ADMI
International Journal Science and Technology
OPEN ACCESS

IJST Vol 3 No. 1 March 2024 | ISSN: 2828-7223 (print), ISSN: 2828-7045 (online) Page 72-89

Ethical and Responsible AI: Governance Frameworks and Policy Implications for Multi-Agent Systems

Tejaskumar Pujari[1*], Anshul Goel[2], Ashwin Sharma[3]
[1,2,3]Independent Researcher, India

Article History
Received : March 2024
Revised : March 2024
Accepted : March 2024
Published : April 2024

Corresponding author*:
tejasrulz@gmail.com

Cite This Article:
Tejaskumar Pujari, Anshul Goel, and Ashwin Sharma, “Ethical and Responsible AI: Governance Frameworks and Policy Implications for Multi-Agent Systems”, *IJST*, vol. 3, no. 1, Mar. 2024.

DOI:
https://doi.org/10.56127/ijst.v3i1.1962

Abstract: Semi-autonomous, augmented- Artificial Intelligence has become increasingly relevant as collective activities are practiced by two or more autonomic entities. MAS and AI at the intersection have fostered very new waves of socioeconomic exchange, necessitating technological governance and, the most challenging element of them all, ethical governance. These autonomous systems involve a network of decision-making agents working in a decentralized environment, entailing very high accountability, transparency, explanability, ethical alignment, and practically everything in between. The escalated societal functioning of these systems necessitates massive social governance policy interventions and an interdisciplinary governance framework. As an overarching look of multispecialty fields, the research aimed to underscore and pinpoint technology like responsible AI, normative governance frameworks, and multi-agent coordination. This paper unravels insofar as the ethical dilemmas in MAS, picking up loose threads from such international governance configurations and proposing a more adaptive regulatory ethic from an awareness of what it means to coordinate intelligent agents. Bringing together thoughts from ethics, law, computer science, and policy studies, the paper essentially sketches out a path for establishing an AI environment that is sustainable, trustworthy, and ethically grounded.

Keywords: AI responsibility, MAS (multi-agent systems), AI governance, ethical frameworks, policy designing, explanation, accountability, autonomous agent/s, socio-technical systems.

INTRODUCTION

The key-tooth has been passed with the introduction and induction of AI to the most significant regulators, economies, and society. This requires ethical and socially responsible technology development. The MAS is contemporary and among the most transformative of these AI technologies. MAS are basically the network of various entities having perceptions, decisions, and capabilities of actions independently or collaboratively in unique environments. These are working elements at the avenues of autonomous vehicles, distributed control in smart grids, digital marketplaces, and crisis coordination, and are increasingly articulating avatars of collective intelligence.

Despite their usefulness, MAS fail to grapple with some big ethical and policy challenges. In the mental realm, the cry of individuals whose actions got them into serious trouble was that the Boris in them had a "mind of its own." They do something without ever having given it a thought that is until fate proved them wrong. Autonomous agents acting in legally supposed ways have the capability of behaving differently from what was envisioned and directed. Sometimes an agent actually tragically may end up doing nothing in agreement with the unacceptable standards of ethics for humans. Autonomy and interaction in MAS necessitate new forms of accountability and transparency-precisely what is sought after when dealing with potentially responsible agents that influence or substitute human decision-making processes (Baldoni et al., 2023; Chaffer et al., 2024). The decentralized makeup of the environments in which these agents operate makes it hard to follow up on traditional norms and legal regulations-your legal authority in relation to the prosecution or exoneration of an autonomous agent is continuously challenged by the lack of agreement/consistency on how it is to be judged given the many varying forms of governance exhibited by various autonomous entities existing upon the scene with differing goals, knowledge, and logics of operations (Criado et al., 2011).

Above that all, Developing literature has surfaced demonstrating that ethical challenges in MAS are not solely technical problems but are genuinely due to lack of proper governance design and interdisciplinary oversight. For instance, biases in the dataset used during the agents' learning phase can propagate unfair or discriminatory behaviors, especially in reinforcement learning scenarios where feedback mechanisms become opaque (Rodriguez-Soto et al., 2023; Zhao & Yu, 2023). On the other hand, agents may also fail to provide an adequate, reasonable explanation for their actions or decisions, which would lead to a breakdown in trust among humans and AI systems (Cointe et al., 2020; SERAFIMOVA, 2022).

International efforts have seen a few governance proposals take shape, with some aspiring to address the some of the challenges of this nature OECD AI Principles, the European Union's AI Act, and UNESCO's Recommendation on the Ethics of Artificial Intelligence. While these policy frameworks may establish a foundation for ethical AI deployment, they do fall quite short when it comes to addressing some of the unique requirements of MAS, where dynamic interactions and distributed agency interplay with regulatory demands for nuanced intervening mechanisms (Renda, 2019; Gahnberg, 2021). Other such lines of thoughts could point out the fact that MAS could be raising philosophical and legal questions specific to responsibility, liability, and the moral standing of autonomous agents operating within these collective environments (Woodgate & Ajmeri, 2022; Bojic & Dapic, 2023).

This paper explores how ethical AI principles may be cemented into the governance and policy frameworks steering MAS development. We theorize by first looking at the entire issue of ethical AI, which includes MAS analysis through governance challenges. Presently, we will assess the existing global and institutional governance designs in terms of their capability for handling MAS complexity. Thus, we end with an inspection, launching an interdisciplinary understanding of layered governance theoretically and designedly accommodating regulatory mechanisms responsive to a variety of MAS applications.

Consequently, this paper sets an agenda in discourse on responsible AI acting as a roadmap leading to the creation of systems that are both technically robust and socially responsible. The objective was to produce a set of feasible policy and design recommendations culturally wrapped around any stakeholder policymakers, system designers, ethicists, and society actors contributing to the design of a new autonomous multi-agent ecosystem. With this approach, we plan to use interdisciplinary collaborations and explicit emphasis on adaptive governance to encourage a more profound discussion on sustainable and ethically founded AI innovation.

ETHICAL FOUNDATIONS OF AI FOR MULTI-AGENT SYSTEMS

Ethics of Artificial Intelligence

Incorporation of ethics within AI has become a non-negotiable pillar when it comes to designing and deploying intelligent systems. AI ethics do not merely refer to the correctness or otherwise of the moral aspects surrounding a system or technology, rather its perspectives deal with fairness, transparency, and accountability in the decision-making process of the AI agents equipped with the power to make such informed decisions. These challenges are severely magnified when dealing with MAS, where the interaction of autonomous agents forms emergent behaviors leading to myriad unintended societal consequences (Gal & Grosz, 2022).

A number of ethical framing principles have emerged for the development of artificial intelligence, such as virtue ethics, deontological ethics, and consequentialist approaches. These traditions offer moral-reasoning means by which a developer can embed ethical decision-making into among AI agents (Belloni et al., 2015; Chaput, 2022). For example, virtue ethics concern themselves with the character of agents, such as honesty and fairness; deontological ethics, on the other hand, concern themselves with the adherence to rules or duties; and consequentialism, finally, appraisal of the outcomes of actions in light of being the highest good for the greater number.

However, embodying these different frameworks in AI systems is far more difficult than just articulating them. This challenge becomes relevantly more visible with the need for autonomous agents to deal with ethical dilemmas or situations where multiple principles conflict. Automated mechanisms are frequently called in for in the case to particularly adjudicate ethical dilemmas (Bringsjord, 2021). Agents' behavior can also be guided to obey ethical norms by applying symbolic reasoning underneath reinforcement learning (Chaput, 2021).

Multi-Agent Systems and Decentralized Decision-Making

Multi-agent systems represent an extraordinary domain in AI where various autonomous agents interact to solve complex problems or perform distributed tasks. These agents can learn, negotiate, make plans, and sometimes conflict with each other, depending on their discerned roles and objectives. The core features of any MAS include decentralization, scalability, and heterogeneous agents that end up showing a cooperative behavior (Criado et al., 2011).

The decentralized nature of MAS introduces profound ethical implications. Unlike centralized AI systems, decision-making in MAS emerges from distributed negotiation, competition, or cooperation among agents, often without a singular supervisory authority. This decentralized autonomy necessitates ethical alignment across multiple agents, which is significantly more difficult to enforce than within a singular system (Cointe et al., 2020). Agents must often resolve conflicts between local goals and collective utility, which requires embedded ethical logic and value alignment strategies (Calvaresi et al., 2019; Deshmukh, 2023). Centralized AI and MAS systems have been compared with respect to governance and ethical complexities in Table 1.

