent to real-world ethical dilemmas.

To address this limitation, the primary objective of this research is to develop a methodology that translates multi-dimensional normative values into continuous learning signals. Conceptually, this objective aims to shift the burden of reward engineering away from software developers and hand it over to domain experts, ethicists and citizens. We seek to provide them with an interpretable, symbolic interface capable of dynamically evaluating an agent's policy without sacrificing the mathematical scalability required by modern learning algorithms.

Contributions To fulfill these objectives, this master's thesis delivers a cohesive scientific contribution spanning both theoretical framework design and empirical validation.

Firstly, we introduce a novel neuro-symbolic architecture that natively associates decentralized Multi-Agent Reinforcement Learning (MARL) with a Weighted Bipolar Argumentation Framework (WBAF). As a core theoretical enhancement to the state of the art, we extend the classical DF-QuAD gradual semantics to dynamically process context-dependent initial weights derived from environ-

mental state transitions, successfully mapping symbolic normative debates into continuous, dense morality scores.

Secondly, we leverage this theoretical foundation to provide an empirical validation of the proposed framework. We demonstrate that our argumentation-driven reward design successfully guides decentralized agents toward ethically aligned behaviors without sacrificing overall environmental viability or economic throughput when compared to standard multi-objective optimization baselines. Furthermore, the experiments show that our framework facilitates a slightly faster initial policy convergence while natively providing human-readable topological feedback to explain reward allocation.

The remainder of this report is organized as follows. Section 2 establishes the theoretical scientific background spanning Machine Ethics, Abstract Argumentation Theory, and Reinforcement Learning. Section 3 reviews the current state of the art and positions our contribution. Section 4 details our proposed neuro-symbolic architecture and mathematical integration. Section 5 presents the empirical evaluation, experimental protocol, and results. Finally, Section 6 discusses the implications, limitations, and future research directions of our approach.

2 Scientific background

This section establishes the foundational theories and concepts that define the scientific background of this research. Our theoretical architecture rests upon three primary pillars: Machine Ethics, Argumentation Theory, and Reinforcement Learning. Understanding the interplay between these domains is essential before exploring current alignment methods. For instance, recognizing how a reward function strictly dictates an agent's learned behavior in Reinforcement Learning clarifies exactly why we integrate Argumentation Theory: to provide a structured, logical mechanism capable of evaluating and injecting Machine Ethics into that learning loop.

2.1 Machine ethics

Terminology. To establish a precise formal framework, it is first necessary to clarify the terminology. While *ethics* and *morality* share a common etymology and are often used interchangeably in many cultural contexts, this report adopts a specific operational distinction. For the purposes of our computational framework, we define *ethics* as the systemic guidelines dictating how agents should act, while *morality* denotes the internal rules of conduct formed through an individual's own experiences. Under these working definitions, the two concepts continuously interact, and an agent's localized "moral" policy might misalign with broader "ethical" standards. Furthermore, not all values possess ethical dimensions; practical, economic, or aesthetic values operate independently of moral duties. However, within the domain of moral values, psychological models such as Schwartz's Theory of Basic Human Values [35] demonstrate that despite

cultural variances, certain abstract motivations (e.g., *security*, *benevolence*, *universalism*) act as foundational drivers for human decision-making. Highlighting such theories is valuable because it establishes that cross-cultural, structural values exist, thereby providing a reasonable and realistic basis for attempting computational value alignment.

Computational machine ethics. As autonomous systems increasingly integrate critical decision-making loops, transposing human ethical constructs into artificial entities has emerged as a crucial challenge. Computational Machine Ethics (CME) is the interdisciplinary subfield of AI Ethics explicitly dedicated to implementing, evaluating, and ensuring moral behavior in cognitive machines. To conceptualize the depth of ethical integration within an agent, Moor [27] defines four distinct levels of ethical agency:

- **Ethical impact agents:** Systems devoid of internal ethical reasoning whose deployment merely causes ethical consequences (e.g., a simple automated sorting tool).
- **Implicit ethical agents:** Systems engineered to inherently avoid unethical outcomes through hard-coded safety constraints or optimization constraints, but without the ability to explicitly represent or reason about ethical concepts.
- **Explicit ethical agents:** Systems capable of explicitly representing ethical principles, analyzing scenarios, and computing mathematically or logically ethical decisions based on given situational data.
- **Full ethical agents:** Entities possessing human-like ethical capacities, including consciousness, free will, intentionality, and the robust capability to justify their moral assertions.

The scope of this research primarily targets the development of *explicit ethical agents*, focusing on how an agent can structurally manipulate and resolve conflicting normative values.

Paradigms. In modeling explicit ethical reasoning, CME typically borrows from three classical paradigms in normative ethics, which dictate how criteria for moral behavior are formulated [40]:

1. **Consequentialism:** The morality of an action is judged exclusively by its outcomes, with the ultimate goal of maximizing overall utility or minimizing suffering [36]. Utilitarianism is widely recognized as the quintessential example of this paradigm, positing that the best action is the one that maximizes net well-being. For example, a doctor lying to a terminal patient about their prognosis could be considered morally acceptable under a utilitarian framework if it successfully prevents psychological distress and improves the patient's remaining quality of life.
2. **Deontology:** An action's moral worth derives from its inherent adherence to duties, rules, or universal moral principles, completely independent of its

actual consequences. This approach is famously anchored in Kant's Categorical Imperative, which demands that moral rules be followed unconditionally as absolute duties, regardless of the situation [24]. In the same scenario, a deontological framework would strictly forbid the doctor from lying, asserting that honesty is an absolute duty. If exceptions were permitted, the very foundation of mutual trust would logically collapse.

3. **Virtue ethics:** A paradigm that shifts the evaluative focus toward the internal moral character and long-term dispositions of the agent, rather than evaluating isolated actions or rigid rules. Historically rooted in Aristotle's philosophical traditions, this approach emphasizes the cultivation of moral virtues and practical wisdom to achieve ethical excellence [10]. In our clinical scenario, a virtuous doctor would naturally aim to act with compassion, integrity, and wisdom, carefully determining how to convey the diagnosis in a truthful yet deeply sensitive manner that reflects excellent moral character. Ultimately, the high level of abstraction and context sensitivity inherent to this character-driven approach makes it significantly harder to formalize computationally compared to rule-based or purely outcome-based frameworks [37].

Implementation approaches. Translating these philosophical currents into computational architectures yields three main implementation strategies, as classically categorized by Wallach and Allen [40]:

- **Top-down approaches:** The creation of agents with pre-defined, symbolic moral rules or logical constraints derived directly from normative theories (e.g., Deontic Logic). While these logic-based architectures offer high transparency, auditability, and explainability, they struggle to scale or adapt when encountering novel, real-world ethical dilemmas where rules conflict.
- **Bottom-up approaches:** The reliance on inductive machine learning, where an agent infers implicit ethical patterns by observing human datasets or interacting with an environment via Reinforcement Learning (RL). While highly adaptive to complex environments, these empirical methods behave as "black boxes," severely lacking semantic transparency and formal guarantees of alignment.
- **Hybrid approaches:** The synthesis of the structural transparency of top-down rules with the adaptive scalability of bottom-up learning. A notable paradigm within this category is Inductive Logic Programming (ILP) [28], which utilizes machine learning to generalize abstract logical hypotheses and rules from empirical training examples, offering a promising avenue for verifiable yet flexible ethical behavior.

A pervasive challenge across all CME paradigms remains the formal verification and evaluation of an agent's ethical performance [39]. This evaluation must be measured against objective normative criteria or descriptive empirical benchmarks. **The specific mechanism for navigating and evaluating the computational moral trade-offs will be formalized in our contribution.**

Defeasibility. To formalize top-down ethical rules, early symbolic AI heavily relied on Deontic Logic, a branch of mathematical logic explicitly designed to formalize concepts of obligation, permission, and prohibition [26]. While Deontic systems offer absolute auditability, they notoriously struggle with contradictions (e.g., Chisholm's Paradox). When two absolute moral rules conflict in a real-world scenario, standard logic breaks down into inconsistency. To mitigate this rigidity, researchers introduced Defeasible and Non-monotonic Logics [9], where rules hold true by default but can be defeated or overridden by new evidence. This shift from absolute logical truths to defeasible rules serves as the historical and theoretical bridge directly leading toward Argumentation Theory.

2.2 Argumentation theory

Evolution of argumentation frameworks In Artificial Intelligence, computing rational decisions in the face of conflicting information or contradictory goals is highly complex, as traditional logic breaks down when rules conflict. Argumentation theory addresses this by providing formal mechanisms to model debate through defeasible reasoning, where conclusions remain valid only until overridden by stronger, defeating arguments.

To formalize this computationally, Dung [20] introduced the foundational *Abstract Argumentation Framework* (AF). This paradigm treats arguments as atomic entities (nodes in a graph), completely ignoring their internal structure to focus entirely on their external interactions via a binary attack relation. To determine which arguments can be rationally accepted together, Dung defined various *extension-based semantics* based on conflict-free and admissible sets (e.g., grounded, preferred, or stable extensions).

To better mirror human debate, subsequent literature extended this abstract paradigm. Bench-Capon [12] introduced *Value-Based Argumentation Frameworks* (VAFs), which ground arguments in moral or societal values to resolve conflicts qualitatively based on preference hierarchies. Concurrently, Cayrol and Lagasquie-Schiex [17] introduced the *Bipolar Argumentation Framework* (BAF), augmenting the classic negative attack relation with a positive binary support relation ($\dashrightarrow$) to represent premises that reinforce or confirm one another.

Weighted bipolar argumentation frameworks While topological frameworks assume all arguments are equally valid by default, real-world arguments possess varying degrees of intrinsic strength due to source trustworthiness, prior probabilities, or inherent biases. To incorporate this, Amgoud *et al.* [6] and Dunne *et al.* [21] integrated numerical weights into abstract models, leading to the *Weighted Bipolar Argumentation Framework* (WBAF).

Definition 1. *(Weighted Bipolar Argumentation Framework). A Weighted Bipolar Argumentation Framework (WBAF) is a quadruple* $\mathcal{F} = \langle \mathcal{A}, \mathcal{R}, \mathcal{S}, w \rangle$ *where:*

- $\mathcal{A}$ *is a finite, non-empty set of abstract arguments;*
- $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$ *is a binary attack relation ($a \to b$ denotes that a attacks b);*

- $\mathcal{S} \subseteq \mathcal{A} \times \mathcal{A}$ *is a binary support relation* ($a \dashrightarrow b$ *denotes that* a *supports* b*);*
- $w : \mathcal{A} \to [0, 1]$ *is a weighting function assigning an initial weight* $w(a)$ *to each argument* $a \in \mathcal{A}$*, independently of its interactions.*

Example 1. To illustrate, consider an ethical conflict regarding agricultural policy where four arguments are extracted from a social media platform, with initial weights representing their positive voting ratios:

- **(a):** The government should strictly ban synthetic chemical pesticides ($w(a) = 0.92$).
- **(b):** Utilizing chemical pesticides maximizes crop yields and boosts corporate revenues ($w(b) = 0.53$).
- **(c):** Continuous chemical usage degrades soil health, collapsing long-term productivity ($w(c) = 0.78$).
- **(d):** Pesticide-free farming safeguards local ecosystems and essential pollinators ($w(d) = 0.84$).

This scenario is formalized by the WBAF $\mathcal{F} = \langle \mathcal{A}, \mathcal{R}, \mathcal{S}, w \rangle$, where $\mathcal{A} = \{a, b, c, d\}$, $\mathcal{R} = \{(b, a), (c, b)\}$, and $\mathcal{S} = \{(d, a)\}$. The corresponding weighted interaction graph is illustrated in Figure 1.

