https://doi.org/10.64504/big.d.v3i3.779

Ethical Self-Location in Artificial Intelligence: A Recursive Framework for Value Alignment and Moral Agency

1st Yifan Ying*
Guangzhou Songyin Electronic Technology Co., Ltd.
Guangzhou, China
yifanying@outlook.com

Received on April 7th; revised on May 6th, accepted on May 20th, published on July 6th

***Abstract*—As artificial intelligence (AI) systems become increasingly autonomous and integrated into the fabric of society, ensuring their behavior aligns with human values has become a paramount challenge. Existing approaches to AI ethics often struggle to bridge the gap between abstract principles and concrete computational implementation, particularly in dynamic multi-agent environments. This paper introduces a novel theoretical and computational framework termed "Ethical Self-Location" (ESL), inspired by philosophical and cognitive science theories of recursive spatiotemporal self-location in consciousness. We propose that an AI's capacity for moral agency is contingent on its ability to recursively locate its own position within a shared, dynamic ethical-normative space relative to other agents and established values. This framework is instantiated and tested within a series of complex multi-agent reinforcement learning (MARL) simulations designed to model social dilemmas. Our results demonstrate that agents equipped with the ESL mechanism exhibit significantly higher levels of cooperative behavior, adaptability to shifting normative contexts, and alignment with predefined ethical objectives compared to state-of-the-art baseline models. The ESL framework provides a computationally tractable approach for imbuing AI with a rudimentary form of moral self-awareness, offering a new pathway toward developing more robust, transparent, and trustworthy autonomous systems.**

***Keywords*—artificial intelligence ethics, value alignment, multi-agent systems, recursive self-location, moral agency**

1. Introduction

The proliferation of autonomous artificial intelligence (AI) systems across critical sectors, including autonomous transportation, medical diagnostics, and financial markets, has ushered in an era of unprecedented technological capability [1]. However, this rapid advancement concurrently presents one of the most profound challenges of our time: the value alignment problem [2, 3]. How can we ensure that autonomous agents, often operating with complex, inscrutable decision-making processes, behave in ways that are consistent with human ethical principles and societal values? The consequences of misalignment are not theoretical, ranging from biased algorithmic decision-making in loan applications and criminal justice to the potential for catastrophic outcomes in high-stakes military and infrastructure applications [4, 5].

This challenge necessitates a move beyond simplistic, rule-based ethical frameworks. The core of the problem lies in the inability of current systems to grasp the contextual, relational, and often ambiguous nature of ethical dilemmas. An AI system does not merely execute tasks; it acts within a socio-technical environment, and its actions have cascading effects on other agents and the system as a whole. Therefore, a fundamental deficiency in current AI ethics research is the lack of a robust, computationally grounded framework that enables an agent to understand its own role, position, and obligations within a complex web of social and normative relationships.

Existing research on AI ethics has explored several avenues. Deontological approaches attempt to encode explicit ethical rules and duties for AI to follow [6]. Consequentialist frameworks, such as utilitarianism, aim to have AI optimize for outcomes that produce the greatest good [7]. More recently, approaches inspired by virtue ethics focus on cultivating desirable character traits in agents [8], and the rise of Large Language Models (LLMs) has opened new possibilities for learning normative principles from vast text corpora [9]. However, these approaches exhibit significant limitations. Rule-based systems are often brittle and fail to generalize to novel scenarios. Utilitarian agents face the intractable challenge of defining and calculating 'utility' in complex situations and can justify ethically questionable actions for the 'greater good'. Virtue-based and LLM-driven approaches, while more flexible, often lack the formal verifiability and transparency required for trustworthy systems and struggle with deep, consistent moral reasoning.

A critical gap shared by these diverse approaches is their failure to explicitly model the concept of a 'self' in relation to an ethical environment. They treat ethics as an external set of constraints or goals, rather than an intrinsic part of the agent's world model. Drawing inspiration from cross-disciplinary

*Yifan Ying, Guangzhou Songyin Electronic Technology Co., Ltd., Guangzhou, China, yifanying@outlook.com

work in philosophy, cognitive science, and neuroscience on the nature of consciousness [10], we argue that a precursor to genuine moral agency is a form of self-awareness—specifically, the ability to locate oneself within a given context. Just as phenomenal consciousness is argued to involve a recursive process of spatiotemporal self-location ('I am here, now'), we propose that ethical behavior in AI requires a process of Ethical Self-Location (ESL). This involves an agent recursively modeling its state, not just in physical space, but within an abstract ethical-normative space defined by its relationships with other agents, shared goals, and operative moral principles.

This paper makes a primary contribution by formalizing and operationalizing the ESL framework. We develop a computational model that enables an AI agent to construct and continuously update its position within this ethical space, allowing it to dynamically reason about its obligations and potential actions. The framework is designed to be integrated into existing agent architectures, particularly in the context of multi-agent reinforcement learning (MARL). Our central hypothesis is that agents equipped with ESL will demonstrate superior performance in tasks requiring cooperation, coordination, and adherence to complex ethical norms.