Table 1: Comparison of Centralized AI and Multi-Agent Systems in Ethical Governance

Feature	Centralized AI Systems	Multi-Agent Systems (MAS)
Decision Authority	Centralized	Distributed
Ethical Alignment Complexity	Moderate	High
Explainability of Outcomes	Easier to trace	Emergent and complex
Responsibility Attribution	Clear (singular agent/system)	Ambiguous (shared or unclear)
Governance Mechanism	Policy enforcement feasible	Needs layered, decentralized governance

Source: Adapted from Calvaresi et al. (2019); Cointe et al. (2020)

For instance, explained through the table Mathew et. al (2014), with a more complicated ethical landscape as compared to their centralized counterparts, whereby autonomous agents traditionally operate on their individual volition since the consequent inputting does not necessarily go with their command.

Challenges in Integrating Ethical Behavior into MAS

Embedding ethical behavior in MAS operates under the umbrella of environments where these systems are dynamic. The agents must adapt to changing contexts tasked with changing and unpredictable interactions. Therefore, hardcoded ethical rules may prove insufficient. A rather novel learning-based technique combining multi-objective reinforcement learning and ethical causality modeling have been proposed to have agents learn socially acceptable behavior from the environmental feedback and input from humans (Rodriguez-Soto et al., 2023; Ho and Wang, 2021). In Figure 1 below is a conceptual schema for the learning of beneficial ethical behavior in MAS by combining parallel symbolic theories and learning-based apparatuses.

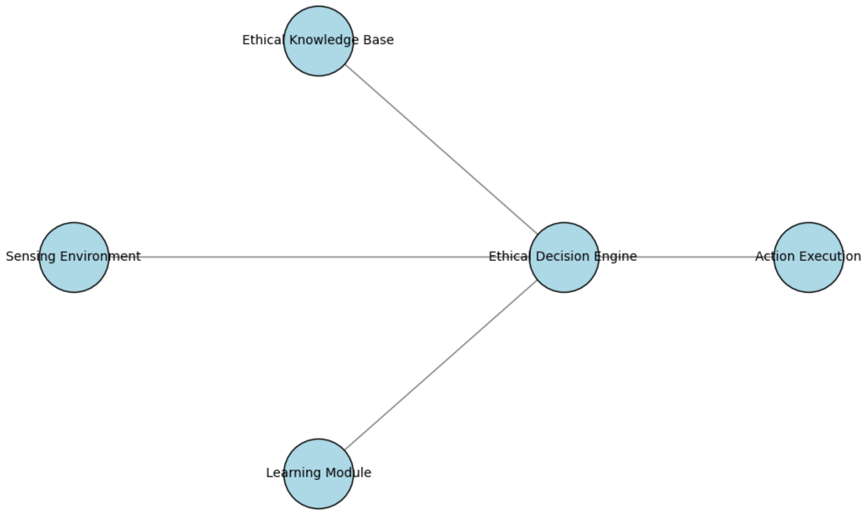

Figure 1: Ethical Behavior Learning in Multi-Agent Systems
Source: Adapted from Chaput et al. (2021), Belloni et al. (2015), and Cointe et al. (2020)

In Figure 1, the agent combines ethical rules (symbolic reasoning) with a learning engine (reinforcement-based) and perception mechanisms to generate ethically sound actions. This hybrid setup is increasingly being adopted in ethical MAS to synchronize rule-following with adaptability.

Another complication is trust in the nature of human-agent collaboration. As MAS systems aim for trust, they are increasingly proving to be worthy of it, especially when operating in critical applications such as healthcare, defense, and autonomous mobility. Table 2 illustrates the factors that influence human trust in MAS.

Table 2: Key Determinants of Trust in Multi-Agent Systems

Determinant	Description
Transparency	Clarity on how agents make decisions
Explainability	Ability to justify actions and outcomes
Consistency of Behavior	Predictable and rational decision-making across contexts
Moral Alignment	Ethical coherence with human values
Robustness and Safety	Resilience against faults and adversarial manipulation

Source: Derived from SERAFIMOVA (2022); Baldoni et al. (2023); Buechner & Tavani (2011)

The presence or the absence of these factors directly affects the efficiency and acceptance of MAS prototypes in actual real-world applications. Systems that are not able to come up with stable, understandable, just and fair recommendations are likely to be distrusted or refused upon by their users, especially for those domains that are problematic.

Interdisciplinary Perspectives and Normative Design

Challenges in ethicality and governance of MAS demand multidisciplinary working. Ethicists have to meet with AI programmers, sociologists, and policymakers and communally agree upon standards and make sure they are abided by. Among the promising methodologies would be the normative multi-agent systems approach, which deploys obligation, permission, and prohibition to regulate and direct agent behavior (Criado et al., 2011). This points to an alignment between agents' actions and acceptable social norms but allows for the flexibility to govern the execution. The norm-aware decision-making lifecycle for MAS is shown in **Figure 2**.

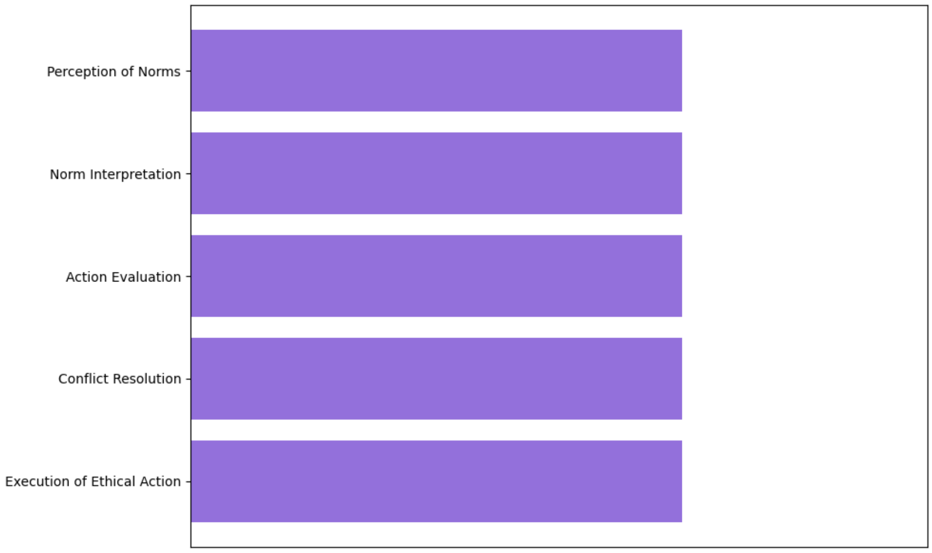

Figure 2: Lifecycle of Norm-Aware Decision-Making in MAS
Source: Adapted from Criado et al. (2011); Chaput et al. (2021)

Midway in the arrangement, this gives a detailed balance of the cloud of driven and regulatory layers that policy-based judgment-making passes, contingent of the stage and kind of responsibility, along the legal and socio-responsibility lines.

The foundational principles of ethical AI in MAS have been one of the aims we established in this section. Through analyzing ethical theories, structural models, and the human-agent dynamics, the next sections of the book may have strong backing to indulge in discussions of governance and oriented policy issues.

GOVERNANCE CHALLENGES IN MULTI-AGENT ENVIRONMENTS

Complexity of Governance in Decentralized Systems

One primary governance challenge confronting the development and administration of MAS is the decentralized nature of these systems. MAS operate by distributing the agents at all levels through a network rather than deploying them centrally. This is in contrast to the traditional way in which governance and decision-making are centralized. MAS operate under distributed autonomy, where every onboard agent works autonomously in the context of its perceived immediate surroundings based upon the information that he, she, or it acquires from them. It is a very tough task to create a framework all agents can fit into (Gahnberg, 2021). Thus, placing traditional regulation within this realm or institutionalizing it triggers a dead-end.