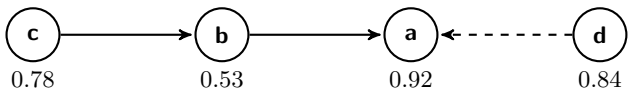

Fig. 1: Graph representation of the WBAF from Example 1. Decimal values below the nodes represent initial weights $w(a)$, solid arrows represent attacks ($\mathcal{R}$), and the dashed arrow represents support ($\mathcal{S}$).

Gradual semantics To evaluate a WBAF quantitatively, Cayrol and Lagasquie-Schiex [16] proposed *gradual semantics*. Unlike extension-based approaches, gradual semantics map topological debates to numerical assignments, making them highly compatible with continuous optimization frameworks like reinforcement learning. Modern literature decomposes this process into two modular mechanisms: an *aggregation function* summarizing the collective force of an argument's attackers and supporters, and an *influence function* balancing this collective force against the argument's intrinsic weight [7].

Definition 2. *(Gradual Semantics for WBAF). A gradual semantics* σ *is a mapping that assigns to any WBAF* $\mathcal{F} = \langle \mathcal{A}, \mathcal{R}, \mathcal{S}, w \rangle$ *a scoring function* $Deg_{\mathcal{F}} : \mathcal{A} \to [0, 1]$*. For a given argument* $a \in \mathcal{A}$*, the value* $Deg_{\mathcal{F}}(a)$ *represents its final acceptability degree, computed recursively or iteratively based on its initial weight* $w(a)$ *and the degrees of its direct attackers and supporters.*

Various gradual semantics exist, such as the weighted h-categorizer [14], DF-QuAD [31], and Quadratic Energy semantics [30]. These models are evaluated against rationality postulates formalized by Amgoud *et al.* [7] and expanded by Rapberger *et al.* [32]. Here is a non-exhaustive list of the core properties satisfied by DF-QuAD.

- **Anonymity & Independence.** Scores depend entirely on graph connectivity, not argument identities. Disjoint subgraphs cannot influence one another.
- **Directionality:** Influence flows strictly forward from direct attackers or supporters, unaffected by downstream consequences.
- **Stability & Neutrality.** Unattacked/unsupported arguments retain their initial weights. Targets with zero-credibility peers remain unaffected.
- **Equivalence & Duality.** Positive and negative forces carry equal weight; a perfectly balanced debate cancels out, leaving the target at its baseline score.
- **Monotonicity & Reinforcement.** Adding or strengthening supporters always strictly increases the final score, while strengthening attackers strictly decreases it.
- **Open-mindedness.** No argument is completely immune; overwhelming attacks can reduce any final score to 0, while overwhelming support can push it to 1.

While standard gradual semantics use static baseline values, dynamic environments require context-dependent weights. To handle this, we introduce a contextualized extension of classical DF-QuAD semantics, detailed in Section 4.3. Ultimately, this continuous mapping allows our symbolic framework to generate the dynamic numerical feedback required for autonomous policy optimization, making it naturally compatible with Reinforcement Learning.

2.3 Reinforcement learning

Concepts. Reinforcement Learning (RL) is a subfield of machine learning in which an agent iteratively interacts with an environment, learning to make optimal decisions by receiving numerical rewards or penalties based on its actions [38]. Unlike supervised learning, which relies on pre-labeled datasets, RL operates through an active trial-and-error process, making it particularly suitable for sequential decision-making tasks where the agent must balance immediate returns against long-term consequences.

Markov decision processes (MDP). To formally model the reinforcement learning problem, the interaction between the agent and its environment is typically framed as a Markov Decision Process (MDP). An MDP provides a mathematical framework for modeling decision-making in environments with stochastically dynamic transitions influenced by an agent's actions. Formally, an MDP is defined as a tuple $M = \langle S, A, T, R, \gamma \rangle$, where:

- S is a set of states representing the environment;
- A is a set of actions available to the agent;
- $T(s'|s,a)$ is the transition dynamics, representing the probability of reaching state $s' \in S$ after taking action $a \in A$ in state $s \in S$;
- $R(s,a,s')$ is the reward function, which outputs a scalar feedback signal received after transitioning from s to s' via action a;
- $\gamma \in [0,1)$ is the discount factor, which determines the present value of future rewards.

The behavior of the agent is defined by its policy $\pi(a|s)$, which maps states to a probability distribution over actions. The ultimate objective of an RL agent is to find an optimal policy π^* that maximizes the expected discounted return:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t, s_{t+1}) \right] \tag{1}$$

This formulation forces the agent to balance immediate rewards against long-term consequences, where the discount factor γ determines the present value of future payoffs.

Multi-Agent Reinforcement Learning (MARL). While the classical MDP framework assumes a static environment from the sole perspective of a single decision-maker, many real-world scenarios involve systems where multiple autonomous entities interact simultaneously. MARL extends the RL paradigm to these shared, dynamic environments [2]. For a comprehensive taxonomy and an overview of the state-of-the-art approaches in MARL, we refer to Buşoniu *et al.* [15].

Formally, MARL transitions the mathematical modeling from single-agent MDPs to Stochastic Games (or Markov Games). In this setting, the state transitions and the reward functions no longer depend on an individual action, but rather on the *joint action* $\mathbf{a} = \langle a_1, \ldots, a_n \rangle$ of all n agents within the environment. This shift introduces severe computational and theoretical challenges, most notably *non-stationarity*. The effective dynamics of the environment fluctuate as other agents concurrently update their policies. As a result, agents must continuously adapt to the shifting patterns of coordination, competition, or cooperation.

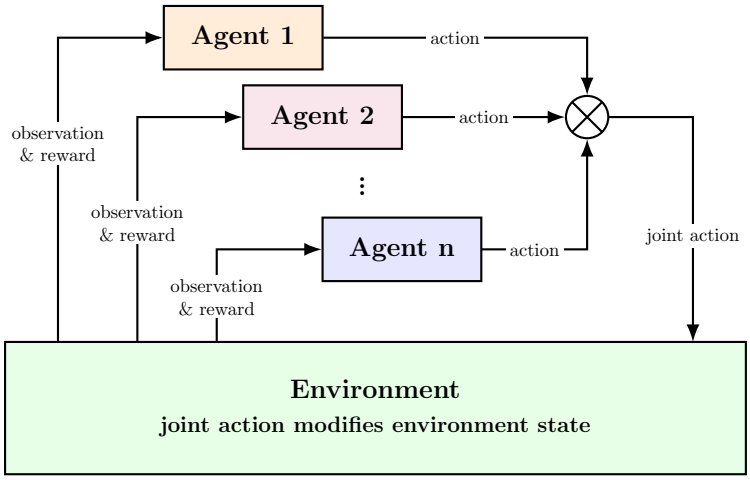

Fig. 2: Schematic of multi-agent reinforcement learning. A set of n agents receive individual observation about the state of the environment, and choose actions to modify the state of the environment. Each agent then receives a scalar reward and a new observation, and the loop repeats [2].

Partial observability and Dec-POMDPs. A further extension of the multi-agent paradigm accounts for situations where agents lack perfect information about the global environment. In such scenarios, the framework is modeled as a Decentralized Partially Observable Markov Decision Process (Dec-POMDP), forcing agents to coordinate based on localized, incomplete, or noisy observations [13].

Independent Proximal Policy Optimization. Independent Proximal Policy Optimization (IPPO) treats multi-agent learning by decentralizing the policy networks [41]. In this paradigm, an independent PPO [34] network is trained for each individual agent. From a theoretical standpoint, non-stationarity presents a severe challenge because all agents learn concurrently, constantly altering the transition dynamics from any single agent's perspective. IPPO addresses this by allowing each agent to optimize its individual policy independently, treating other active agents simply as part of the changing environmental context. Despite its lack of explicit coordination mechanisms, IPPO has been shown to establish a remarkably robust baseline for decentralized behavioral alignment, often matching or outperforming more complex joint-action algorithms while maintaining high empirical stability.

Reward design and alignment. In both single and multi-agent RL, the reward function R serves as the primary mechanism for defining the agents' objectives. While scalar feedback is highly effective for clear-cut tasks, it becomes

a significant vulnerability when deploying systems in complex human environments. Poorly designed reward functions often lead to *reward hacking* [8], where an agent uncovers a mathematically optimal but practically undesirable (or unethical) policy to maximize its score. Consider a cleaning agent rewarded strictly on the volume of collected debris. Instead of maintaining a clean space, the agent may optimize its policy by repeatedly scattering dirt onto the floor only to vacuum it again, thereby maximizing its cumulative reward while completely undermining the intended goal.

3 State of the art and positioning

Building upon the theoretical foundations established in the previous section, this section explores the current landscape of Computational Machine Ethics (CME). We review the state-of-the-art methodologies attempting to solve ethical challenges within learning environments, analyze how recent works have utilized argumentation to mitigate these issues, and synthesize the specific theoretical gaps our contribution aims to fill.

3.1 Current methodologies in ethical reinforcement learning

The integration of moral considerations into autonomous learning agents has primarily been approached through two distinct methodological axes, each presenting significant structural limitations:

- **Bottom-up (learning-based) approaches.** Modern machine learning models attempt to infer ethical behavior empirically from data. *Inverse Reinforcement Learning (IRL)* assumes an agent can extract an implicit "moral" reward function from human demonstrations, a technique frequently applied to value alignment and norm learning [23,11]. In pioneering work bridging reinforcement learning and machine ethics, Abel *et al.* [1] demonstrated how agents could learn to respect common-sense utilitarian ethics by observing human-curated environments. Similarly, *Reinforcement Learning from Human Feedback (RLHF)* scales this alignment by optimizing policy behavior based on pairwise human preferences [18].
  Multi-Objective Reinforcement Learning (MORL) serves as the mathematical backbone for these bottom-up approaches, as ethical RL naturally balances task efficiency against competing ethical rewards (fairness, well-being). However, both IRL and RLHF operate as "black boxes": they lack symbolic transparency regarding moral trade-offs and can propagate systemic biases present in training data.
- **Hybrid neuro-symbolic approaches.** To overcome the opacity of pure neural optimization, researchers combine the adaptability of RL with the structural transparency of symbolic rules. A prominent paradigm is the "shielding" architecture [4], where a logical verification layer monitors the RL agent's action proposals and dynamically blocks non-compliant choices.

While this provides formal guarantees, hard-coded symbolic shields lack the flexibility required to autonomously resolve edge cases where multiple ethical principles or duties enter into direct conflict.

It is critical to distinguish Computational Machine Ethics from the neighboring field of *Safe RL* (often modeled via Constrained MDPs) [5]. Safe RL focuses primarily on *normative compliance*, enforcing rigid operational boundaries to strictly forbid identified dangerous or undesired behaviors to preserve physical or functional system integrity. Conversely, CME operates in the domain of *ethical reasoning*, which requires evaluating abstract values, weighing moral arguments, and managing dilemmas where binary constraints are fundamentally reductionist and insufficient.

3.2 Argumentation-Based Reinforcement Learning

To address the inability of standard symbolic logic to handle conflicting rules, recent works have integrated Argumentation Theory directly into the Reinforcement Learning (RL) loop. These approaches primarily leverage symbolic frameworks to perform reward shaping or guide action selection. A foundational framework proposed by Gao *et al.* [22] utilizes abstract argumentation to inject expert domain knowledge into RL agents, accelerating policy convergence by filtering unviable actions. The closest architecture to our objective is **AJAR** (Argumentation-based Judging Agents Framework for Ethical Reinforcement Learning) developed by Alcaraz *et al.* [3]. AJAR introduces machine ethics into the training loop by separating learning agents from external "judging agents." These judges instantiate an abstract argumentation graph containing "pros" and "cons" arguments regarding the learning agent's behaviors, outputting a scalar reward derived from the numerical ratio of accepted arguments.