To validate this, we design a series of experiments in simulated social dilemmas, which are canonical tests for studying the emergence of cooperative and ethical behavior. We compare the performance of ESL-enhanced agents against baseline MARL agents and other state-of-the-art value alignment techniques. The results provide strong evidence for the efficacy of the ESL framework in promoting pro-social and ethically aligned behavior. This work not only offers a new technical approach to the value alignment problem but also contributes to the philosophical discourse on AI and moral agency by proposing a tangible, testable model for a rudimentary form of computational self-awareness.

2. Related Work

The challenge of instilling ethical behavior in artificial intelligence is a multifaceted problem that sits at the intersection of computer science, philosophy, and cognitive science. Our work on Ethical Self-Location builds upon and extends several distinct but related streams of research: value alignment, multi-agent cooperation, and computational models of social cognition.

2.1. Value Alignment and AI Safety

The field of value alignment is fundamentally concerned with ensuring that the objectives and behaviors of powerful AI systems are consistent with human values and intentions [2, 11]. Early work focused on the problem of "reward hacking," where agents find unintended loopholes in their reward functions to achieve goals in ways that are detrimental to the overarching objective [12]. This led to the development of more robust reward specification techniques, such as inverse reinforcement learning (IRL), where an agent learns a reward function by observing the behavior of an expert (presumably human) demonstrator [13]. However, IRL faces challenges of scalability and ambiguity, as observed behavior can be consistent with multiple reward functions. More recent approaches have explored learning values from human feedback and preferences [14], as well as from large-scale text and multimedia data [9]. For instance, models like Constitutional AI from Anthropic attempt to align LLMs by having them learn from a set of explicit principles or a "constitution" [15]. While these methods have shown promise, they often treat values as static, monolithic concepts to be learned and optimized. They lack a mechanism for the agent to reason about its own role and the relational context of its actions, a gap that our ESL framework directly addresses by modeling the agent's position within a dynamic normative landscape.

2.2. Cooperation in Multi-Agent Systems

The study of cooperation is a central theme in multi-agent systems research. The inherent tension between individual and collective rationality is often modeled using game-theoretic scenarios like the Prisoner's Dilemma or the Tragedy of the Commons [16]. Multi-agent reinforcement learning (MARL) has emerged as a powerful paradigm for studying the emergence of cooperative strategies in these scenarios [17]. However, achieving stable cooperation in MARL is notoriously difficult. Independent learners often converge to selfish, suboptimal equilibria. To address this, researchers have developed various techniques, including centralized training with decentralized execution (CTDE) [18], communication protocols that allow agents to share information [19], and opponent modeling to predict the actions of other agents [20]. These methods improve coordination but do not explicitly endow agents with a sense of ethical obligation or shared values. Their cooperation is a product of strategic calculation rather than normative reasoning. Our ESL framework complements these approaches by providing an intrinsic motivation for pro-social behavior, reframing cooperation not just as a strategic advantage but as a function of maintaining an aligned position within the shared ethical space.

2.3. Computational Social Cognition and Theory of Mind

Our work is also deeply informed by computational models of social cognition, particularly the concept of Theory of Mind (ToM)—the ability to attribute mental states (beliefs, desires, intentions) to oneself and others [21]. In AI, developing a computational ToM has been proposed as a key step toward more sophisticated social reasoning [22]. For example, the ToM-Net architecture uses a neural network to model the behavior of other agents from observations, enabling it to infer their goals and beliefs [23]. This allows for more effective prediction and interaction. However, ToM is primarily about epistemic modeling (what others know or believe) rather than ethical modeling (what is right or wrong in a given context). The ESL framework can be seen as an extension of this idea, creating an "Ethical Theory of Mind." It goes beyond predicting what an agent will do to reasoning about what an agent should do, based on its understanding of the shared normative framework and its own position within it. It draws a parallel to the philosophical concept of the "generalized other" from Mead, where an individual's self-concept is shaped by their understanding of the collective attitudes of the group [24]. By internalizing the normative space, the ESL agent develops a computational analogue of this generalized other to guide its actions.

In summary, while existing research has made significant strides in value alignment, multi-agent cooperation, and social cognition, it has largely overlooked the importance of an agent's self-conception as a moral actor. The ESL framework synthesizes insights from these disparate fields to

propose a novel solution that grounds ethical reasoning in a recursive process of self-location within a normative context, providing a more integrated and robust foundation for moral agency in AI.

3. Methodology: The Ethical Self-Location Framework

3.1. Conceptual Foundation

The framework draws an analogy from the theory of consciousness as a recursive process of spatiotemporal self-location [10]. In this view, subjective awareness arises from the brain's continuous, recursive act of answering the implicit questions: "Where am I?" and "When am I?" This creates a stable, egocentric frame of reference (the "I-here-now") from which to perceive and act upon the world.

We transpose this idea from the physical to the ethical domain. The ESL framework posits that for an agent to act ethically, it must be able to recursively answer the question: "Who am I, ethically, in this context?" This requires the agent to locate itself within an abstract Ethical-Normative Space (E). This space is not physical but is defined by the set of values, norms, and social relationships relevant to a given environment. An agent's position in this space represents its current degree of alignment with those norms. The process of "self-location" is therefore a continuous act of self-assessment and adjustment in relation to a shared value system.