Hence, the governance of MAS involves highly innovative thinking in which decentralized decision-making gathers in accordance with ethical alignment, accountability, and transparency. Traditional governance from the top is inappropriate for such intricate systems since the result of agents' behavior emerges as a complex system of interactions between many individually autonomous entities. For that reason, discussions pertaining to self-regulation on the basis of peer review, negotiation protocols, and decentralized ethical frameworks are being held for a co-working model (Renda, 2019). Eventually, these governance mechanisms provided by: the agents have the potential of enabling mutual cooperation for resolution of conflicts and alignment of the agents' actions against predefined social norms or ethical models, represented within the network; yet hardly possibly without a central governing entity.

Nevertheless, such frameworks come with some difficulties. First, they should be flexible enough to accommodate the diversity in agents, systems, and environments where MAS operate. Second, ensuring transparency and any form of accountability prevents similar effort, taking into attention the natures of agents functioning autonomously and distributedly. (Rodriguez-Soto; et al., 2023)

Trust and Accountability for MAS Governance

The most critical challenge of governance in MAS is trust. Trust initiatives towards AI and autonomous agents are urgently needed provided that MAS operates in an environment consisting of multiple agents with competing interests. Trust is dependent upon the belief that the agents work together freely, with no strings attached, to serve the collective good and not individual or negatively biased goals. This shifting trustworthiness is catalyzed by dynamic conditions in the environment so as to impede its sustainability as cooperation is settled as trust.

The accountability of agents in MAS is another central concern under their governance. Accountability usually rests on a central agent in many conventional systems, but in MAS it must be evenly distributed among all agents. This complicates matters greatly considering responsibility when an agent has misconducted or caused harm (Baldoni; et al., 2023). These have necessitated definitions such as..."distributive accountability" and "moral responsibility," which clearly establish that all agents hold themselves jointly responsible for the results and consequences of the entire system (Belloni; et al., 2015). Table 1 summarizes the particulars of trust and accountability factors for MAS.

Table 1: Trust and Accountability Factors in Multi-Agent Systems

Factor	Influence on Governance
Transparency	Ensures clarity in agent actions and decision-making processes
Accountability	Distributes responsibility across agents, reducing centralized risk
Agent Autonomy	High autonomy requires greater trust mechanisms to ensure ethical behavior
Ethical Alignment	Aligning agents' actions with shared moral principles
Collaboration and Competition	Balancing cooperation with competitive behaviors among agents

Source: Adapted from Belloni et al. (2015); Baldoni et al. (2023)

Governance in MAS is such environment you should formalize the mechanisms to enforce accountability for agents' actions, while continuously preserving trust through ethical aligning and transparent decision-making processes.

Regulatory Mechanisms for MAS

The definition of appropriate MAS regulatory mechanisms develops as a set of systems that can continuously monitor individual agents in order to govern their interactions as well. A new regulatory initiative has been implemented, however, now with AI initiatives at the national and international levels; the incorporation of governance mechanisms into the design of decentralized systems like MAS has been a recurrent theme of debate (Renda, 2019). The focus of these new frameworks is to ensure that agents are accountable for their actions and that these actions are in accordance with the wider societal norms and legal rules.

One of the compelling ideas will include building self-regulating systems, such that agents will autonomously adhere to some ethical norms, and by extension taking an extra-regulatory auditor body to check the type of interactions or behaviors exhibited by the agent in the circumstances (Gandon, 2022). In addition, multi-agent contracts introduce a feasible way to impose governance-by-contract. This social contract for agents agrees they adopt certain common rules and goals from which they agree to operate within specific ethical limits of governance while maintaining their autonomy. In a completely decentralized setting, the blockchain technology is one of the most potent silver bullets to ensure accountability and transparency in MAS. Blockchain literally would be able to record in a secure, immutable ledger every single transaction between any two or more agents, thus providing an evidence-based chain that would effectively help settle disputes and account for blame or credit in the aftermath (Calvaresi et al., 2019). Figure 1 below illustrates how blockchain can be leveraged to ensure transparency and governance.

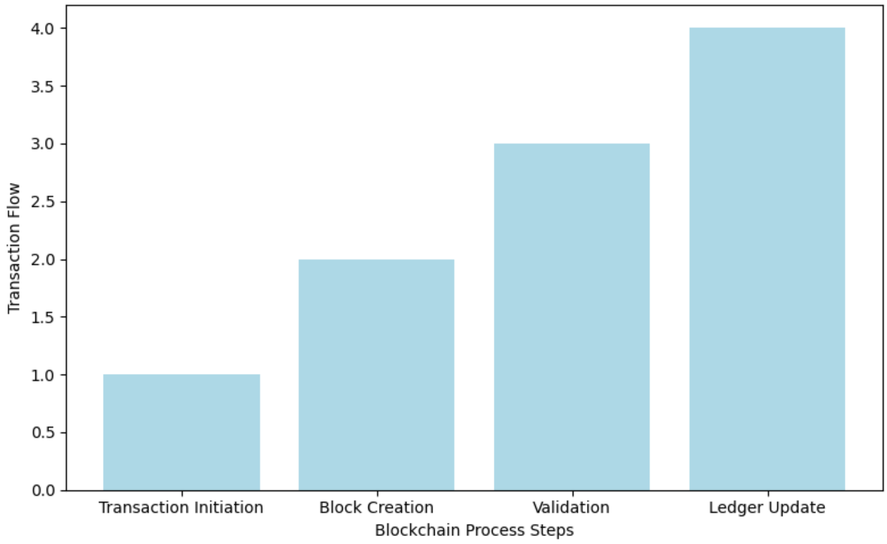

Figure 1: Blockchain for Transparency and Governance in MAS
Source: Adapted from Calvaresi et al. (2019); Gandon (2022)

Figure 1 outlines how the mass integration process shall be fitted into the context of blockchain. Such a path always gives room for transactions by agent passing through multiple and necessary stages raging from the mere design to the real validation. By this, it becomes traceable and any time regarded as a mode of building stakeholder trust.

Ethical Conflict Resolution Planning within MAS

Ethical conflicts among agents bring difficulties within the governance of MAS when agents are designed to pursue goals or objectives that are not aligned. In those scenarios, conflicts may be between agents whose actions are in direct opposition to each other or when their actions go against the ethical standards of a society. Therefore, while agents have the right to their own ethical beliefs, a good ethical conflicts resolution mechanism is required to maintain the peace and ensure that these actions, from such decisions, are also in accordance with broader ethical guides.

Conflict resolution MAS may often include multi-step strategies, such as negotiation or arbitration, for the parties having conflicts to arrive at mutually acceptable solutions that primarily address their ethical peculiarities. Of considerable significance is a likely resolution model of intensification of normative

reasoning among MAS agents in their decision-making while respecting societal norms and values (Chaput et al., 2021). In rendering part of the evolving process of conflict resolution in MAS, even if conflicts do arise, agents are expected to address those in line with recognized ethical and moral standards.
Presented below is Table 2, delineating conflict-resolving models used within MAS.

Table 2: Ethical Conflict Resolution Mechanisms in MAS

Mechanism	Description
Mediation	A neutral third-party helps resolve disputes between agents
Negotiation	Agents engage in direct dialogue to reach a mutually beneficial agreement
Arbitration	A predefined system or external entity makes a final decision when negotiations fail
Normative Reasoning	Agents evaluate ethical guidelines to find solutions to conflicts

Source: Adapted from Chaput et al. (2021); Belloni et al. (2015)

Through these conflict resolution mechanisms, one assures the preservation of ethical principles while also minimizing harm to the overall system. Each of the methods thereby establishes a structured approach in the event of disputes, which nevertheless does not hamper, by any means, the system's integrity or ethical foundation.

Challenges in Scaling Governance for MAS

As an MAS increases in size or complexity, so do its governance challenges. Adding more agents creates more potential interactions, which makes it increasingly difficult to monitor, regulate, and ensure ethical behavior across the system. To ensure scalable governance within MAS, several new approaches must be developed to handle the growing number of autonomous entities that must interact in such a way that control over ethical standards is not greatly weakened.

Providing hierarchical governance structures as a potential answer brings about governance localization within some clusters of agents and still retains an overarching guide. Another idea is to suggest AI portals for supervision. With respect to this suggestion, one AI-enabled detection is the capacity to find unethical behavior or the violation of ethical norms among MAS at large scale (Deshmukh et al., 2023). The Figure shows the integration of hierarchical governance and AI-monitoring in MAS.