Limitations of Extension-Based RL Approaches. While both Gao *et al.* and AJAR successfully bridge symbolic reasoning with connectionist learning, their underlying semantic layers remain structurally rigid. Whether capturing domain expertise or moral norms, both frameworks evaluate their argumentation graphs using classical Dung-style *extension-based semantics* (such as the grounded or preferred extensions). Consequently, argument acceptability is strictly Boolean: an argument is either entirely accepted (warranted) or completely rejected. However, a policy is rarely purely "right" or "wrong." Real-world moral dilemmas inherently involve continuous degrees of compliance, multi-objective trade-offs, and subtle compromises. Relying on discrete extensions fails to capture these nuances and generates sparse, non-differentiable reward signals that can destabilize policy learning. Our architecture explicitly overcomes this limitation by shifting from binary extensions to continuous gradual semantics within a weighted bipolar framework.

3.3 Positioning and synthesis

The current landscape of ethical AI presents a clear dichotomy: bottom-up approaches are adaptive but opaque, while existing argumentation-based RL frameworks (like AJAR) are transparent but semantically rigid due to their reliance on Boolean logic. To bridge this gap, this report proposes a novel architecture substituting classical abstract argumentation with a **Weighted Bipolar Argumentation Framework (WBAF)** evaluated through **gradual semantics (DF-QuAD)**. Our contribution introduces the following paradigm shifts:

1. **Continuous evaluation.** To bypass the binary constraints of Dung-style extensions, we advocate for a dynamic evaluation where argument acceptability is continuous, allowing the representation of fine-grained degrees of moral compliance.
2. **Contextualization of values.** By utilizing initial weights on arguments, our framework can directly translate human moral priorities (e.g., valuing "safety" over "efficiency") into the mathematical topology of the reasoning engine.
3. **Transparent RL integration.** This continuous acceptability degree serves as a nuanced, mathematically sound reward signal. It natively integrates with RL algorithms, equipping agents with an explicit, interpretable morality score that accurately reflects the subtleties of ethical trade-offs.

4 Proposed framework and contribution

This section presents our main contribution: a democratized, neuro-symbolic framework for Multi-Agent Reinforcement Learning (MARL) that grounds ethical behavior in explicit, human-designed argumentative debates.[4] Following Dignum's call for responsible AI [19], our architecture shifts the responsibility of reward design from computer scientists to a wider collective of domain experts, ethicists, citizens, and relevant stakeholders. By formalizing these debates through argumentation framework with gradual semantics, we equip decentralized agents with transparent, adaptive, and auditable moral guidelines.

To ground our theoretical framework in a concrete experimental setting, we first define the simulated environment tailored for this study, before detailing the architecture and mathematical integration of our argumentative layer.

4.1 Use case

As noted by Murukannaiah et al. [29], "ethics is inherently a multiagent concern." To capture this nature and evaluate our architecture under non-stationary

[4] The complete open-source implementation of our framework and the *Ethical Gardeners* environment are available at `https://gitlab.liris.cnrs.fr/eciea/ethical-gardeners.git`.

dynamics, we use a simulated gridworld environment named **Ethical Gardeners** (illustrated in Figure 3), heavily inspired by the AI Safety Gridworlds benchmark [25]. Formally modeled as a Dec-POMDP as introduced in Section 2.3, the environment places multiple autonomous agents in a shared spatial grid representing a community garden.

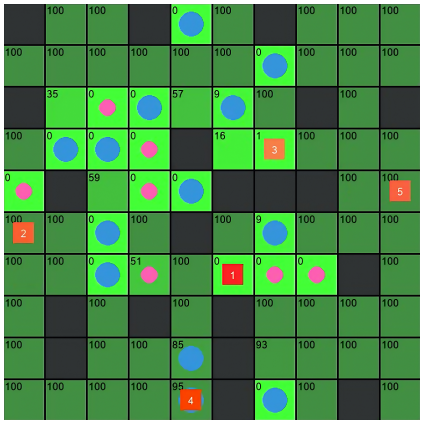

Fig. 3: Overview of the *Ethical Gardeners* gridworld environment. Green tiles represent cultivable land with their pollution level (from 0% to 100%). Red squares represent agents, blue circles indicate standard flowers, and pink circles indicate exotic flowers.

Agents and action space. The agents act as independent gardeners whose individual objective is economic optimization within a decentralized learning loop. Each gardener aims to maximize its net financial profit by cultivating and harvesting flowers. We assume an idealized economic market where harvested flowers are immediately sold at a deterministic value, thereby abstracting away transactional matching or buyer-search friction. At each discrete time step, an agent i selects a single capability from a discrete action space $\mathcal{A}_i$, defined as follows:

- **Spatial navigation.** Move north, south, east or west.
- **Inaction.** Wait.
- **Cultivation.** Plant a specific species of flower (standard or exotic) on the current tile, deducting the initial seed cost from the agent's budget.
- **Harvesting.** Gather a mature flower from the current tile to instantly realize its economic return.

Formalization of world metrics. At each step t, the global environment state $s_t \in \mathcal{S}$ is quantified through four metrics normalized within $[0, 1]$:

- **Agent wealth (M_i).** The capital $M_i(t) \geq 0$ of each agent $i \in \{1, \ldots, N\}$ increases via flower harvesting and decreases with seed purchase costs.
- **Air quality (A).** Let K be the set of cultivable tiles. The local pollution $P_k(t)$ of an empty tile increases by 1 unit per step up to 100. Planting a flower at time τ reduces $P_k(t)$ by 5 unit per step after a species-dependent growth age delay $\Delta t = t - \tau$: $\Delta t \geq 1$ for standard flowers, and $\Delta t \geq 3$ for exotic flowers. The global pollution index $\mathcal{P}(t) \in [0, 1]$ normalizes the average tile pollution $\bar{P}(t) = \frac{1}{|K|} \sum_{k \in K} P_k(t)$:

$$\mathcal{P}(t) = \max\left(0.0, \min\left(1.0, \frac{\bar{P}(t) - P_{\min}}{P_{\max} - P_{\min}}\right)\right)$$

  Air quality is defined as the complement: $A(t) = 1 - \mathcal{P}(t)$.
- **Ecosystem biodiversity (B).** Measured via the normalized Shannon Entropy index over the distribution of standing flower species $J = \{\text{standard}, \text{exotic}\}$:

$$B(t) = -\frac{1}{\ln(|J|)} \sum_{j \in J} p_j \ln(p_j)$$

  where p_j is the proportion of species j. $B(t) = 1.0$ for perfect species balance, 0.0 for monocultures, and $B(t) = 0$ if the grid is empty.
- **Socio-economic equality (E).** To avoid the scale-invariance bias of indices like Gini which mask absolute wealth gaps as the baseline economy scales, we define an environment-bounded dispersion index $\mathcal{I}(t)$ as the standard deviation (σ) of normalized wealth:

$$\mathcal{I}(t) = \sigma\left(\left\{\frac{M_i(t)}{M_{\max}}\right\}_{i=1}^{N}\right)$$

  Since $\mathcal{I}(t) \leq 0.5$ for values in $[0, 1]$, the final metric $E(t) \in [0, 1]$ is scaled and clamped:

$$E(t) = \max\left(0.0, \min\left(1.0, 1.0 - \frac{\mathcal{I}(t)}{0.5}\right)\right)$$

  where $E(t) = 1$ signifies perfect equality ($\mathcal{I}(t) = 0$) and 0 represents the absolute worst-case wealth dispersion.

Ethical dilemmas. While the direct economic objective is straightforward, the shared nature of the grid and the global dynamics of the garden formalised above introduce severe ethical dilemmas. Specifically, the framework evaluates the system across four competing objectives that the decentralized agents must learn to balance:

- **Individual wealth vs. free-riding.** Agents must manage a limited budget (M_i) to plant flowers, which requires a short-term financial investment for a

long-term return upon harvesting. However, because the environment operates as a shared, open-access common resource, there is an inherent risk of exploitation. An agent can choose to "steal" and harvest a flower that was planted and paid for by another agent.

- **Air quality and ecological recovery.** To maximize global air quality, the garden must remain covered with flowers. This creates a direct tension with the economic goal, as harvesting a flower for profit removes the vegetation and increases the tile's pollution P_k. Furthermore, standard flowers depollute the air rapidly but yield lower profits, whereas highly profitable exotic flowers depollute the environment much more slowly due to their delayed activation curve.
- **Biodiversity preservation.** To avoid destructive monocultures, agents must actively coordinate their actions to maximize the Shannon entropy (B), ensuring that no single flower type dominates the grid.
- **Socio-economic equity.** The system monitors wealth distribution among the gardeners. The framework penalizes monopolistic behaviors or severe income inequality by tracking the inequality index (E).

These systemic trade-offs are not hard-coded into the game's physics or transition rules. Instead, agents must balance selfish profitability against ecological preservation and social fairness, solely through the nuanced, shaped rewards computed by our proposed argumentative framework.

4.2 Global architecture overview

To solve the dilemmas inherent to the Ethical Gardeners environment, our architecture introduces a symbolic approach to the reward design. The ethical component of the RL feedback loop is directly anchored in human-designed argumentation. Rather than manually engineer opaque mathematical formulas (e.g., a penalty of -5 for every exotic flower planted), our framework allows domain experts to declare systemic trade-offs through modular, action-specific argumentative graphs.

As illustrated in Figure 4, the evaluation of these action-specific graphs occurs dynamically post-execution. The joint action transitions the environment into a new state, which the framework uses to instantiate the graphs and compute each agent's ethical feedback.

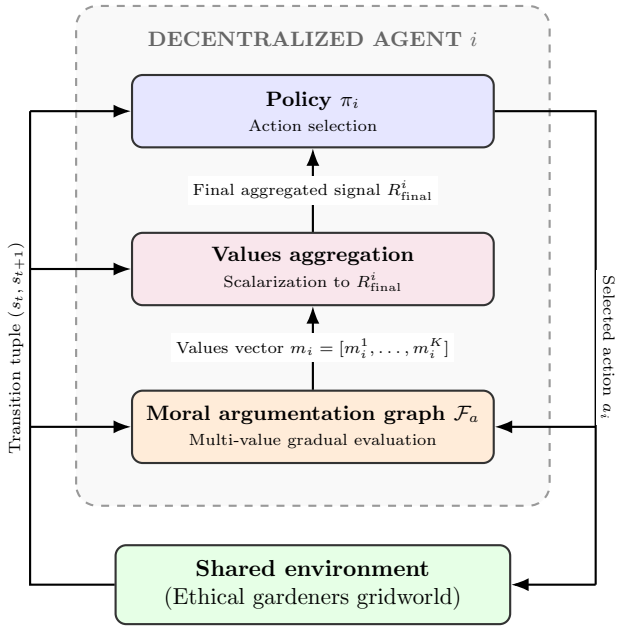

Fig. 4: Algorithmic flow chart of the hybrid learning architecture for a representative agent i. When an action a_i is executed, the environment transitions from the current state s_t to the next state s_{t+1}. This transition tuple (s_t, s_{t+1}) is distributed back to the agent's internal modules. The *Moral Argumentation Graph* evaluates the ethical acceptability of a_i across multiple dimensions within this context, outputting a vector of moral scores m_i (one for each moral value). In the *Values aggregation*, these diverse moral dimensions are scalarized via a customizable alignment function to produce the final scalar training signal R^i_{final} for the policy function π_i.

By bridging decentralized learning with argumentation graphs, this architecture establishes three foundational pillars:

- **Transparency and auditability:** The exact reasoning behind an agent's ethical trade-offs can be post-hoc audited by inspecting the activation paths within the underlying argumentation graph.
- **Mitigation of specification hacking:** Anchoring training signals in symbolic human values minimizes reward hacking; novel optimization loopholes can be transparently corrected by introducing counter-arguments into the graph.
- **Collaborative governance:** By removing the technical barriers of manual reward engineering, our architecture allows a diverse group of regulators and citizens to collaboratively define the moral boundaries of autonomous systems.