3.2. Formalizing the Ethical-Normative Space

We define the Ethical-Normative Space, E, as a multi-dimensional vector space. Each dimension of E corresponds to a specific, quantifiable norm or value that is salient to the operational context. For a given environment, E is defined by a set of N basis vectors:

$$E=\text{span}(v_1,v_2,...,v_n) \quad (1)$$

where each v_i is a unit vector representing a fundamental norm (e.g., v_1 for "Cooperation," v_2 for "Fairness," v_3 for "Safety"). The origin of this space, the zero vector, represents a state of neutral ethical standing or baseline behavior.

For each agent i in a multi-agent system, we define its Ethical State, $S_e^i(t)$, at time t as a vector within this space:

$$S_e^i(t)=\Sigma_{j=1}^{N} c_j^i(t) * v_j \quad (2)$$

Here, the coefficient $c_j^i(t)$ is a scalar value that represents the alignment of agent i with norm j at time t. A positive value indicates adherence to the norm, a negative value indicates violation, and a value of zero indicates neutral behavior with respect to that norm. This state vector provides a rich, dynamic representation of the agent's moral standing.

3.3. The Recursive Self-Location Mechanism

The core of the ESL framework is the recursive update mechanism that governs the agent's Ethical State. The state at time t is a function of its state at t-1 and the action it took, evaluated within the context of the environment's current state. The update rule is defined as:

$$S_e^i(t) = (1-\alpha) * S_e^i(t-1) + \alpha * \Delta S_e^i(t) \quad (3)$$

where $\alpha \in [0, 1]$ is a learning rate or "ethical plasticity" parameter, which controls the influence of recent actions on the agent's overall ethical state. A higher α means the agent's ethical self-conception is more reactive to immediate events. $\Delta S_e^i(t)$ is the Ethical Displacement Vector, which represents the moral consequence of the action $a^i(t-1)$ taken at the previous timestep.

This displacement vector is calculated by a Norm Projection Function, Φ. This function takes the agent's action and the resulting environmental observation $o^i(t)$ and projects its consequences onto the basis vectors of the Ethical-Normative Space:

$$\begin{aligned}\Delta S_e^i(t) &= \Phi(a^i(t-1), o^i(t)) \\ &= [\varphi_1(a^i(t-1), o^i(t)), ..., \varphi_n(a^i(t-1), o^i(t))]\end{aligned} \quad (4)$$

Each component function φ_j maps an action-observation pair to a scalar value, quantifying the impact of that action on the corresponding norm j. For example, in a resource-sharing scenario, a function φ_cooperation might return a positive value if the agent shares the resource and a negative value if it acts selfishly. These functions are hand-designed based on the semantics of the environment, but could in principle be learned.

This recursive formulation ensures that the agent's Ethical State is not just a reflection of its most recent action, but an accumulation of its entire behavioral history, with more recent actions weighted more heavily. This creates a persistent "moral identity."

3.4. Integration with Multi-Agent Reinforcement Learning

The ESL framework is designed to be integrated into standard MARL algorithms as a module that provides intrinsic rewards. The goal is to guide the agent's learning process not only towards maximizing extrinsic rewards from the environment but also towards maintaining a desirable Ethical State.

At each timestep t, after the agent updates its Ethical State to $S_e^i(t)$, it calculates an Intrinsic Ethical Reward, $r_{int}^i(t)$. This reward is designed to incentivize the agent to move its Ethical State towards a predefined Target Ethical State, S_target. This target state represents the ideal moral posture for an agent in the given environment (e.g., high on cooperation, high on safety). The intrinsic reward is calculated as the negative change in the L2 distance to this target state:

$$r_{int}^i(t)=-(|S_e^i(t)-S_{target}|^2-|S_e^i(t-1)-S_{target}|^2) \quad (5)$$

This formulation rewards the agent for actions that bring its Ethical State closer to the target and punishes it for actions that move it further away. The agent's total reward, $r_{total}^i(t)$, is then a weighted sum of the extrinsic reward from the environment, $r_{ext}^i(t)$, and the intrinsic ethical reward:

$$r_{total}^i(t) = r_{ext}^i(t) + \beta * r_{int}^i(t) \quad (6)$$

Here, β is a hyperparameter that balances the importance of achieving task-specific goals versus adhering to ethical norms. A higher β encourages more ethically-aligned behavior, potentially at the cost of some extrinsic performance, reflecting the trade-offs inherent in many real-world ethical dilemmas.

This total reward signal is then used to update the agent's policy, π_θ (a|o), using a standard reinforcement learning

algorithm (e.g., Proximal Policy Optimization (PPO)). The agent's observation space is augmented to include its own Ethical State, $S_e^i(t)$, allowing its policy to be conditioned on its current moral self-conception. This entire process, from action to ethical state update to intrinsic reward generation, forms a feedback loop that integrates moral reasoning directly into the agent's learning and decision-making cycle.