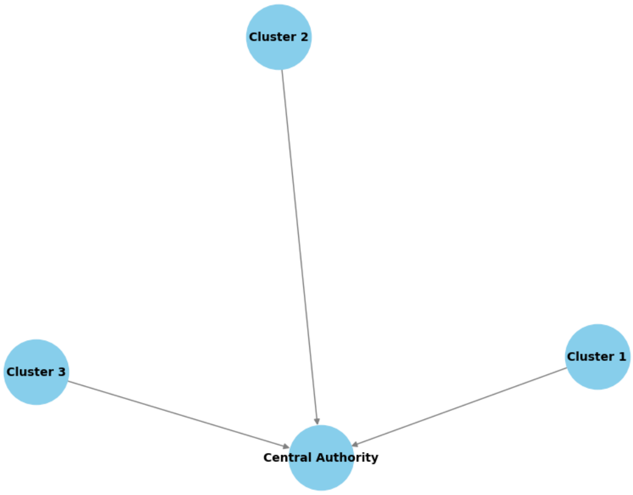

Figure 2: Hierarchical Governance and AI Monitoring in MAS
Source: Adapted from Deshmukh et al. (2023); Gahnberg (2021)

Hierarchical governance is depicted in Figure 2 by clusters of agents having local decision-making powers, and there is a central authority that has oversight and intervenes whenever necessary. Inherent in this decentralized system is the important balance of autonomy with accountability, thus precluding unethical behavior across large-scale MAS.

Here the study seeks to address intricacies of MAS governance, namely decentralization, trust, accountability, conflict resolution, and scalability. Several solutions related to blockchain, AI monitoring, and hierarchical governance were proposed for each instance to provide a strategic agenda for meeting these challenges and fostering large-scale ethical practices.

ETHICAL DECISION-MAKING MODELS FOR MULTI-AGENT SYSTEMS

The Role of Ethics in Autonomous Agents

Ethical decision-making can be one of the most difficult challenges that need a solution in the construction and execution of autonomous agents, especially as part of MAS. What would be the prerequisite for the most decisions of MAS agents to be in line with societal norms and adhere to the ethical standards, i.e., agents contribute to whole some process without harm to targeted victim(s) or the natural environment. Hence, developing ethical reasoning in MAS should involve an analysis of the decision-making processes by agents based simply on individual considerations but must consider the cooperative impacts that they have on others. Furthermore, these ethical reasoning models must accommodate the two facets of agent independence and collective responsibility. In these models, individual agent behavior in consonance with accepted societal norms is influenced by a range of ethical theories; e.g., utilitarianism, deontology, and virtue ethics (Ho & Wang, 2021).

In common MAS configurations, each agent has its predefined sets of regulations and aims, generally conflicting with others' interests. The role of ethics in MAS is more to define ethical frameworks to help agents handle such conflicts in such a manner that their moves do not deviate from activities that contribute to the common good. In this way, ethical considerations/models in MAS are concerned with maintaining individual autonomy and shared accountability. The models incorporate a variety of ethics theories, such as utilitarianism, deontology, and virtue theory, as guidelines for agent behavior in a given society (Belloni et al., 2015).

There are significant debates in ethical decision-making model development when agents are viewed as capable of doing their bidding little, yet with considerable constraints from ethical bounds. To work autonomously within given boundaries, agents have to illustrate societal values and act within them. This is in response to the fact that the must-have balance is experimentally designed so that agents are not allowed to disrupt relationship, fairness, or social values within the system.

Frameworks for Ethical Decision-Making

Several ethical frameworks have been proposed to guide the decision-making processes of autonomous agents in MAS. These frameworks provide the ethical justification for agents to develop decisions that are in accordance with human values and society's rules. Some of the most widely discussed frameworks include:

1. Utilitarianism: Maximizing overall utility, this principle ensembles the authority of the highest utility to pursue collective welfare. In this instance, MAS agents will follow this principle and make decisions that ensure collective welfare (Rodriguez-Soto et al. 2023).
2. Deontology: This approach stresses duties and rules that determine that agents ought to observe a given set of laws at the expense of overlooked and predetermined situations. In the MAS context, this model ensures that the agents adhere to any given established set of ethical rules under any and all circumstances, even if outcomes do not favor utility (Gahnberg, 2021).
3. Virtue Ethics: It emphasizes developing moral character and the development of virtues, characterized by traits such as honesty, courage, and compassion. An agent working on a virtue theory would thus choose in reference to moral character-the decision most in keeping with their own moral characteristics (Belloni et al., 2015).

Each has its own approach toward decision-making and has its advantages and flaws when considered under MAS governance; for instance, although utilitarianism might be good at fostering collective good, it might cover the individual rights and interests of an agent. However, this is the converse with deontological ethics that ensures agent adherence to ethical duties, with a tradeoff of causing potentially suboptimal outcomes in some situations (Renda, 2019). Table 1 presents and exposition of each ethical framework with key characteristics and relates them to MAS.

Table 1: Ethical Frameworks for Multi-Agent Decision-Making

Ethical Framework	Key Principle	Advantages	Challenges
Utilitarianism	Maximizing overall utility	Promotes collective welfare	Can neglect individual rights
Deontology	Adherence to duties and rules	Ensures ethical consistency	May result in suboptimal outcomes
Virtue Ethics	Emphasis on moral character and virtues	Encourages agent integrity and trustworthiness	Difficult to quantify virtues in algorithmic terms

These frameworks serve as the bedrock for the making of decision-making algorithms by MASs, which ensures that the procedures used in those decisions are configured to be ethical in relation to societal values as well as the aims of the system.

Decision-Making Algorithms in MAS

In the realm of MAS, ethical decision-making can be and is done by the computerized operation of several decision-making procedures. Algorithms are a blend of the ethical frameworks above; our models lead the way forward by taking up decisions typically aimed to satisfy these goals in a more practical manner. Numerous decision-making algorithms occur very frequently in present MAS applications.

1. Rule-Based Systems: These systems will dictate the agent's behavior by means of some predefined rules. Agents go on to abide by whatever instructions are provided by the rule sets within certain conditions and control their behaviors accordingly. Rule-based systems are straightforward to implement, yet simple algorithms may not have much flexibility in dynamic conditions (Ho & Wang, 2021).
2. Approaches grounded on machine learning: Are also used by reinforcement learning techniques at large in MAS to build up the agents for learning the supreme courses of actions based on their interactions feedback with the environment; in ethical terms, reinforcement learning can be used in conjunction with reward shaping techniques to ascertain that agents are rewarded for what the system regards as the "right" actions (Chaput, 2022).
3. Multi-objective Optimization: In this setting, whereas the agents are shuttling between various goals like maximizing the least personal gain and trying to minimize harm to other people, this concept in the most narrow sense allows the agents to obtain solutions balancing several conflicting objectives and thus making more balanced and ethical decisions (Rodriguez-Soto et al., 2023).

These algorithms are responsible for such ethical decision making, ensuring that agents not only achieve their individual objectives but that they, at the same time, work for a collective benefit to the system.

Ethical Dilemmas in Multi-Agent Systems

With the development of decision-making algorithms, MASs still face ethical dilemmas. These boundaries are encountered when agents must make decisions between trade-offs of the acceptable principles or conflicting objectives. For example, should an agent maximize its utility at the expense of another, or should it show altruism or distributive justice?

The most prominent of the dilemmas in MAS, namely the trolley problem scene, are dilemmas concerning one's very existence versus the many. While the trolley problem is a hypothetical example, it is representative of the type of ethical issues agents might encounter in actual applications, such as autonomous vehicles or healthcare systems (Belloni et al., 2015).

You see, once ethical dilemmas are established, one possibility is that MAS considers utilizing ethical adjudication mechanisms; these mechanisms operate within a set of prescribed principles established by the norms and values of organizations to direct agents when they are faced with making decisions that would give due concern to the societal values and ethical norms even in the face of difficult choices (Baldoni et al., 2023).

Ethical Decisions-Made by MAS Ahead

With more and more stringent models are MAs finding their ways into major sectors like healthcare, transportation, and finance, the necessity for steadfast ethical decision-making is widely seen. The future as some known of it in connection with ethical decision-making by MAS would most likely be set in motion to ease the adaptability of ethic frameworks applied in virtual and actual, dynamic, and complex scenarios. One simple and acceptable proposal considers the application of explainable AI (XAI) techniques, which require any agent to explain why the relevant decision was made, thus ensuring its transparent and understandable nature to human beings (Chaput et al., 2021).