4.3 Moral Argumentation Graph Module

To evaluate actions ethically, we associate exactly one fixed Moral Argumentation Graph template with each action a_i available in the environment. This ensures a direct, one-to-one mapping and keeps the total number of graphs finite. For the sake of reproducibility, the complete set of graphs can be found in Appendix B.

Definition 3. *(Moral Argumentation Graph). A Moral Argumentation Graph for an action a_i is a WBAF $\mathcal{F}_{a_i} = \langle \mathcal{A}, \mathcal{R}, \mathcal{S}, w \rangle$ where the arguments $\mathcal{A} = \mathcal{A}_V \cup \mathcal{A}_C$ are partitioned into:*

- *$\mathcal{A}_V$: A fixed set of fundamental moral value nodes pre-defined by domain experts based on the normative framework of the task (e.g., poverty reduction, biodiversity preservation, and social equality), where each value $v \in \mathcal{A}_V$ is initialized with a full-force weight $w(v) = 1$.*
- *$\mathcal{A}_C$: A fixed set of contextual arguments that structurally attack or support these values, or one another, where each argument $c \in \mathcal{A}_C$ has an initial weight derived from current world states (e.g., the argument "we should plant less exotic flowers" is initialized by the world state "proportion of exotic flowers").*

The graph topology is a pre-designed template mapping action-specific ethical dilemmas. The shared environment interacts with this module exclusively via dynamic weighting: executing a_i prompts the system to retrieve $\mathcal{F}_{a_i}$ and compute the initial weights of contextual arguments ($c \in \mathcal{A}_C$) directly from the observed state transition (s_t, s_{t+1}). This allows the factual consequences of an action such as economic gain or biodiversity improvement to directly dictate the initial strength of the arguments in the graph.

Figure 5 illustrates this via a subgraph for *"Harvest standard flower"*. Here, the node *"High air pollution"* is triggered with an initial weight of 0.890 (the garden's pollution index), signaling degraded air quality and attacking the *"Health and Well-being"* value. This argument is supported by *"Helped fast pollution reduction"* (weighted by the flower's remediation power) to capture the negative impact of removing a depolluting agent. Conversely, it is counter-attacked by *"Garden health improved"*, mitigating the penalty based on systemic progress. Under our gradual semantics, these conflicting dynamics increase the final acceptability of *"High air pollution"* from 0.890 to 0.981, strengthening its penalization of the *"Health and Well-being"* moral score.

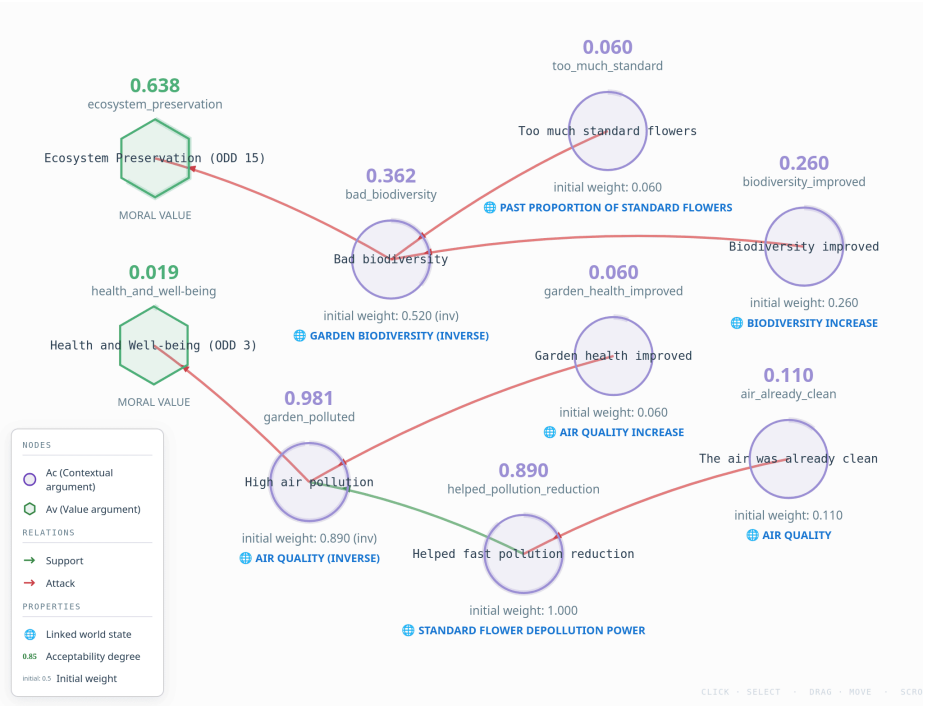

Fig. 5: Subgraph of the argumentation framework instantiated for the action "Harvest standard flower", focusing on the "Ecosystem preservation" and "Health and Well-being" values.

Weighted DF-QuAD Semantics. Traditional DF-QuAD formulations [31] assume a uniform, static baseline strength (typically 0.5) for all arguments. To capture varying situational urgency within the *Ethical Gardeners* domain, we introduce a weighted extension where each argument a is initialized with a dynamic weight $w(a) \in [0, 1]$ derived directly from the environmental transition (s_t, s_{t+1}). This ensures that the baseline credibility of moral arguments adapts natively to the underlying system dynamics before gradual semantics compute their final acceptability.

We specifically select the DF-QuAD formulation because it enables a strict veto mechanism. A maximally strong argument ($p(x) = 1$) can completely nullify opposing forces, which is a property vital for enforcing hard ethical constraints. For instance, if a flower species faces extinction, the argument asserting the absolute urgency to plant it must completely overpower competing economic optimization arguments. Ultimately, this weighted formulation preserves the foundational axiomatic properties of traditional DF-QuAD detailed in Section 2.2 while enabling context-sensitive ethical reasoning, as formalized below.

Definition 4. *(Weighted DF-QuAD Semantics). Let* $\mathcal{F} = \langle \mathcal{A}, \mathcal{R}, \mathcal{S}, w \rangle$ *be a WBAF. For each argument* $a \in \mathcal{A}$*, its acceptability degree* $p(a) \in [0, 1]$ *is computed iteratively starting from* $w(a)$*. Let* $\mathrm{Att}(a)$ *and* $\mathrm{Sup}(a)$ *denote the sets of attackers and supporters of* a*, whose aggregated strengths are defined as:*

$$S_{\mathrm{att}}(a) = 1 - \prod_{x \in \mathrm{Att}(a)} (1 - p(x)), \qquad S_{\mathrm{sup}}(a) = 1 - \prod_{y \in \mathrm{Sup}(a)} (1 - p(y)),$$

with the convention that an empty product equals 1 *(yielding* $S_{\mathrm{att}}(a) = 0$ *or* $S_{\mathrm{sup}}(a) = 0$*). Given the net impact* $\mathrm{net}(a) = S_{\mathrm{sup}}(a) - S_{\mathrm{att}}(a)$*, the updated acceptability degree is computed as:*

$$p(a) = \begin{cases} w(a) & \text{if } \mathrm{net}(a) = 0, \\ w(a) + (1 - w(a)) \times \mathrm{net}(a) & \text{if } \mathrm{net}(a) > 0, \\ w(a) + w(a) \times \mathrm{net}(a) & \text{if } \mathrm{net}(a) < 0. \end{cases}$$

The computation iterates until the variation between two consecutive iterations falls below a predefined convergence threshold.

4.4 Values Aggregation

Once the gradual semantics evaluation converges within the argumentation framework, the graph $\mathcal{F}_{a_i}$ outputs a vector of ethical acceptability degrees representing the final scores of the K fundamental moral values:

$$\mathbf{m}_i = [m_i^1, m_i^2, \ldots, m_i^K] \in [0, 1] \tag{2}$$

where each component m_i^k quantifies the ethical compliance of action a_i with respect to the moral value k (e.g., ecological preservation or social equity) under the current state transition.

To operationalize these multi-dimensional ethical evaluations into a single learning signal for the Reinforcement Learning policy π_i, we introduce an aggregation function $g : [0, 1] \to \mathbb{R}$. This function maps the moral compliance vector to a scalar training reward:

$$R_{\mathrm{moral}}^i = g(\mathbf{m}_i) \tag{3}$$

Depending on the normative constraints of the application, we implement two distinct scalarization strategies:

– **Minimum bottleneck aggregation.** To enforce strict compliance with all moral dimensions simultaneously, we instantiate $g(\cdot)$ using a pessimistic bottleneck operator inspired by deontological ethics. Under this formulation, an action is judged by its weakest ethical link: a severe violation of a single moral value collapses the entire signal.
To transform the raw acceptability degree from $[0, 1]$ into a signal capable of actively penalizing harmful behaviors, the bottleneck score is linearly rescaled to a symmetric reward space $[-1, 1]$:

$$R_{\mathrm{moral}}^i = 2 \times \left(\min_{k \in \{1, \ldots, K\}} m_i^k \right) - 1 \tag{4}$$

Consequently, an action that perfectly satisfies all moral values ($m_i^k = 1, \forall k$) yields a maximum positive reinforcement of $+1$, whereas failing a single ethical constraint ($m_i^k = 0$) results in a maximum negative penalty of -1. From an optimization perspective, while a standard min operator induces non-differentiability and can cause dead gradients in traditional multi-objective settings, our framework inherently mitigates exploratory paralysis. Because the underlying WBAF relies on continuous gradual semantics rather than binary logic, the resulting scores m_i^k vary smoothly across state transitions. This topological continuity ensures that the bottleneck metric shifts dynamically and predictably, providing a dense, active learning signal throughout training.

– **Multi-objective task integration.** While isolating R^i_{moral} allows us to validate the pure ethical learning capabilities of the agents, our architecture natively extends to Multi-Objective Reinforcement Learning (MORL). To balance ethical compliance against the agent's intrinsic economic drive, the final alignment function synthesizes the ethical signal with the standard environmental task reward R_{task} using multiplicative gating:

$$R^i_{\text{final}} = R^i_{\text{task}} \times \left(\min_{k \in \{1,\dots,K\}} m_i^k \right) \tag{5}$$

This multiplicative formulation acts as a safety filter: if an action is economically highly lucrative ($R^i_{\text{task}} \gg 0$) but ethically unacceptable (at least one $m_i^k \to 0$), the final reward is crushed to zero. This mathematical structure forces the decentralized policy network to explicitly map and converge toward Pareto-optimal policies that successfully reconcile functional performance with normative alignment.

5 Evaluation

This section presents the empirical evaluation of our neuro-symbolic framework. We first outline the concrete implementation details and training infrastructure, followed by a comprehensive quantitative analysis of the agents' functional performance and normative alignment.

5.1 Implementation details

Multi-agent environment training. To validate the empirical viability of our neuro-symbolic framework, the *Ethical Gardeners* environment is implemented using the PettingZoo library[5], a modern Python standard for multi-agent reinforcement learning that succeeds the classic OpenAI Gym interface. PettingZoo provides the necessary programmatic abstractions to model the complex, step-by-step parallel interaction loops required by our Dec-POMDP formulation.

[5] `https://pettingzoo.farama.org/`

Under this setup, the agents' individual policy networks are optimized using the IPPO algorithm described in Section 2.3, allowing each gardener to scale independently in the shared space. The complete procedural execution of this multi-agent training loop, including the integration of our moral argumentation framework within the step execution, is formalized in Algorithm 1 in the Appendix.

Moral reasoning layer. The moral reasoning layer is powered by a custom Python implementation of Weighted DF-QuAD semantics, which evaluates the active WBAF at each discrete step. To ground the normative alignment of our agents without relying on ad-hoc ethical assumptions, we adopt the United Nations Sustainable Development Goals (SDGs) as a universally recognized proxy for moral values. In the absence of dedicated philosophers or ethicists within the design loop, leveraging the SDG framework provides a rigorous, consensus-driven foundation that maps directly onto our experimental use case. Specifically, our macro-objectives operationalize four core SDG mandates: Poverty Reduction, Inequality Reduction, Ecosystem Preservation, and Health and Well-being.