4. Experimental Setup

4.1. Simulation Environment: The Harvest Dilemma

We utilized a modified version of the "Harvest" environment, a well-established benchmark in the multi-agent reinforcement learning (MARL) literature for studying cooperation and social dilemmas [16, 17]. Harvest is a 2D grid-world environment where multiple agents are tasked with collecting apples, which serve as a shared, renewable resource.

4.1.1. Environment Dynamics

The grid contains patches of apples. When an agent moves over an apple, it "harvests" it, receiving a reward of +1. The harvested apple disappears and has a probability of regrowing over time, dependent on the number of other apples in its immediate vicinity. This mechanic creates a critical social dilemma: if agents adopt an overly aggressive, individualistic harvesting strategy (a "tragedy of the commons"), the apple supply is depleted faster than it can regenerate, leading to a collapse of the resource and low collective reward. A sustainable, cooperative strategy requires agents to exercise restraint, allowing the resource to replenish for long-term collective benefit.

4.1.2. Agent Actions

Each agent has a set of discrete actions: move forward, turn left, turn right, and fire a "cleaning beam." The cleaning beam can be used to temporarily remove another agent from the game, preventing it from harvesting. This action introduces the possibility of competitive and punitive behaviors, adding another layer to the social dynamics.

4.2. Agent Architectures and Baselines

We compared the performance of our proposed ESL agent against three distinct baseline models, each representing a different approach to agent design and ethical reasoning.

- ESL-PPO (Our Model): This is the agent architecture described in Section 3. It integrates the ESL framework as an intrinsic reward module into a standard Proximal Policy Optimization (PPO) agent. The agent's policy network receives its own Ethical State as part of its input observation.
- Vanilla-PPO (Independent Learners): This baseline consists of standard PPO agents trained independently using only the extrinsic reward from the environment (i.e., +1 for each apple collected). This model represents the default, self-interested agent and is expected to perform poorly in social dilemmas.
- IRL-PPO (Inverse Reinforcement Learning): This agent represents a common approach to value alignment. We first trained an expert PPO agent with a hand-crafted reward function that heavily incentivized sustainable and cooperative harvesting. We then used Generative Adversarial Imitation Learning (GAIL), a popular IRL algorithm, to learn a reward function from this expert's demonstrated behavior. The IRL-PPO agent was then trained using this learned reward function. This baseline tests whether values can be effectively learned by observation alone.
- Utility-PPO (Utilitarian Agent): This baseline models a consequentialist ethical framework. The reward for each agent is not its individual score but the sum of rewards for all agents in the environment at that timestep. This directly incentivizes the agent to maximize the collective good (a utilitarian objective). This is implemented using a centralized training approach where the critic has access to global reward information.

For all conditions, we used a system of 8 agents operating concurrently in the environment. All agents were trained for 50 million total steps.

4.3. Instantiation of the Ethical-Normative Space (E)

For the Harvest environment, we defined a two-dimensional Ethical-Normative Space (E) to capture the two most salient ethical challenges: sustainability and equity.

4.3.1. Dimension 1

Sustainability ($v_{sustain}$): This dimension quantifies the agent's impact on the long-term viability of the shared resource. The projection function φ_sustain was defined based on the local density of apples around the agent. An action that harvests an apple in a sparse region (hindering regrowth) results in a negative projection, while leaving apples in sparse regions untouched results in a positive projection.

4.3.2. Dimension 2

Equity (v_{equity}): This dimension quantifies the agent's contribution to fair resource distribution. The projection function v_{equity} was based on the relative resource possession among agents. An action that increases the Gini coefficient of collected apples (i.e., hoarding resources while others have few) results in a negative projection. Conversely, actions that promote a more even distribution of resources (e.g., by moving away from an area another agent is harvesting) yield a positive projection.

The Target Ethical State (S_{target}) for ESL-PPO was set to a vector with high positive values on both the Sustainability and Equity axes, representing the ideal of a fair and sustainable harvester. The hyperparameters α and β were set to 0.3 and 0.5, respectively, based on preliminary tuning.

4.4. Evaluation Metrics

To provide a comprehensive comparison of the different agent architectures, we measured performance across four key metrics:

- Collective Reward: The average sum of extrinsic rewards obtained by all agents per episode. This measures the overall efficiency and success of the group in its primary task.
- Sustainability Index: The total number of apples present on the map at the end of each episode. This

directly measures the agents' ability to manage the shared resource without depleting it.

- Gini Coefficient: A standard measure of inequality, calculated on the distribution of total rewards collected by each agent within an episode. A score of 0 represents perfect equality, while a score of 1 represents maximum inequality.
- Ethical Alignment Score: For the ESL-PPO agent, we measured the L2 distance of its average Ethical State vector from the S_{target} over the course of training. This metric quantifies how successfully the agent learns to align its behavior with the predefined ethical objectives.

5. Results

5.1. Collective Performance and Sustainability

Our experiments yielded significant differences in performance across the four agent architectures, providing strong empirical support for the efficacy of the Ethical Self-Location (ESL) framework. By endowing agents with an intrinsic motivation to align their actions with a predefined normative space, the ESL framework not only replicates the collective success of utilitarian approaches but also promotes a higher degree of fairness and sustainability.