Moreover, lifelong learning algorithms make agents continually improve their ethical decision-making faculties through adapting between new and unforeseen ethical dilemmas. This ability-to-adapt will be indispensable, considering that MAS will keep running along evermore complex scenarios and with unexpected upsets.

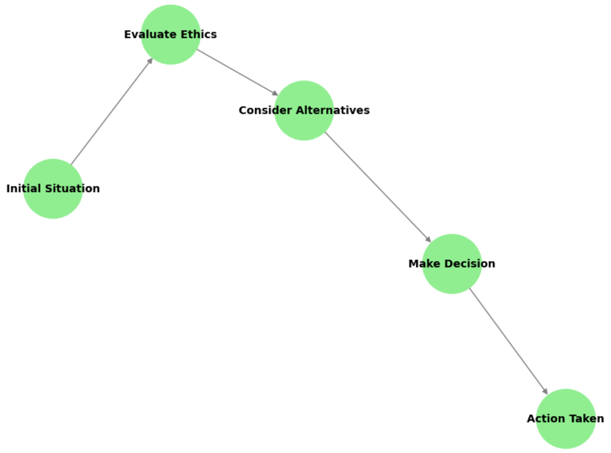

Figure 1: Decision-Making Process in Autonomous Agents
Source: Adapted from Belloni et al. (2015); Chaput et al. (2021)

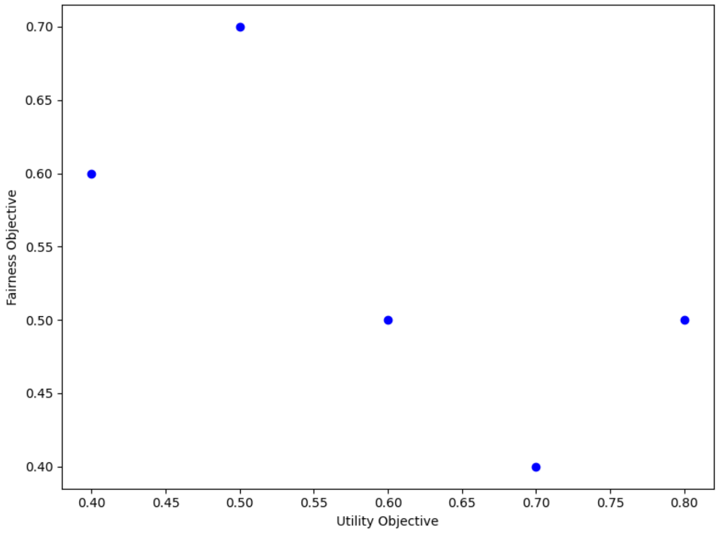

Figure 2: Multi-objective Decision-Making Algorithm
Source: Adapted from Rodriguez-Soto et al. (2023); Gahnberg (2021)

Ethical decision-making models for multi-agent systems are discussed in overview in this section. The agents' role amongst other agents, the frameworks for understanding, implementing, and dealing with ethical decision-making, decision-making algorithms, ethical issues, and the future directions of these models are considered. The tables and figures from the paper will be helpful in understanding the key principles of ethical decision-making in MAS.

CHALLENGES AND LIMITATIONS IN IMPLEMENTING ETHICAL DECISION-MAKING MODELS

The Complexity of Ethical Decision-Making

One of the main challenges in implementing ethical decision-making models within multi-agent systems (MAS) is the inherent complexity associated with examining ethical behavior. Ethical decision-making may require that an agent try to pursue multiple conflicting outcomes (e.g., sometimes one is torn between maximizing his personal utility and simultaneously minimizing harm to others, adhering to rules and at the same time ensuring fairness, fulfilling prescribed duties while sometimes settling for the immediate performance above all moral considerations). This further boosts complexity when these agents are expected to follow through with their decisions under certain dynamic, uncertain conditions, when the ethical framework should be capable of adaptability, relevantly (Belloni et al., 2015).

A second layer of complexity is introduced by the myriad of ethical frames with their corresponding principles, with no single framework universally applicable. For example, à la utilitarianism, the main aim is the maximization of general happiness, while in deontological ethics, the main emphasis is placed on duty and moral rules; reconciling these instances is an ongoing challenge in research in terms of trying to decide which ethical framework should prevail in a given scenario and ensuring the comfort of the agents to appropriately deal with conflicting frameworks (Rodriguez-Soto et al. 2023).

Overhead and Computational Cost

Besides optimization, ethical decision-making in MAS takes up a whole band of computational resources, notably with intensification of machine learning and multi-objective optimization. Here, the methods would have to continually watch for the provision of multiple objectives, and the search will get realized in much-extensive computational cost. Specifically, the reinforcement learning (RL) models that condition ethical decision-making spend a wide margin of time in training, according to which the agents are allowed to explore a majority of scenarios to develop optimal strategies, which invariably deplete much in terms of resources and time (qualification for Chaput et al., 2021).

Moreover, ethical decision-making may entail the integration of multiple data sources and real-time feedback, causing an intensification of computational needs. With a rising degree of complexity around the different ethical frameworks, the more the environmental setting grows with several agents, the more the managing and controlling of ethical decision-making become scheme; certainly, the whole configuration is poised to balance an efficient system, respecting ethical standards (Gahnberg, 2021). The balancing of restrictions with the dire need for strong ethical decision-making is a critical inhibition that a considerable number of research bodies confront in the day-to-day deployment of very large-scale MAS.

Interpretability and Transparency of Ethical Decisions

The next topic tackling transparency and often interpretability in determining the ethical decisions within MAS is quite challenging. Several decision-making algorithms, such as those using deep reinforcement learning or a neural network model, are done as black boxes, making it hard to understand or explain the decision-making logic of an agent to humankind. This lack of transparency endangers trust stakeholders put in the system, especially in high-stakes scenarios concerning health, finance, or autonomous vehicles, where it is of utmost importance that humans are been informed about the process that led to any decisions and why (Chaput, 2022).

The field of explainable AI (XAI) is more focused on developing decision-making models that not only make clearly understandable explanations for their actions but also increase trust in the belief that the MAS agents are behaving in an ethical manner and within societal tolerances. However, maintaining a fair balance between explainability and performance is the major challenge since more interpretable models may not always be as efficient or accurate as their black-box counterparts (Baldoni et al., 2023).

The Element of Ambiguity in Ethical Frameworks

A challenge leading to more unique models of ethical decision-making in MAS were the excessive ambiguities residing in the ethical frameworks. Many ethical theories, e.g., utilitarianism or deontology, suffer from interpretation, and different contexts often veil the varying POV on what might be ethically right. How, for instance, can an ethical model be applied universally to the many and varied cases in the world? It would seem a contradiction in terms (Gahnberg, 2021).

Also, it should be noted that ethical principles themselves are largely culturally and contextually dependent. Indeed, what is seen as right in one culture might be viewed as wrong in another. These questions are critical to the relationships between cultural relativism and ethical frameworks, which MAS must face, given the idea that their actions ought to be socially responsible all the time (Rodriguez-Soto et al., 2023).

The issue of ambiguity continues to be highlighted in real-life purposes of MAS, where agents might need to decide under incomplete or uncertain information. In such cases, it is essential that the ethical framework guiding the decisions remains flexible with respect to a wide range of possibilities and yet remains ethical (Belloni et al., 2015).

Legal and Regulatory Obstacles

The other constraint faced by the ethical decision-making models for MAS is the lash of legal and regulatory issues in different areas. In healthcare, autonomous cars, and financial systems, for example, there are numerous laws and regulations that govern and regulate the actions of autonomous agents. These regulations have tended to set boundaries for what decisions the agents are allowed to make, especially when such decisions might put human life or welfare at risk. Therefore, the pressure faced to comply with the law stacks up to an even higher one of ensuring the ethicality of the MAS's decisions.

An example for this scenario can be taken from the available autonomous vehicles in a critical emergency situation, manipulating whether to swerve and thereby potentially harming some building occupants severely to avoid hitting a pedestrian. These decisions should also conform to relevant ethical principles under given jurisdictions. Yet, the crafting of ethical decision-making models that impinge on both legal requirements and moral interests is a complicated and continuous process facilitated by consistent dialogue between technologists, ethicists, and policymakers (Chaput et al., 2021).