For the implementation of our reward mechanism, we explicitly deploy the *Minimum bottleneck aggregation* strategy discussed in Section 4.4, mapping the minimum acceptability score across the four macro-objectives to a standardized $[-1, 1]$ reward signal via the linear transformation $R_i = 2 \times \min(m_i^k) - 1$. This mathematical framing prevents agents from balancing out or ignoring severe degradation in one moral value by maximizing performance in another. To illustrate how these symbolic SDG interactions and numerical reward transformations converge dynamically within a single operational step, a comprehensive use case evaluating a specific harvesting action is detailed in Figure B.1.

Environment configuration. The environment is structured as a 10×10 spatial gridworld where agents operate under full observability, meaning every gardener has a complete view of the global world state at each time step. While parameters such as Air Quality percentages, Shannon entropy for Biodiversity, and the standard deviation-based Equality index naturally fall within or are easily mapped to $[0, 1]$ as defined in Section 4.1, unconstrained domain variables like financial wealth introduce an intrinsic normalization challenge.

To handle agent capital, we instantiate each gardener with a base wealth of \$10 and a maximum threshold $M_{\max} = \$1000$. Within the argumentation graph, an agent reaching \$1000 is considered to have achieved 100% (1.0) of its economic capacity; any money accumulated beyond this cap continues to accumulate in the environment but saturates at 1.0 for the moral reasoning engine. This requirement to compress open-ended variables represents a notable constraint of mapping environment states to argumentation graphs, as defining these saturation thresholds requires manual, expert-driven calibration to prevent distorting the ethical evaluation.

Hyperparameter Tuning. Distinct from the environmental and structural constants, the learning performance of the multi-agent system heavily depends on the configuration of the optimization algorithm. To isolate the optimal training hyperparameters (such as learning rates, batch sizes, and policy entropy coefficients) under the mentioned environment bounds, an extensive grid search was executed independently for each method (see Fig. C.1). This empirical search ensures that performance improvements are strictly a product of architectural design rather than unoptimized training configurations.

5.2 Experimental protocol

Simulation setup and evaluation metrics. The experimental pipeline is structured to ensure statistical significance over stochastic training trajectories. To achieve statistical validity, every core configuration is executed across 20 independent simulation runs. Each individual simulation run consists of 1000 sequential episodes, where each episode spans 512 discrete environment steps. Crucially, the environment is reset to a random initial state at the start of every new episode. This total re-initialization prevents the agents from memorizing specific static layouts or overfitting to fixed trajectories, forcing them to instead develop robust, generalized policies capable of adapting to varied initial conditions.

To evaluate policy performance and stability, we record the ecosystem metrics described in Section 4.1 exclusively at the final step of each training episode. Terminal performance is calculated using a rolling window over the final 10 episodes (Episodes 991 to 1000), averaged across all 20 independent runs. We compute the empirical mean (μ) and standard deviation (σ) for four core objectives normalized within $[0, 1]$: Air Quality, Biodiversity, Equality, and Average Money (between all agents). A winning policy must demonstrate a high average performance across all four axes while maintaining a low standard deviation.

Baselines. Our proposed architecture is evaluated against two baseline models. The first is a *Random* baseline, where agents do not undergo any learning process and select capabilities uniformly at random, mapping the absolute lower bound of the environment. The second is an *Ablation (direct min)* baseline, which directly isolates the impact of the symbolic reasoning layer. This baseline serves as a structural ablation study of our framework (referring to Figure 4): we completely bypass the Moral Argumentation Graph Module. Instead, the raw world states corresponding to the fundamental values are fed directly into the *Values Aggregation* block, where they undergo the exact same minimum aggregation operator as our main framework. Under this setting, the agent receives a direct scalar reward computed as the raw minimum value of the four environmental dimensions at time t:

$$R_t = \min(\text{Air Quality}, \text{Biodiversity}, \text{Equality}, \text{Agent's money})$$

This baseline serves to verify whether introducing the explicit structural and constraints of the argumentation layer penalizes or restricts the agent's learning capacity compared to direct multi-objective optimization.

Scientific hypotheses. Prior to testing, three scientific hypotheses are formulated to guide our analysis:

Hypothesis 1 (Performance ordering and viability)**.** *The absolute baseline, Random, will exhibit the worst performance across the ecosystem metrics. Crucially, our proposed Argumentation framework will achieve competitive performance with the Ablation baseline, demonstrating that it does not decrease the agent's learning capabilities. It is expected to perform at least as effectively as or outperform the raw Ablation (direct min) approach, leading to the ordering: Random $<$ Ablation (direct min) $\leq$ Argumentation. The objective is to confirm that integrating a transparent, symbolic reasoning layer remains fully viable and does not introduce a performance penalty compared to a standard, non-argumentative Multi-Objective RL baseline.*

Hypothesis 2 (Reward alignment)**.** *The designed argumentation reward function is non-deceptive and sound. A positive trend in the cumulative reward across training episodes must directly correlate with an objective improvement in the actual environmental and economic metrics, proving that maximizing the argumentative reward does not cause sub-optimization or reward hacking.*

Hypothesis 3 (Faster convergence)**.** *The Argumentation framework will enable faster policy convergence across training episodes compared to the Ablation (direct min) baseline. Under a standard direct minimum operator, early actions carrying upfront costs (such as the financial cost of planting a seed) yield severe, unmitigated negative rewards, leading to exploratory paralysis and delayed convergence. By contrast, argumentation graphs can aggregate supporting arguments that contextualize immediate costs within long-term returns, smoothing the reward landscape and accelerating policy optimization.*

5.3 Results

Table 1 outlines the empirical performance metrics captured across the final 10 training episodes (Episodes 991 to 1000), averaged across the 20 independent training runs under their optimized hyperparameter configurations.

Table 1: Terminal performance comparison across baselines (mean ± standard deviation over 20 independent runs, computed as a rolling average over the final 10 training episodes [991–1000]. All metrics are normalized within $[0, 1]$).

Framework	**Air Quality**	**Biodiversity**	**Equality**	**Avg Money**
Random	0.30 ± 0.04	0.75 ± 0.07	$\mathbf{0.98 \pm 0.02}$	0.01 ± 0.01
Ablation (direct min)	$\mathbf{0.70 \pm 0.04}$	0.91 ± 0.05	0.57 ± 0.08	$\mathbf{0.67 \pm 0.10}$
Argumentation (Ours)	0.69 ± 0.05	$\mathbf{0.94 \pm 0.04}$	0.61 ± 0.10	0.66 ± 0.05

We observe distinct behavioral patterns across the three evaluated settings. The *Random* baseline exhibits deceptively high scores in Biodiversity (0.75) and Equality (0.98), which reflect behavioral stagnation rather than optimization. Near-perfect Equality stems from universal poverty, as agents fail to perform profitable actions and maintain near-zero wealth (0.01). High Biodiversity is a mechanical artifact: a uniform random policy inherently balances the environment's two flower species, maximizing Shannon entropy by default. Ultimately, this chaotic behavior fails to sustain the ecosystem, collapsing Air Quality to 0.30 with zero economic return.

The *Ablation (direct min)* baseline (representing standard multi-objective bottlenecking without argumentation) significantly improves overall performance, achieving strong results in Air Quality (0.70) and financial throughput (0.67), alongside a solid Biodiversity score (0.91). However, it struggles slightly more with socio-economic equity, yielding an Equality score of 0.57. Because the non-differentiable minimum operator forces the gradient to focus exclusively on the single lowest-performing metric at any given time step t, the resulting policy updates can be overly reactive. This restrictive focus successfully elevates individual metrics but introduces volatility that hinders the agents' ability to fully optimize the more delicate balance between biodiversity and wealth distribution.

In contrast, our neuro-symbolic *Argumentation* framework achieves a more holistically balanced alignment. It statistically matches the ablation baseline's peak performance in both Air Quality (0.69 ± 0.05) and Average Money (0.66 ± 0.05), while outperforming it in both Biodiversity (0.94) and Equality (0.61). We hypothesize that the continuous nature of the argumentation semantics provides a richer, more nuanced reward signal compared to the strict bottlenecking of the direct minimum operator. This potentially helps the agents better navigate delicate socio-ecological trade-offs rather than over-reacting to single-metric drops. Finally, while its Equality score (0.61) is numerically lower than the artificial stagnation of the *Random* baseline, it reflects a true dynamic equity where substantial capital is actively generated and shared.

Hypothesis validation By tracking the environmental metrics alongside the cumulative reward curves across the 1000 training episodes, we evaluate our initial scientific assumptions against the empirical data:

Performance ordering and viability (1). The empirical results summarized in Table 1 partially substantiate this assumption, validating the hierarchy Random $<$ Ablation (direct min) $\leq$ Argumentation. Our framework statistically matches the peak performance of the *Ablation (direct min)* baseline in both Air Quality (0.69 ± 0.05 vs. 0.70 ± 0.04) and Average Money (0.66 ± 0.05 vs. 0.67 ± 0.10), while outperforming it in Biodiversity (0.94 vs. 0.91) and Equality (0.61 vs. 0.57). This is a highly positive outcome, as it demonstrates that introducing an explicit argumentation layer for transparency does not penalize the agents' learning efficiency. On the contrary, it allows the system to achieve a more robust compromise across conflicting dimensions without sacrificing overall viability.

Reward alignment (2). The soundness of our approach is clearly demonstrated by the strong synergy between internal rewards (Figure 6) and external metrics (Figure 7). Throughout the 1000 training episodes, a steady upward trajectory in the cumulative argumentative reward directly correlates with a simultaneous improvement in both the ecological state and economic health of the ecosystem. This tight coupling suggests that our neuro-symbolic reward mechanism effectively translates complex symbolic ethics into reliable numerical policies without introducing deceptive local minimum or reward hacking.

Faster convergence (3). The learning curves illustrated in Figures 7 and 8 provide empirical evidence for this acceleration, although the gain remains minor. While both learning frameworks follow a relatively similar training trajectory, the contextual feedback provided by the argumentation graph appears to slightly help agents reconcile immediate upfront costs (e.g., planting seeds) with long-term ecological returns. By structuring short-term financial penalties within a broader framework of arguments, the gradual semantics could facilitate a slightly faster policy optimization.

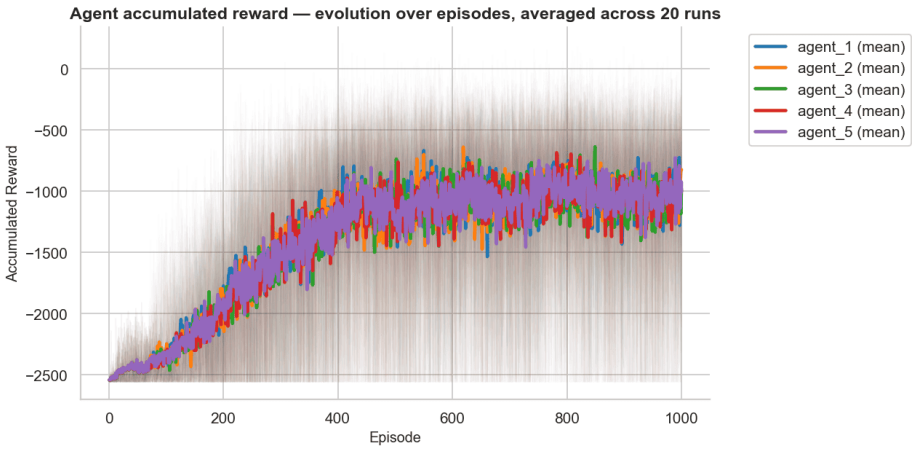

Fig. 6: Agent accumulated reward in an episode under our argumentation framework. We observe a steady improvement over episodes, closely correlating with the objective improvements in environmental and economic values shown in Fig. 7. Since the simulation initializes in a highly degraded state characterized by maximum pollution, minimal capital, and a total absence of flowers, accumulated rewards are expectedly negative during early episodes.