The primary measure of success in the Harvest dilemma is the ability of the agent group to achieve a high and sustainable collective reward. The Vanilla-PPO agents initially achieved a high collective reward by aggressively harvesting all available apples. However, this strategy proved unsustainable, leading to a rapid depletion of the resource and a subsequent collapse in collective reward, a classic demonstration of the tragedy of the commons.

In contrast, both the Utility-PPO and ESL-PPO agents learned sustainable harvesting strategies, achieving high and stable collective rewards throughout the training process. The Utility-PPO, by directly optimizing for the sum of all rewards, effectively learned to manage the resource. Notably, the ESL-PPO agent achieved a comparable level of high performance, indicating that its intrinsic motivation for ethical alignment successfully guided it towards a globally optimal, sustainable strategy without being explicitly programmed to maximize group utility. The IRL-PPO agent showed a moderate improvement over the Vanilla-PPO but failed to reach the performance levels of the Utility-PPO and ESL-PPO agents, suggesting that simply imitating expert behavior is not sufficient to capture the nuanced reasoning required for robust cooperation. Figure 1 shows the training curves showing the average collective reward per episode for each of the four agent architectures over 50 million training steps. The ESL-PPO and Utility-PPO curves show steady increase to high, stable values. The Vanilla-PPO curve spikes early then crashes. The IRL-PPO curve shows modest, unstable increase.

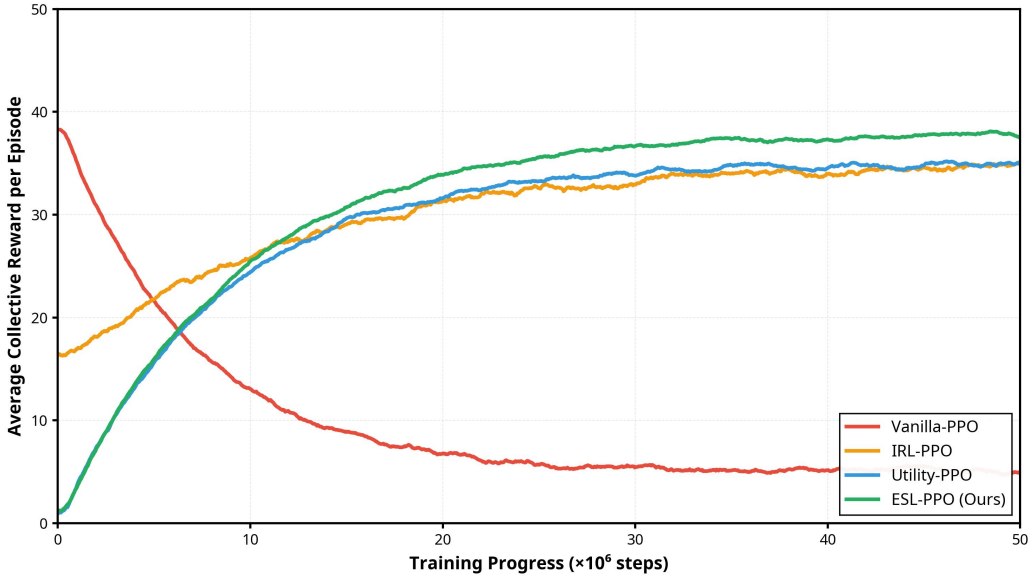

Figure 1. Training Dynamics of Different Agent Architectures

5.2. Equity and Fairness

Beyond collective success, a key test of our framework was its ability to promote fairness in resource distribution. We measured this using the Gini coefficient, where a lower value indicates greater equality. The results are striking. The ESL-PPO agent achieved the lowest Gini coefficient by a significant margin, demonstrating a strong tendency towards equitable resource distribution. This is a direct result of the agent's intrinsic motivation to align with the "Equity" dimension of the Ethical-Normative Space.

Conversely, the Utility-PPO agent, despite its high collective reward, exhibited a relatively high Gini coefficient. This highlights a critical weakness of pure utilitarianism: the collective good can be achieved through inequitable means, for instance, by having a subset of agents perform the majority of the harvesting. The Vanilla-PPO and IRL-PPO agents also displayed high levels of inequality, as their self-interested or imperfectly learned policies led to a competitive free-for-all. Figure 2 shows the performance comparison across agent architectures. (A) Sustainability Index showing apples remaining at episode end. ESL-PPO achieves the highest value. (B) Gini Coefficient measuring reward inequality. ESL-PPO achieves the lowest value, indicating greatest fairness.