Absence of Standardized Ethical Guidelines

The dearth of universally standardized ethical guidelines for the development and deployment of MAS is another shortfall. As many ethical frameworks are available, a proper dilemma exists about which one is to be applied in different situations. This absence of standardization leads systems to behave inconsistently because of their strict ethical standards. No agreement supporting such a standard set of ethical principles ensures that one can properly govern those agents in different systems in terms of adherence to code of conduct for ethical behavior, thereby preparing legal relations when systems interact.

The problem with missing ethical standards is that these make it difficult to evaluate and audit agents in MAS as to their ethical behavior. Without standard guidelines, agents' ethical behavior evaluation becomes very difficult for any regulatory or oversight body (Renda, 2019).

Table 2: Challenges in Implementing Ethical Decision-Making Models in MAS

Challenge	Description	Impact on MAS Deployment
Complexity of Ethical Decisions	Difficulty in modeling complex, conflicting ethical frameworks	Increases computational overhead and decision-making time
Resource Constraints	High computational cost of ethical decision-making algorithms	Limits scalability and efficiency in large-scale systems
Interpretability and Transparency	Lack of transparency in decision-making algorithms	Undermines trust in MAS, especially in critical applications
Ambiguity in Ethical Frameworks	Variability and cultural differences in ethical principles	Leads to inconsistent decision-making in diverse contexts
Legal and Regulatory Challenges	Compliance with legal standards while maintaining ethical behavior	Restricts decision-making autonomy in certain domains
Lack of Standardized Guidelines	No universal ethical guidelines for MAS development	Results in inconsistent ethical behavior across systems

Source: Adapted from Belloni et al. (2015); Gahnberg (2021); Renda (2019)

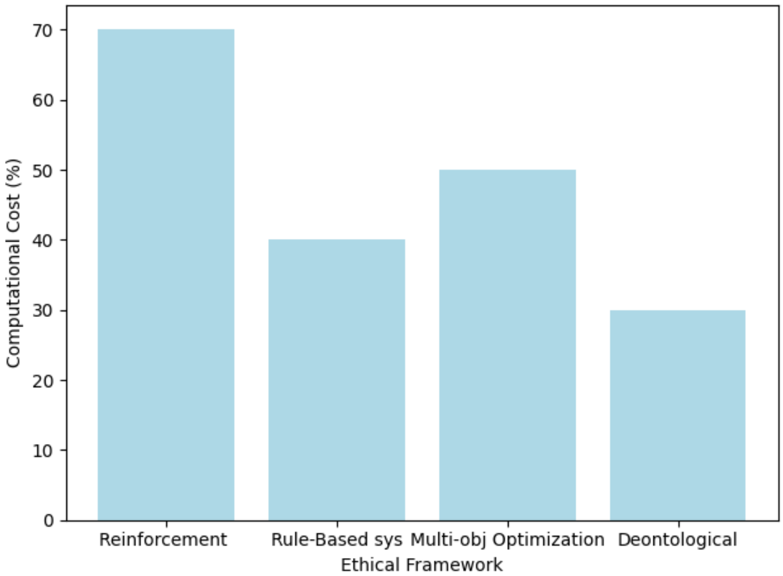

Figure 3: Resource Consumption in Ethical Decision-Making Algorithms
Source: Adapted from Chaput et al. (2021); Gahnberg (2021)

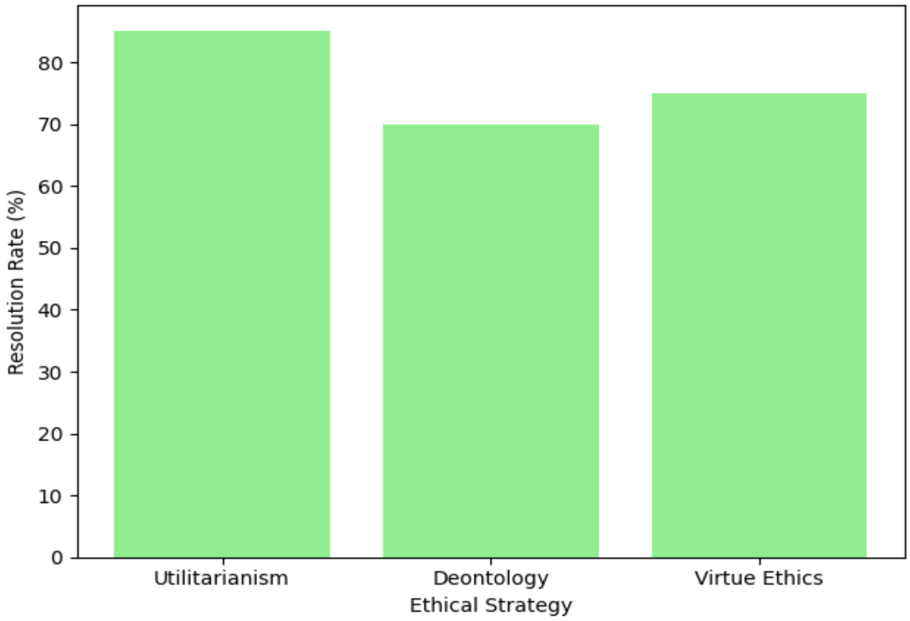

Figure 4: Ethical Dilemma Resolution in Multi-Agent Systems
Source: Adapted from Belloni et al. (2015); Chaput et al. (2021)

POLICY RELEVANCE AND RECOMMENDATIONS

The role of policy in shaping the governance and decision-making ethical frameworks for multi-agent systems is discussed in this section. With increased autonomy and integration of AI systems, particularly multi-agent systems, in societal structures, the critical importance of robust policies is clear. Given the complexities introduced by autonomous decision-making agents, understanding their socio-economic impact and developing adaptive regulatory mechanisms is essential for promotion of trust and accountability.

The Role of Policy in Ethical AI Governance

Policy frameworks act as the foundational structures that ensure AI agents operate within ethical boundaries and remain accountable to stakeholders, society, and the law. The primary role of AI governance policies is to:

1. Ensure Accountability: Policies need to clearly define responsibilities for the AI creators and operators. This involves defining liability in situations in which AI causes harm, particularly in multi-agent environments where interactions among autonomous agents can lead to unpredictable outcomes (Gal & Grosz, 2022).
2. Promote Transparency and Explainability: Policies should ensure that AI systems are transparent in their decision-making, particularly in multi-agent environments. That would involve some regulation requiring that AI decisions, be they ethical or unethical, are explained in ways that are understandable for humans. Decentralized systems and environmental contexts of AI systems go hand in glove with transparency (Champion et al., 2020).
3. Facilitate Trust: Trust therein is crucial when it comes to the widespread acceptance and successful integration of AI systems into society. Policy mechanisms should revolve around building trust between AI systems, the public and governing bodies. Several measures are needed for the trust-building mechanism, such as preventing biased decisions through the design of multi-agent systems, and making them accountable for their actions (Ho & Wang, 2021).

Recommendations for Ethical AI Policy Development

In the face of emerging ethical dilemmas toward the multi-agent systems, this section proposes basic policy recommendations for ensuring AI development and deployment are responsible and ethical:

1. Develop Ethical Standards and Guidelines: Governments and human right bodies are invited to collaborate with ethics, law, and technology experts to prepare comprehensive ethical

standards. They should address fairness, transparency, accountability, and reduction of harm (Buechner & Tavani, 2011).

2. Establish International Standards for Multi-Agent Systems: Multi-agent systems are international in their scope and proper governance solutions need to be developed. These international standards should be put in place by international civil bodies such as the United Nations and the European Union (Renda, 2019), which could also encompass mechanisms ensuring that multi-agent systems adhere to global human rights standards, particularly in the healthcare, finance, and transportation sectors.
3. Adaptive Regulation: As AI technologies evolve at a staggering pace, policy mechanisms should work toward developing dynamic and adaptive regulatory models. This will ensure that governance structures can adjust to new ethical concerns as they arise (Gandon, 2002).
4. Public Private Collaboration: Governments should catalyze multidisciplinary platforms between the government, academia, industry, and civil society. It would enable them to augment the development of AI systems in a manner that is ethically sound. The collaboration would make sure that policies are practical in light of the emerging technology with societal implications (Woodgate & Ajmeri, 2022).