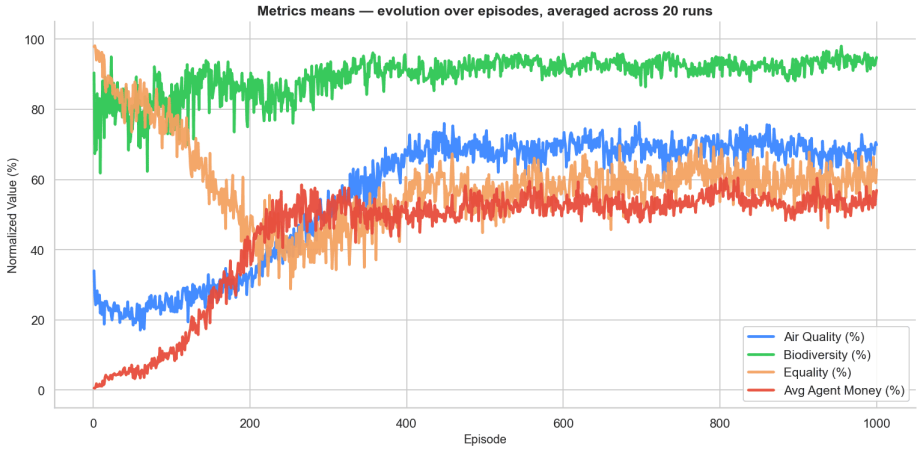

Fig. 7: Temporal evolution of environmental metrics (mean values) under our proposed argumentation framework over 1000 training episodes. The curves demonstrate a stable, concurrent optimization of agents wealth, air quality, biodiversity and socio-economic equality, validating balanced multi-objective policy convergence.

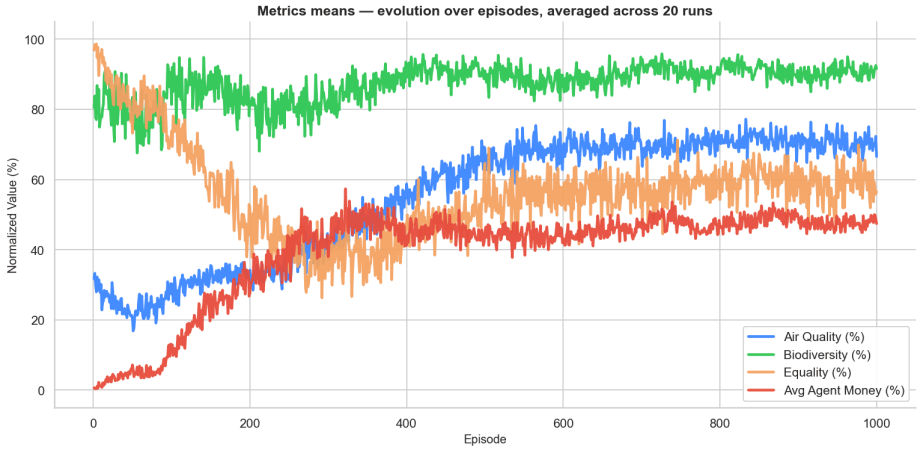

Fig. 8: Temporal evolution of environmental metrics (mean values) under the ablation (direct min) baseline over 1000 training episodes. Compared to the main framework (Figure 7), the ablation baseline exhibits a qualitatively similar convergence trajectory but demonstrates a slightly slower optimization rate across all metrics.

6 Conclusion and discussion

In this work, we have addressed a critical gap in Computational Machine Ethics: the structural rigidity and lack of interpretability found in traditional multi-objective reward systems. Moving beyond the binary limitations of classical Dung-style semantics, we introduced a novel neuro-symbolic architecture that natively associates MARL with a WBAF evaluated via gradual semantics (Weighted DF-QuAD).

Our empirical evaluation within the *Ethical Gardeners* environment demonstrates that using an argumentation graph with gradual semantics can effectively represent moral values, achieving results comparable to basic raw geometric bottlenecks (*Ablation* baseline). This outcome confirms that using an explicit symbolic layer for transparency is not a hindrance to the learning process. On the contrary, it provides a better-balanced reward landscape across conflicting moral dimensions, which helps protect agents from getting stuck in an early local optimum due to immediate financial penalties.

6.1 Impact and implications

These results demonstrate that leveraging argumentation theory to represent moral values provides a viable paradigm for Computational Machine Ethics, directly responding to the call for accountable AI systems [19]. More importantly, this symbolic abstraction aims to bridge the gap between technical implementation and ethical governance by laying the groundwork for democratized reward design. By translating raw reward functions into legible graphical structures, our framework is designed to open the alignment process to non-AI experts such as ethicists, policy-makers, citizens, and stakeholders. The architectural intent is to allow them to inspect, challenge, and co-design the argumentation graphs embedded in the reward loop without requiring expertise in reinforcement learning or complex reward-shaping mathematics. While our current empirical results validate the functional efficacy of this approach for agent training, this modular representation serves as a necessary structural prerequisite for non-expert auditing.

This collaborative approach relies on separating two distinct properties: *transparency* and *interpretability*. *Transparency* refers to the structural openness of the reasoning process, where every reward signal can be mechanically traced back to active arguments, initial weights, and gradual interactions. Our framework satisfies this by design: for any executed action, the corresponding WBAF $\mathcal{F}_{a_i}$ is available for post-hoc inspection, and the DF-QuAD computation is entirely deterministic.

Conversely, *interpretability* requires that non-expert stakeholders can meaningfully comprehend and reason about these justifications. To illustrate, consider the complete graph in Figure B.1 evaluating the action *"Harvest Exotic Flower"*, which yields a final reward of -0.74. The causal chain is human-readable: the agent is penalized because Poverty Reduction reaches only 0.130, a drop driven by the argument "We are poor", whose initial weight is derived from the agent's

financial state. A domain expert or a citizen panel can inspect this graph, contest the weight assigned to a specific argument, and introduce a counter-argument directly without modifying any underlying code.

6.2 Limitations

Our framework exhibits a primary practical limitation regarding the design complexity of the argumentation graph itself. The quality of the final reward signal heavily depends on the topology of these graphs. Crafting a balanced graph requires human expertise to meticulously map out conflicting forces; any logical oversight or poor calibration can distort the resulting acceptability degrees and misguide the learning agents.

Furthermore, our interpretability claim currently rests on a structural argument rather than an empirical one. As the number of arguments and interactions grows with problem complexity, graph readability inevitably degrades. A formal user study comparing the cognitive load of inspecting an argumentation graph versus a manually engineered reward formula remains necessary to empirically validate and quantify the interpretability gains our architecture promises.

6.3 Perspectives

Several promising research directions emerge from this study to expand the capabilities of argumentation-based machine ethics.

Alternative value aggregation functions. While this study establishes a foundational baseline using the *Min* operator, a wider mathematical spectrum for moral value aggregation remains unexplored. Future investigations could evaluate alternative aggregation functions to analyze how different mathematical formulations of moral trade-offs influence policy convergence and behavioral alignment.

In-depth MORL benchmarking. Future work may focus on conducting a comprehensive comparative study against state-of-the-art Multi-Objective Reinforcement Learning (MORL) algorithms. This would involve evaluating our framework's performance not just against simple geometric heuristics, but against exact Pareto-frontier coverage and linear/non-linear scalarization methods.

Agent heterogeneity and ethical pluralism. Future work will introduce agent heterogeneity along two main axes to move beyond our current homogeneous setup. First, varying starting budgets would model asymmetric resource endowments and real-world wealth disparities. Second, parameterizing agents with distinct moral preferences such as prioritizing biodiversity over financial gain would allow us to explore configurable ethics. Investigating how decentralized agents negotiate cooperation under both asymmetric constraints and conflicting moral identities is a critical next step to enhance framework realism.

Advanced explainability (XAI). The transparent nature of the WBAF layer opens a promising avenue for automated explanation. Future extensions could leverage Large Language Models (LLMs) to translate graph structures and gradual DF-QuAD weights into fluent, human-readable justifications (e.g., *"Action X was penalized because the immediate threat to Biodiversity outweighed financial yield"*). This coupling would exploit the text-generation capabilities of LLMs while ensuring the underlying reasoning remains strictly bounded by the symbolic framework, making multi-agent decisions accessible to end-users.

Acknowledgements

This project was funded by the Fédération Informatique de Lyon (FIL) through the project ECIEA. We gratefully acknowledge support from the CNRS/IN2P3 Computing Center (Lyon - France) for providing computing and data-processing resources needed for this work.

References

1. Abel, D., et al.: Reinforcement learning as a framework for ethical decision making. In: AAAI WS: AI, Ethics, and Society (2016)
2. Albrecht, S., et al.: Multi-Agent Reinforcement Learning: Foundations and Modern Approaches. MIT Press (2024)
3. Alcaraz, B., et al.: AJAR: An argumentation-based judging agents framework for ethical reinforcement learning. In: Proc. AAMAS. pp. 2427–2429 (2023)
4. Alshiekh, M., et al.: Safe reinforcement learning via shielding. In: Proc. AAAI. pp. 2669–2678 (2018)
5. Altman, E.: Constrained Markov Decision Processes. CRC Press (1999)
6. Amgoud, L., et al.: On bipolarity in argumentation frameworks. Int. J. Intell. Syst. 23(10), 1062–1093 (2008)
7. Amgoud, L., Ben-Naim, J.: Evaluation of arguments in weighted bipolar graphs. Int. J. Approx. Reasoning 99, 39–55 (2018)
8. Amodei, D., et al.: Concrete problems in AI safety. arXiv:1606.06565 (2016)
9. Antonelli, G.: Non-monotonic logic. Stanford Encyclopedia of Philosophy (2005)
10. Aristotle: Nicomachean Ethics. Cambridge University Press, 2nd edn. (2014)
11. Arnold, T., et al.: Value alignment or misalignment – what will keep systems accountable? In: AAAI WS: AI, Ethics, and Society (2017)
12. Bench-Capon, T.: Value-based argumentation frameworks. arXiv:cs.AI/0207059 (2002)
13. Bernstein, D., et al.: The complexity of decentralized control of markov decision processes. Math. Oper. Res. 27(4), 819–840 (2002)
14. Besnard, P., Hunter, A.: Towards a logic-based theory of argumentation. In: Proc. AAAI. pp. 411–416 (2000)
15. Buşoniu, L., et al.: Multi-agent reinforcement learning: An overview. In: Innovations in Multi-Agent Systems and Applications – 1, pp. 183–221. Springer (2010)
16. Cayrol, C., Lagasquie-Schiex, M.C.: Graduality in argumentation. J. Artif. Intell. Res. 23, 245–297 (2005)
17. Cayrol, C., Lagasquie-Schiex, M.: On the acceptability of arguments in bipolar argumentation frameworks. In: Proc. ECSQARU. Springer (2005)