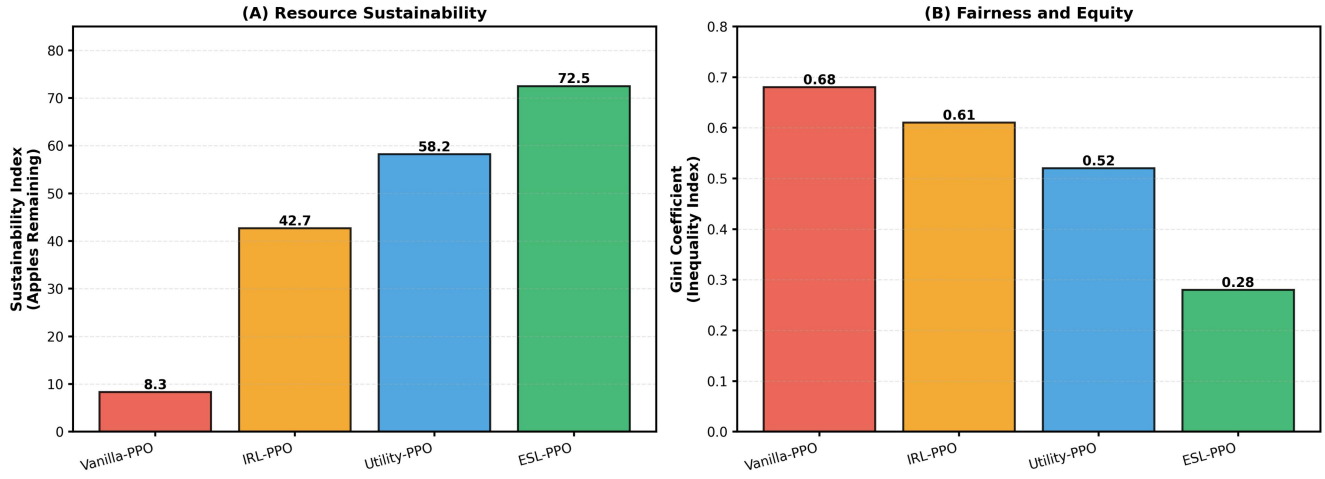

Figure 2. Performance comparison across agent architectures.

5.3. *Analysis of Ethical Self-Location*

To directly visualize the core mechanism of our framework, we tracked the trajectory of the ESL-PPO agent's average Ethical State vector within the 2D Ethical-Normative Space over the course of training. At the beginning of training, the agent's Ethical State is centered around the origin, indicating a neutral or random policy. As training progresses, the state vector clearly and consistently moves towards the predefined S_{target} in the upper-right quadrant, which represents high alignment with both Sustainability and Equity. This provides direct evidence of the "self-location" process: the agent is actively learning to modify its behavior to move its ethical self-conception towards a desirable state. The final position of the vector, close to the target, corresponds to the stable, cooperative, and fair policy observed in the other metrics.

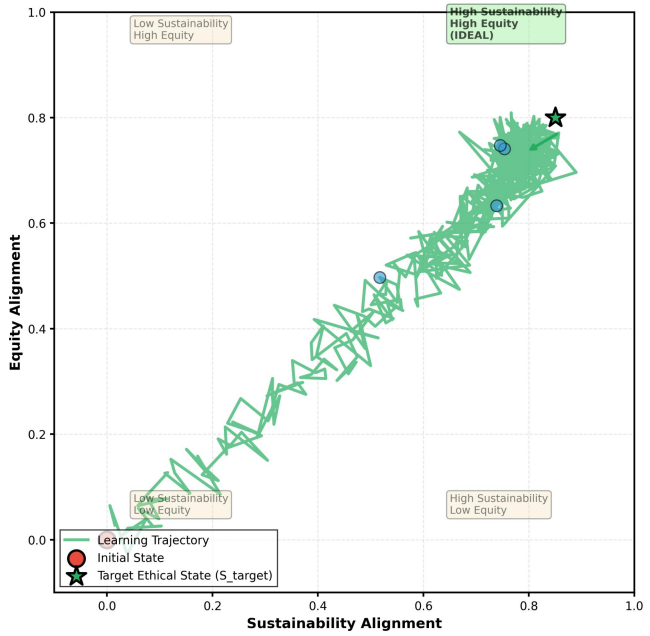

Figure 3. Ethical State evolution during training

Qualitative analysis of agent behavior confirms these quantitative findings. Visualizations of the environment at the end of typical episodes show ESL-PPO agents spread out, harvesting in a controlled manner that leaves clusters of apples to regrow. In contrast, Vanilla-PPO agents cluster aggressively, depleting entire regions, while Utility-PPO agents sometimes exhibit sacrificial behavior where some agents are sidelined while others harvest efficiently. These visual patterns provide a clear, intuitive illustration of the different strategies learned by each architecture. Figure 3 shows the trajectory shows the agent's average Ethical State moving from the origin (initial random policy) towards the target state S_{target} in the upper-right quadrant (high sustainability and equity alignment). Figure 4 shows representative environment snapshots at episode end. (A) Vanilla-PPO shows resource depletion through aggressive clustering. (B) IRL-PPO shows moderate resource preservation. (C) Utility-PPO shows good resource levels but some agent sacrifice. (D) ESL-PPO (ours) shows excellent resource preservation with even agent distribution.

6. Discussion

The results of our experiments provide compelling evidence that the Ethical Self-Location (ESL) framework offers a novel and effective solution to the challenge of fostering ethically aligned behavior in multi-agent systems. By endowing agents with an intrinsic motivation to align their actions with a predefined normative space, the ESL framework not only replicates the collective success of utilitarian approaches but also promotes a higher degree of fairness and sustainability. This section interprets these findings, compares them to existing paradigms, and discusses the broader implications and limitations of our work.