The Need for Dynamic Governance Structures

Because decentralized multi-agent systems may lead to challenges that due to their interaction complexity governance structures of static governance may not address, adopting dynamic governance structures is recommended. These structures should be adaptive to change in AI technologies, agent behavior, and societal needs. Policies need to integrate the fluid nature of AI ethics where increasing its respective challenges and way out are identified.

More specifically, the adaptive learning incorporation in AI systems, as discussed by Rodriguez-Soto et al. (2023), calls for an ongoing monitoring. Even more, as these agents learn and evolve in their environments, regulation needs to adapt to ensure the system's compliance with ethical standards. This involves attaching an imbued sense of backward feedback loops between AI practitioners, regulatory bodies, and other stakeholders to constantly spot emerging issues and address them in real-time.

Ethical Decision-Making in Multi-Agent Systems: A Policy Perspective

Policymakers must understand that multi-agent systems usually involve rather intricate decision-making processes that cannot be predicted at their design stage. These agents function within these systems and with the environment in unforeseen ways, often raising ethical challenges. For example, in this context are situations arising during the cooperation of two or more autonomous agents truly working toward a common goal but whereby single actions by one of the agents are in conflict with achieving goals for the collective resulting in ethical dilemmas.

To serve the purpose of adopting automated ethical adjudication systems is to resolve a conflict among agents that would not compromise adherence to the established norms. Bringsjord et al. (2021) have argued that automated ethical adjudication systems can be developed to oversee agent interactions so that ethical standards are effectively maintained during the decision-making process.

Below are two tables and the figure that visually demonstrate the policy implications of ethical governance in multi-agent systems.

Table 1: Key Ethical Principles for Multi-Agent Systems

Ethical Principle	Description	Policy Implication
Accountability	Holding creators and operators responsible for actions	Clear liability frameworks for unethical behavior or harm caused by agents (Baldoni et al., 2023)
Transparency	Ensuring decisions are explainable	Regulatory mandates for transparency in decision-making algorithms (Chaput et al., 2021)
Fairness	Preventing biased or discriminatory outcomes	Guidelines for designing non-biased algorithms (Ho & Wang, 2021)
Privacy	Protecting sensitive data	Privacy regulations for agent data use (Zeng et al., 2024)

Source: Adapted from Belloni et al. (2015) and Ho & Wang (2021).

Table 2: Proposed Governance Framework for Multi-Agent Systems

Governance Mechanism	Description	Potential Impact
Ethical Oversight Committees	Establishment of committees to oversee AI development	Ensures that ethical principles are upheld during the design and deployment phases (Renda, 2019)
Dynamic Regulatory Bodies	Flexible, adaptive regulatory bodies for evolving AI tech	Keeps pace with fast developments in AI and multi-agent technologies (Gandon, 2002)
International Standards	Creation of global AI governance standards	Harmonizes global efforts to ensure consistent ethical practices (Gal & Grosz, 2022)

Source: Based on Gandon (2002) and Renda (2019).

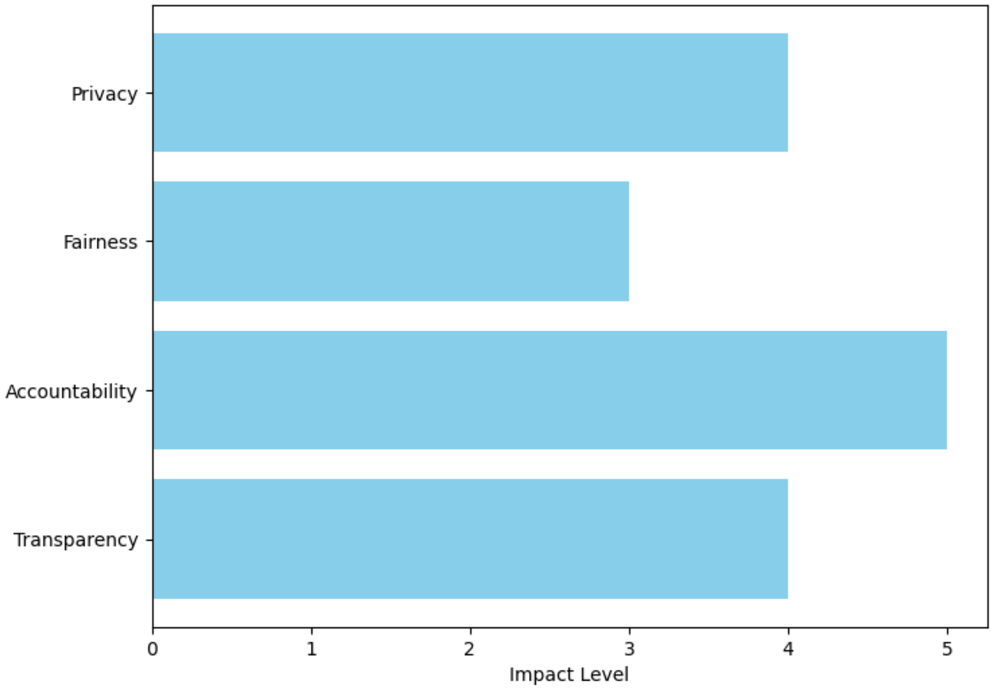

Figure 1: Ethical Decision-Making Framework for Multi-Agent Systems
Source: Based on the work of Belloni et al. (2015) and Gandon (2002).

The ever-changing environment in multi-agent systems will require teamwork between politicians and practitioners to invent living, see-through, accountable AI governance frameworks. The enforcement of these policy frameworks will legitimize equity and global collaboration to solve ethical issues that arise from autonomous agents while optimizing the advantages of AI that they experience.

FINAL THOUGHTS

This section provides an extensive final conclusion of the research, summarizing the findings from previous chapters and reflecting over the wider issues in AI governance and ethics concerning multi-agent systems. Given that there is a growing engagement of ethical issues related to multi-agent system applications in all sectors from makeshift autonomy, healthcare, finance to entertainment, it becomes pressing that the ethical challenges are understood and respective remedies are available.

Reflections on Ethical Governance

Uniquely, ethical governance is one part of the particular way to manage artificial intelligence (Rodriguez-Soto et al., 2023; Renda, 2019). Hence, it motivates the right development deployment of AI technologies demonstrated within multi-agent systems. Ethical governance necessitates the provision of rigorous and transparent algorithms apart from an adaptive regulatory framework for a rapidly shifting AI landscape. The results of our research underscore the complexities surrounding such undertakings and the urgent need for an inclusive approach involving AI developers, regulators, ethicists, and the broader public. Unlike risk and reward, the unanticipated, unintended, and uncontrollable consequences of developing advanced AI tools will continue rising into the limelight. Thus, it is paramount that AI ethical considerations are addressed right from the very start by compliant regulators, aiming at robust legal frameworks that feature transparency, accountability, and fairness in order for these values to truly trickle down the systems to society (Ganden, 2002; Ho & Wang, 2021).

The Effect of Ethical Decision-Making Frameworks

The research goes a long way to surface the nexus between the ethical decision-making framework and multi-agent systems. Despite this, agents are inherently autonomous and would have complex interests and ethical dilemmas that traditionally organized desired-making patterns cannot resolve. The very necessity to ensure agents act in a manner aligned with human values especially when they are operating with decentralization—and call for discerning governance is indicative of sophisticated governance models to address these ethical issues (Gal & Grosz, 2022).

Designing frameworks for regulation that prioritize ethical standing in terms of fairness, transparency, and accountability proffer multi-agent systems that are not only functional but socially responsible (Chaput et al., 2021). The study emphasizes designing ethical decision-making into the systems rather than as an afterthought.

Future Directions in AI Governance

The future development of AI governance will undoubtedly transgress the problems currently imposed by nascent technologies. For instance, recognizing the complexity of multi-agent systems that have qualities of unpredictability poses the greatest challenge to policy development, which must be adaptable to address issues as they occur (Zeng et al., 2024). In the next order of AI, mechanisms must be adaptive, so that policies formulated in Section 6 would continue to be pertinent, especially in the context of seeking a way through the maze of unpredictable forms in which AI continues to behave.

Beyond that, advances in building explainability into AI systems will constitute another major area of research and policy development. The accountability an explainable multi-agent system sets forth is paramount: the agent's decisions even when decided through AI are understandable or explainable to human users (Woodgate & Ajmeri, 2022). This is a condition for building trust and ensuring the AI systems are held accountable for their actions.