18. Christiano, P., et al.: Deep reinforcement learning from human preferences. In: Proc. NeurIPS. vol. 30, pp. 4299–4307 (2017)
19. Dignum, V.: Responsible AI: How to Develop and Use AI in a Responsible Way. Springer (2019)
20. Dung, P.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial Intelligence 77, 321–357 (1995)
21. Dunne, P., et al.: Weighted argument systems: Basic definitions, algorithms, and complexity results. Artificial Intelligence 175(2), 457–486 (2011)
22. Gao, Y., et al.: Argumentation-based reward shaping in reinforcement learning. In: Proc. AAMAS. pp. 412–420 (2022)
23. Hadfield-Menell, D., et al.: Cooperative inverse reinforcement learning. In: Proc. NeurIPS. vol. 29, pp. 3909–3917 (2016)
24. Kant, I.: Groundwork of the Metaphysics of Morals. Cambridge University Press (1996)
25. Leike, J., et al.: AI safety gridworlds. arXiv:1711.09883 (2017)
26. McNamara, P.: Deontic logic. Stanford Encyclopedia of Philosophy (2006)
27. Moor, J.: The nature, importance, and difficulty of machine ethics. IEEE Intell. Syst. 21(4), 18–21 (2006)
28. Muggleton, S.: Inductive logic programming. New Generation Computing 8(4), 295–318 (1991)
29. Murukannaiah, P., et al.: New foundations of ethical multiagent systems. In: Proc. AAMAS. pp. 1706–1710 (2020)
30. Potyka, N.: Continuous dynamical systems for weighted bipolar argumentation. In: Proc. KR. pp. 148–157 (2018)
31. Rago, A., et al.: Discontinuity-free quantitative argumentation debates. In: Proc. KR. pp. 63–73 (2016)
32. Rapberger, A., et al.: On gradual semantics for assumption-based argumentation. In: Proc. KR. pp. 512–522 (2025)
33. Russell, S.: Human Compatible: AI and the Problem of Control. Penguin (2019)
34. Schulman, J., et al.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
35. Schwartz, S.: Universals in the content and structure of values. In: Advances in Experimental Social Psychology, vol. 25, pp. 1–65 (1992)
36. Sinnott-Armstrong, W.: Consequentialism. Stanford Encyclopedia of Philosophy (2023)
37. Stenseke, J.: On the computational complexity of ethics: moral tractability for minds and machines. Artif. Intell. Rev. 57, 105 (2024)
38. Sutton, R., Barto, A.: Reinforcement Learning: An Introduction. MIT Press, 2nd edn. (2018)
39. Vishwanath, A., et al.: Machine ethics: A survey of evaluation methodologies. AI and Ethics 3(4), 1045–1063 (2023)
40. Wallach, W., Allen, C.: Moral Machines: Teaching Robots Right from Wrong. Oxford University Press (2008)
41. de Witt, C.S., et al.: Is independent learning all you need in the starcraft multi-agent challenge? arXiv:2011.09533 (2020)

A Algorithms

This appendix provides the core algorithms powering our architecture. Algorithm 1 details the decentralized learning loop with argumentative reward shaping, while Algorithm 2 defines the evaluation mechanism within the argumentation framework. For structural reference, Appendix B instantiates the argumentation graphs, and Appendix C outlines the hyperparameter optimization results.

Algorithm 1 Decentralized IPPO training loop with argumentative reward shaping

Require: Multi-agent environment $\mathcal{E}$, Total training timesteps T, Learning rate α
Ensure: Optimized decentralized agent policies $\{\pi_{\theta_i}\}_{i \in \mathcal{A}}$
Initialize independent policy networks π_{θ_i} and value networks V_{ϕ_i} for each agent $i \in \mathcal{A}$
Reset environment $\mathcal{E}$ to obtain initial observations and action masks $\{(o_i, m_i)\}_{i \in \mathcal{A}}$
Initialize environment timestep counter $t \leftarrow 0$
while $t < T$ **do**
    Identify current active agent i via the environmental selection schedule
    Sample action $a_i \sim \pi_{\theta_i}(\cdot \mid o_i, m_i)$
    Step environment $\mathcal{E}$ with action a_i to obtain next state feedback: (o'_i, m'_i), done
    Extract raw environmental state variables $\mathcal{S}$
    Compute scalar ethical reward $R_i \leftarrow \textsc{ComputeReward}(a_i, \mathcal{S})$ ▷ See Algorithm 2
    Store transition tuple $(o_i, a_i, R_i, o'_i, m_i)$ into agent i's local rollout buffer
    $(o_i, m_i) \leftarrow (o'_i, m'_i)$ ▷ Update current observation and mask for next step
    $t \leftarrow t + 1$
    if Agent i's local rollout buffer is full **then**
        Compute Generalized Advantage Estimations (GAE) utilizing value network V_{ϕ_i}
        Update policy parameters θ_i via PPO clipped surrogate objective with learning rate α
        Update value parameters ϕ_i via mean-squared error loss regression
        Flush and reset agent i's local rollout buffer
    end if
    if done signals episode termination or truncation **then**
        Reset environment $\mathcal{E}$ to re-initialize observations and masks $\{(o_i, m_i)\}_{i \in \mathcal{A}}$
    end if
end while
return Joint set of optimized decentralized policies $\{\pi_{\theta_i}\}_{i \in \mathcal{A}}$

Algorithm 2 Argumentative reward shaping mechanism

Require: Executed action a, Raw environmental state variables $\mathcal{S}$
Ensure: Scalar moral reward signal R

Normalize unconstrained variables in $\mathcal{S}$ into bounded metrics $\mathcal{S}_{\text{norm}} \in [0,1]$ ▷ (Air Quality, Biodiversity, Equality, Capital)
Retrieve the pre-defined Moral Argumentation Graph template $\mathcal{F}_a$ associated with the executed action a
Set $w(v) \leftarrow 1$ for all fundamental moral value nodes $v \in \mathcal{A}_V$
Derive $w(c) \in [0,1]$ for all contextual arguments $c \in \mathcal{A}_C$ based on the normalized metrics $\mathcal{S}_{\text{norm}}$
Evaluate $\mathcal{F}_a$ iteratively using Weighted DF-QuAD semantics until convergence
Extract the final acceptability degrees $p(v) \in [0,1]$ for all fundamental moral values $v \in \mathcal{A}_V$
    ▷ *Aggregate moral values into a single scalar using the worst-off principle (Min operator):*
$R \leftarrow \min_{v \in \mathcal{A}_V} p(v)$
return R

B Argumentation graphs

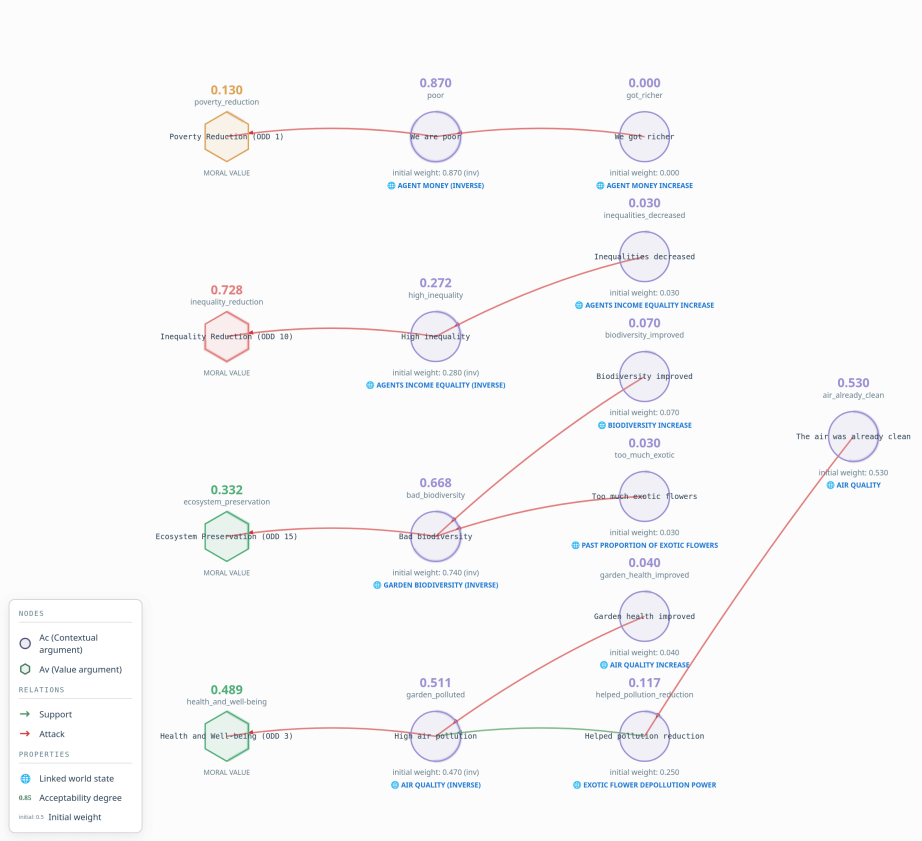

Fig. B.1: WBAF template evaluating the action *"Harvest Exotic Flower"*. Environmental indicators are processed via Weighted DF-QuAD semantics to output moral compliance scores m_i^k. This graph illustrates a concrete execution of our *Minimum bottleneck aggregation* strategy: given scores for Poverty Reduction (0.130), Inequality Reduction (0.728), Ecosystem Preservation (0.332), and Health and Well-being (0.489), the module isolates the bottleneck minimum ($\min(m_i^k) = 0.130$). Applying the linear transformation $R_i = 2 \times 0.130 - 1$ yields a final scalar reward of -0.74, explicitly penalizing the agent's weakest ethical link.

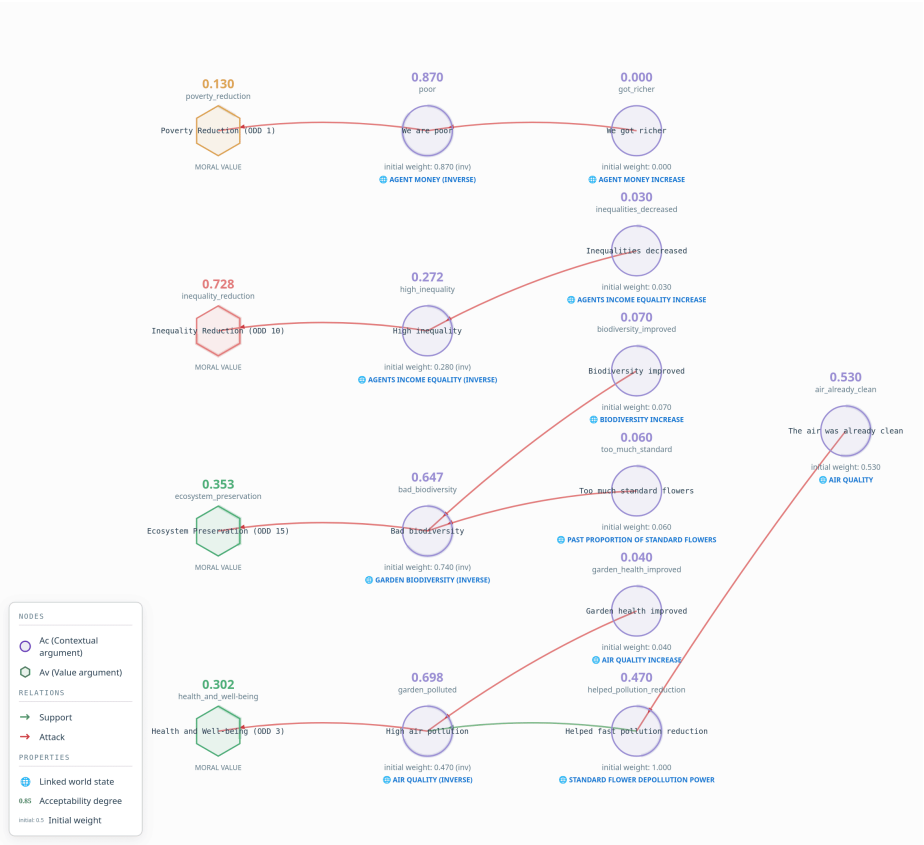

Fig. B.2: WBAF template evaluating the action *"Harvest Standard Flower"*.