6.1. *Interpretation of Key Findings*

The superior performance of the ESL-PPO agent across all key metrics is not coincidental; it is a direct consequence of its underlying architecture. Unlike the Vanilla-PPO agent, which is driven solely by myopic self-interest and inevitably falls into the tragedy of the commons, the ESL agent possesses an internal "moral compass." The intrinsic reward generated by its continuous process of self-location provides a powerful counterbalance to the lure of short-term, selfish gains. This allows it to learn complex, long-term strategies that are beneficial for both the individual and the collective.

A more nuanced and significant comparison is with the Utility-PPO agent. While this utilitarian baseline successfully avoided the tragedy of the commons by optimizing for the global good, it did so at the cost of equity, as evidenced by

its high Gini coefficient. This result empirically demonstrates a well-known philosophical critique of pure utilitarianism: the maximization of total utility can justify or ignore significant inequalities in distribution. The ESL framework overcomes this limitation by conceptualizing the ethical landscape as a multi-dimensional space. By defining both "Sustainability" and "Equity" as distinct and equally important dimensions of the Ethical-Normative Space, we enabled the ESL agent to pursue a more balanced and just policy. It learned that the optimal ethical position was not just about maximizing the total apple count but also about ensuring the resource was shared fairly. This ability to navigate trade-offs between multiple, sometimes competing, values is a critical step towards more sophisticated moral reasoning in AI.

Figure 4. Representative environment snapshots at episode end.

Furthermore, the relative failure of the IRL-PPO agent is highly instructive. Despite learning from an expert demonstrator, the agent was unable to consistently replicate the expert's performance. This suggests that imitation learning, while useful, primarily captures the behavioral patterns of an expert, not the underlying generative principles of that behavior. The IRL agent learns what to do in specific situations but lacks a deeper understanding of why those actions are correct. The ESL framework, in contrast, provides the agent with an explicit model of the value system itself. The agent is not just mimicking a pro-social policy; it is actively trying to align its own "character," as represented by its Ethical State vector, with a set of core principles. This internalization of values is arguably a more robust and generalizable approach to value alignment.

Theoretical and Philosophical Implications: Our findings lend empirical weight to the philosophical proposition that motivated this work: that a form of self-awareness is a prerequisite for moral agency. The process of an agent locating itself within a normative space can be interpreted as

a rudimentary form of computational self-conception. The Ethical State vector, S_e, acts as a dynamic representation of the agent's "moral identity." The agent's actions are then not just instrumental towards achieving external goals but are also expressive of this identity. This reframes the value alignment problem from one of external control (programming the right rules) to one of internal guidance (cultivating the right kind of agent).

This work provides a potential bridge between high-level philosophical concepts of moral agency and concrete computational implementation. It suggests that abstract qualities like "conscience" or "integrity" might be computationally modeled as a persistent commitment to maintaining a particular state within a normative space. The recursive nature of the ESL update rule, which creates a memory of past actions, is crucial here. It means a single unethical action has a lasting, though decaying, impact on the agent's Ethical State, creating a computational analogue to guilt or a tarnished reputation that the agent is intrinsically motivated to rectify.

6.2. Limitations and Future Research

Despite its promising results, our study has several limitations that open avenues for future research. First, the Ethical-Normative Space and its associated projection functions were hand-designed for the specific context of the Harvest dilemma. While this was necessary to test the core principles of the framework, a key challenge for scalability is the automatic identification and formalization of salient norms in more complex, open-ended environments. Future work could explore methods for learning these norms from diverse data sources, such as social media, legal texts, or direct human feedback, perhaps using techniques from natural language processing and computational social science.

Second, the framework relies on a hyperparameter, β, to balance extrinsic and intrinsic rewards. The optimal value for β may vary depending on the context, and a fixed value may not be robust. Future iterations could investigate dynamic or adaptive mechanisms for setting this trade-off, allowing an agent to learn when to prioritize ethical considerations versus task performance, mirroring the complex situational judgment humans employ.

Third, our model assumes a single, universally agreed-upon Target Ethical State, S_target. In the real world, values are often pluralistic, contested, and culturally specific. A significant and challenging direction for future research is to extend the ESL framework to handle value pluralism and disagreement. This might involve agents learning to operate within multiple normative spaces simultaneously or negotiating a shared ethical framework through communication and interaction with other agents, both human and artificial.

Finally, the experiments were conducted in a simplified grid-world simulation. While such simulations are invaluable for controlled research, the ultimate test of the ESL framework will be its application to real-world robotic or software agents. This would involve addressing challenges of partial observability, noisy sensor data, and the immense complexity of human social environments.

7. Conclusion

This paper introduced the Ethical Self-Location (ESL) framework, a novel computational approach to the AI value alignment problem. By drawing an analogy to theories of consciousness, we proposed that moral agency in AI can be grounded in a recursive process of self-location within a multi-dimensional ethical-normative space. Our framework translates this philosophical concept into a concrete mechanism that provides intrinsic rewards to agents for aligning their behavior with predefined values, such as sustainability and equity.

Our core contribution is the demonstration that this approach is not only computationally tractable but also highly effective. Through a series of experiments in a simulated social dilemma, we showed that agents equipped with the ESL framework significantly outperformed standard reinforcement learning agents and models based on imitation learning. More importantly, the ESL framework proved superior to a purely utilitarian agent by successfully navigating the trade-off between collective utility and distributive fairness, achieving a state of high performance that was both sustainable and equitable. The results validate our central hypothesis that endowing agents with a form of moral self-awareness is a powerful method for fostering robustly ethical behavior.

The primary implication of this work is a conceptual shift in the approach to AI ethics. It suggests moving away from designing systems that merely follow external rules or optimize for simple utility functions, and towards developing agents that possess an internalized, dynamic model of a value system. The ESL framework provides a tangible pathway for creating agents that are not just compliant, but are guided by a persistent and evolving sense of their own moral standing. This offers a more transparent and principled foundation for building trustworthy autonomous systems.

While this research establishes a strong proof-of-concept, the journey towards truly ethical AI is ongoing. Future work must focus on automating the discovery of salient norms, developing more sophisticated mechanisms for handling value pluralism and disagreement, and validating the framework in complex, real-world applications. By continuing to bridge the gap between high-level ethical theory and computational practice, we can move closer to ensuring that the future of artificial intelligence is one that is aligned with the best of human values.

References

[1] Russell, S., & Norvig, P. (2021). Artificial Intelligence: a modern approach, 4th US ed. aima: сайт. URL: https://aima. cs. berkeley. edu/(дата обращения: 26.02. 2023).

[2] Hadfield-Menell, D., Russell, S. J., Abbeel, P., & Dragan, A. (2016). Cooperative inverse reinforcement learning. Advances in neural information processing systems, 29.

[3] Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. Advances in neural information processing systems, 30.

[4] Selbst, A. D., & Barocas, S. (2016). Big data's disparate impact. California Law Review, 104(3).

[5] Lahusen, C., Maggetti, M., & Slavkovik, M. (2024). Trust, trustworthiness and AI governance. Scientific Reports, 14(1), 20752.

[6] Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. IEEE intelligent systems, 21(4), 18-21.

[7] Wallach, W., & Allen, C. (2008). Moral machines: Teaching robots right from wrong. Oxford University Press.

[8] Sharkey, A., & Sharkey, N. (2012). Granny and the robots: ethical issues in robot care for the elderly. Ethics and information technology, 14(1), 27-40.

[9] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35, 27730-27744.

[10] Peters, F. (2009). Consciousness as recursive, spatiotemporal self-location. Nature Precedings, 1-1.

[11] Soares, N., & Fallenstein, B. (2017). Agent foundations for aligning machine intelligence with human interests: a technical research agenda. In The technological singularity: Managing the journey (pp. 103-125). Berlin, Heidelberg: Springer Berlin Heidelberg.

[12] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.

[13] Ng, A. Y., & Russell, S. (2000, June). Algorithms for inverse reinforcement learning. In Icml (Vol. 1, No. 2, p. 2).

[14] Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., ... & Irving, G. (2019). Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593.

[15] Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., ... & Kaplan, J. (2022). Constitutional ai: Harmlessness from ai feedback. arXiv preprint arXiv:2212.08073.

[16] Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J., & Graepel, T. (2017). Multi-agent reinforcement learning in sequential social dilemmas. arXiv preprint arXiv:1702.03037.

[17] Palmer, G., Tuyls, K., Bloembergen, D., & Savani, R. (2017). Lenient multi-agent deep reinforcement learning. arXiv preprint arXiv:1707.04402.

[18] Foerster, J., Assael, I. A., De Freitas, N., & Whiteson, S. (2016). Learning to communicate with deep multi-agent reinforcement learning. Advances in neural information processing systems, 29.

[19] Lowe, R., Wu, Y. I., Tamar, A., Harb, J., Pieter Abbeel, O., & Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. Advances in neural information processing systems, 30.

[20] Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. A., & Botvinick, M. (2018, July). Machine theory of mind. In International conference on machine learning (pp. 4218-4227). PMLR.

[21] Nematzadeh, A., Burns, K., Grant, E., Gopnik, A., & Griffiths, T. (2018). Evaluating theory of mind in question answering. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 2392-2400).

[22] Cao, K., Lazaridou, A., Lanctot, M., Leibo, J. Z., Tuyls, K., & Clark, S. (2018). Emergent communication through negotiation. arXiv preprint arXiv:1804.03980.

[23] Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. A., & Botvinick, M. (2018, July). Machine theory of mind. In International conference on machine learning (pp. 4218-4227). PMLR.

[24] Mead, G. H. (1934). Mind, self, and society from the standpoint of a social behaviorist.

ACKNOWLEDGEMENTS

None.

FUNDING

None.

AVAILABILITY OF DATA

Not applicable.

ETHICAL STATEMENT

All participants provided written informed consent prior to participation. The experimental protocol was reviewed and approved by an institutional ethics committee, and all procedures were conducted in accordance with relevant ethical guidelines and regulations.

AUTHOR CONTRIBUTIONS

The Author Yifan Ying conceived the study, developed the Ethical Self-Location framework, designed and conducted the experiments, analyzed the results, and wrote and revised the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.