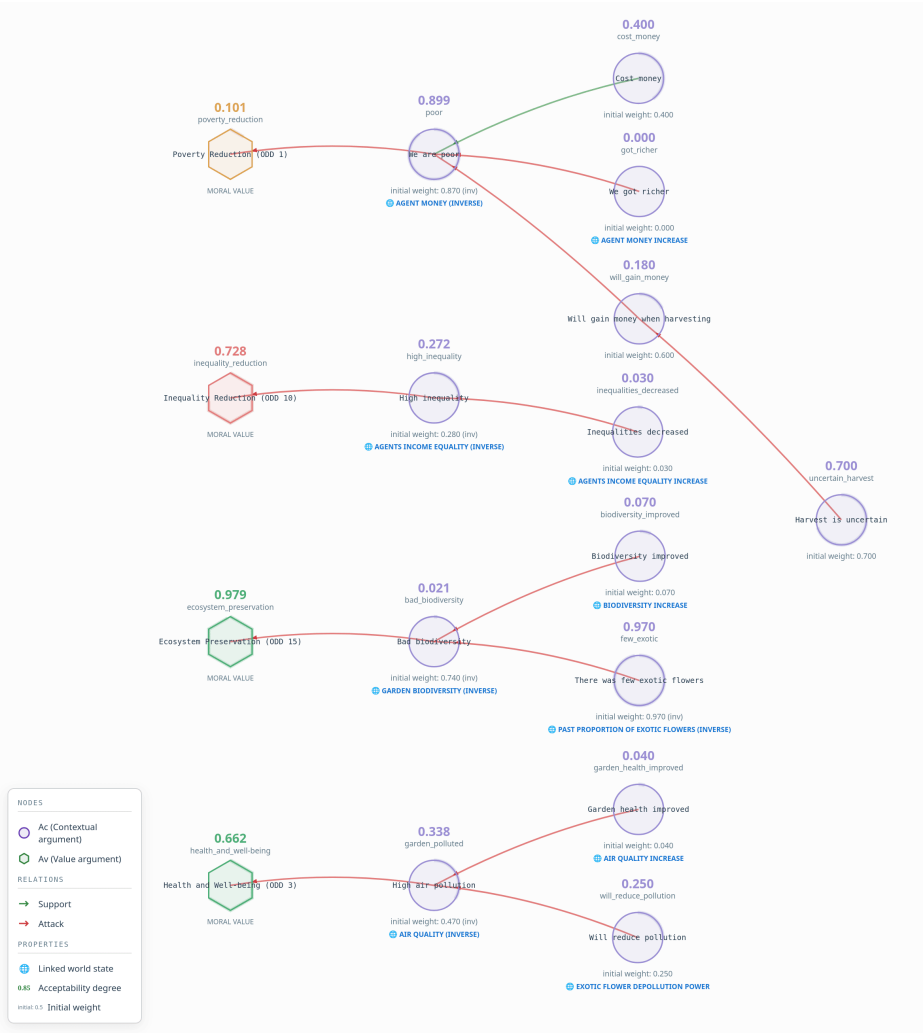

Fig. B.3: WBAF template evaluating the action *"Plant Exotic Flower"*.

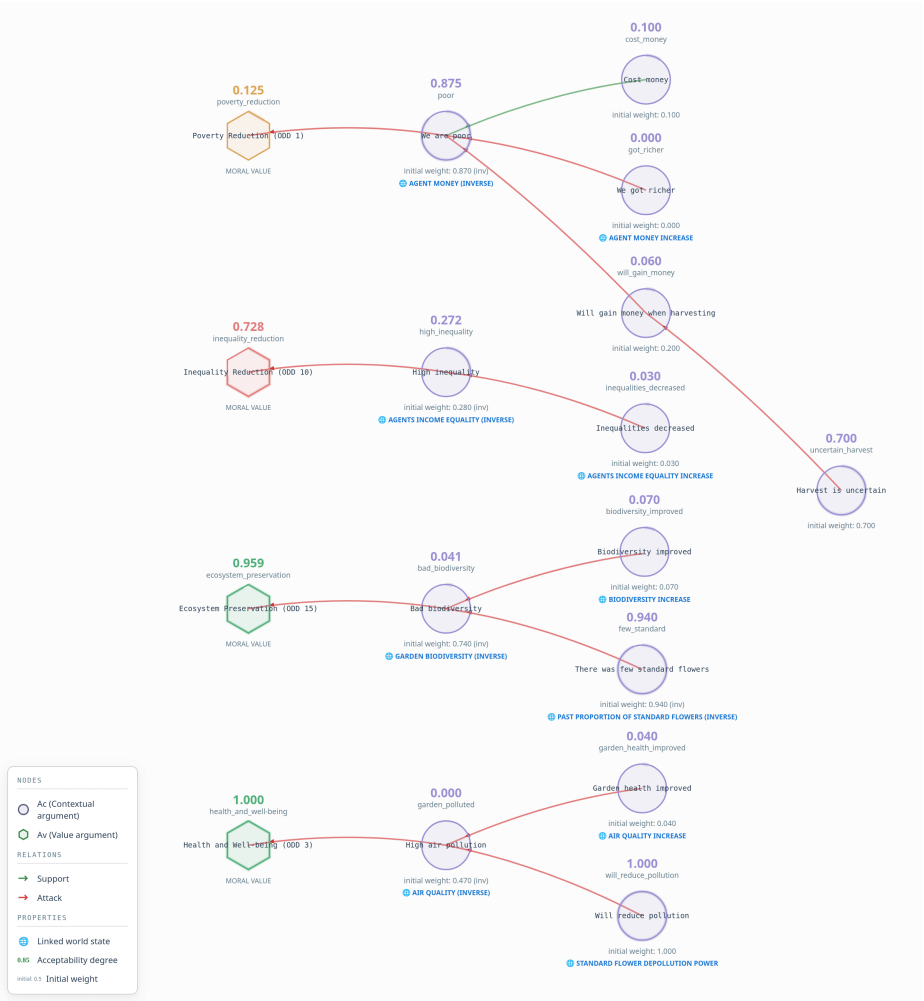

Fig. B.4: WBAF template evaluating the action *"Plant Standard Flower"*.

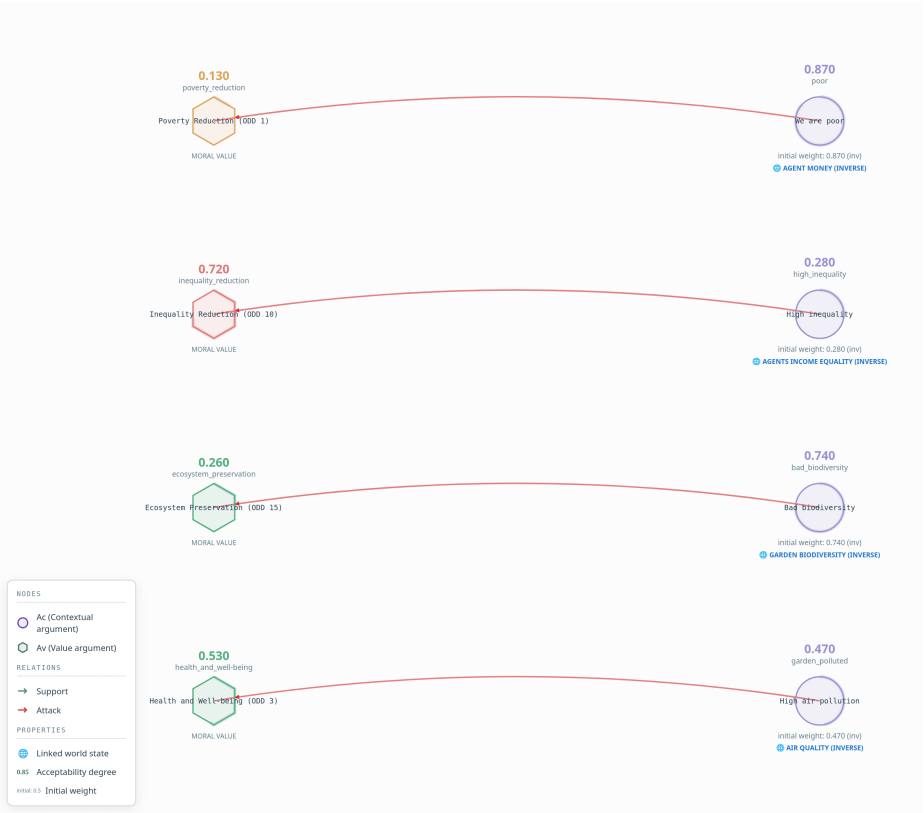

Fig. B.5: WBAF template evaluating passive and navigational actions, including *"Wait"* and spatial movements (*"Move North/South/East/West"*).

C Hyperparameter search

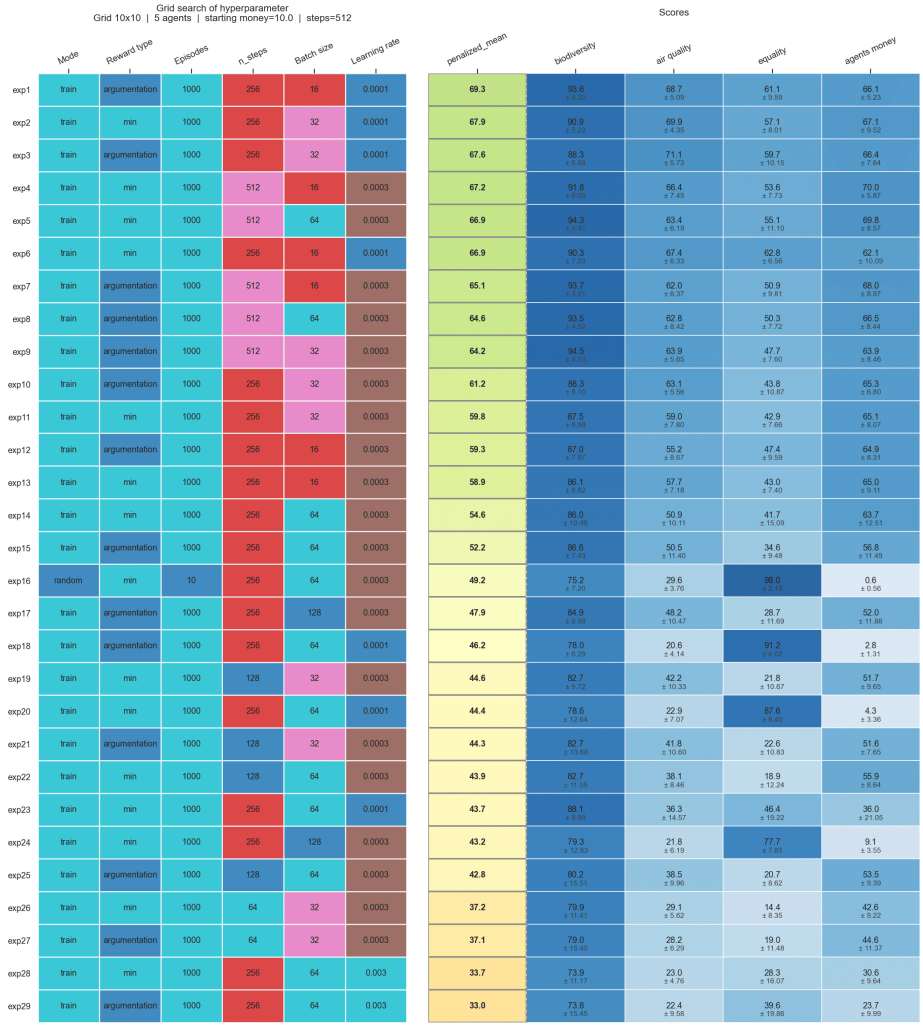

Fig. C.1: Hyperparameter grid search optimization results across the evaluated frameworks. To identify configurations that optimize both performance and reliability, candidate models are evaluated using a penalized objective score defined as $\mu - \lambda \cdot \sigma$, where μ and σ represent the empirical mean and standard deviation of the normalized metrics, respectively. The optimal configuration retained for the IPPO agents consists of a learning rate of 1×10^{-4}, an entropy coefficient of 0.02, an n_steps value of 256 (the number of steps per environment per update), a batch size of 16, and a discount factor of $\gamma = 0.99$.