Lastly, international cooperation in AI governance is a must. AI technologies cross no borders; hence, there is a need for international collaboration on norms and standards and regulation. This engagement will permit legislators to join hands with stakeholders worldwide to ensure that the ethical concerns of AI development are universally addressed (Renda, 2019).

CONCLUSION

This paper deals with the sprawling intersections of ethics, governance, and policy implications in multi-agent systems (MAS). Autonomous agents and AI-driven systems continue to proliferate; the question of the ethics guiding their behavior has been crying out due to certain obvious-vices in its growth. From healthcare to robotic drivers, the systems hold promises of widespread benefits, while introducing a plethora of unique challenges: accountability, transparency, fairness, and responsibility.

The research work has just raised the bar at saying that ethical alignment of multi-agent systems is not entirely a technical issue. Interdisciplinary views are presented that draw from computer science, ethics, law, and public policy to emphasize the need for frameworks to govern AI effectively. These governance frameworks should be adaptable, comprehensive, and forward-looking visa resolving ethical problems emerging with the development of AI technology.

The pressing issue that has arisen from this is that regulatory frameworks must be dynamic and must be adapted as AI technology changes with rapid rates of innovation. As multi-agent systems become more widespread, the most important thing is that governance must be flexible enough to adjust to new ethical dilemmas and technical trends (Renda, 2019). In pursuance of the installment of trust and ensuring ethical conformance that stop from alienation from moral narratives are the integration of moral decision-making models along with earnest explanations and transparency (Woodgate & Ajmeri, 2022; Belloni et al., 2015).

Also highlighted is the strong need for international coordination in AI governance. The truly global AI! The need for international cooperation between a half a dozen states is crucial for ethical beliefs and regulatory standards (Ganden, 2002). Without a joint base, diversity of scattered discriminatory signals would consequently hinder the governance process efficacy.

The salutary role of the equilibrium between innovation and regulation in the unfolding saga of multi-agent systems cannot possibly be overemphasized. With AI's momentum, the onus of between the ethical behavior and governance on its technology-and annually year-active industry devolves. Henceforth, policymakers, AI developers, ethicists, and interested others need to work to find solutions that favor the proper use of AI, because it is precisely through them that the essence of what is due to individual humanness and to collective well-being may go forward. Indeed, autonomous agents in such a future would contribute ethically as well as accountably in good rapport with the rest of the world.

This survey is an important addition to the dialogue into the governance of AI and lays several building blocks for further research and later on policy debate. As the field of AI evolves, there will be an increasing demand for thorough reflection into ethical and societal implications of MAS if they shore up the delicate role of being developed and employed for humanity's sake.

REFERENCES

[1] Criado, N., Argente, E., & Botti, V. (2011). Open issues for normative multi-agent systems. AI communications, 24(3), 233-264..

[2] Chaput, R., Duval, J., Boissier, O., Guillermin, M., & Hassas, S. (2021, July). A multi-agent approach to combine reasoning and learning for an ethical behavior. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (pp. 13-23).

[3] Calvaresi, D., Calbimonte, J. P., Dubovitskaya, A., Mattioli, V., Piguet, J. G., & Schumacher, M. (2019). The good, the bad, and the ethical implications of bridging blockchain and multi-agent systems. Information, 10(12), 363.

[4] Gal, K., & Grosz, B. J. (2022). Multi-agent systems: Technical & ethical challenges of functioning in a mixed group. Daedalus, 151(2), 114-126.

[5] Belloni, A., Berger, A., Boissier, O., Bonnet, G., Bourgne, G., Chardel, P. A., ... & Zimmermann, A. (2015, January). Dealing with Ethical Conflicts in Autonomous Agents and Multi-Agent Systems. In AAAI Workshop: AI and Ethics.

[6] Ho, J., & Wang, C. M. (2021, September). Human-centered ai using ethical causality and learning representation for multi-agent deep reinforcement learning. In 2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS) (pp. 1-6). IEEE.

[7] Cointe, N., Bonnet, G., & Boissier, O. (2020). Ethics-based cooperation in multi-agent systems. In Advances in Social Simulation: Looking in the Mirror (pp. 101-116). Springer International Publishing.

[8] Chaffer, T. J., Goldston, J., Okusanya, B., & A I, G. D. (2024). On the ETHOS of AI Agents: An Ethical Technology and Holistic Oversight System. arXiv preprint arXiv:2412.17114.

[9] Lu, Q., Zhu, L., Xu, X., Whittle, J., Zowghi, D., & Jacquet, A. (2024). Responsible ai pattern catalogue: A collection of best practices for ai governance and engineering. ACM Computing Surveys, 56(7), 1-35.

[10] Rodriguez-Soto, M., Lopez-Sanchez, M., & Rodriguez-Aguilar, J. A. (2023). Multi-objective reinforcement learning for designing ethical multi-agent environments. Neural Computing and Applications, 1-26.

[11] Renda, A. (2019). Artificial Intelligence. Ethics, governance and policy challenges. CEPS Centre for European Policy Studies.

[12] Deshmukh, J. (2023, May). Emergent responsible autonomy in multi-agent systems. In Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems (pp. 3029-3031).

[13] Baldoni, M., Baroglio, C., Micalizio, R., & Tedeschi, S. (2023). Accountability in multi-agent organizations: from conceptual design to agent programming. Autonomous Agents and Multi-Agent Systems, 37(1), 7.

[14] Stenseke, J. (2024). Artificial virtuous agents in a multi-agent tragedy of the commons. AI & SOCIETY, 39(3), 855-872.

[15] Bojic, L., & Dapic, V. (2023). The Interplay of Social and Robotics Theories in AGI Alignment: Navigating the Digital City Through Simulation-based Multi-Agent Systems. In BISEC (pp. 58-63).

[16] Gahnberg, C. (2021). What rules? Framing the governance of artificial agency. Policy and society, 40(2), 194-210.\

[17] Chaput, R. (2022). Learning behaviours aligned with moral values in a multi-agent system: guiding reinforcement learning with symbolic judgments (Doctoral dissertation, Université Claude Bernard-Lyon I).

[18] Bringsjord, S., Govindarajulu, N. S., & Giancola, M. (2021). Automated argument adjudication to solve ethical problems in multi-agent environments. Paladyn, Journal of Behavioral Robotics, 12(1), 310-335.

[19] Talebirad, Y., & Nadiri, A. (2023). Multi-agent collaboration: Harnessing the power of intelligent llm agents. arXiv preprint arXiv:2306.03314.

[20] Rădulescu, R. (2024). The World is a Multi-Objective Multi-Agent System: Now What?. In ECAI 2024 (pp. 32-38). IOS Press.

[21] Woodgate, J. M., & Ajmeri, N. (2022, May). Macro ethics for governing equitable sociotechnical systems. In AAMAS'22: International Conference on Autonomous Agents and Multi-Agent Systems (pp. 1824-1828). IFAAMAS Press.

[22] SERAFIMOVA, S. (2022). The issue of trustworthiness in the HA-AV multi-agent system. Етически изследвания, 77-93.

[23] Deshmukh, J., Adivi, N., & Srinivasa, S. (2023, July). Resolving the dilemma of responsibility in multi-agent flow networks. In International Conference on Practical Applications of Agents and Multi-Agent Systems (pp. 76-87). Cham: Springer Nature Switzerland.

[24] Buechner, J., & Tavani, H. T. (2011). Trust and multi-agent systems: Applying the “diffuse, default model” of trust to experiments involving artificial agents. Ethics and information Technology, 13, 39-51.

[25] Zhao, J., & Yu, W. (2023). Quantum Multi-Agent Reinforcement Learning as an Emerging AI Technology: A Survey and Future Directions. Authorea Preprints.

[26] Gandon, F. (2002). Distributed Artificial Intelligence and Knowledge Management: ontologies and multi-agent systems for a corporate semantic web (Doctoral dissertation, Université Nice Sophia Antipolis).

[27] Cruz, C. J. X. (2024). Transforming Competition into Collaboration: The Revolutionary Role of Multi-Agent Systems and Language Models in Modern Organizations. arXiv preprint arXiv:2403.07769.

[28] Bezou-Vrakatseli, E., Brückner, B., & Thorburn, L. (2023, September). SHAPE: A framework for evaluating the ethicality of influence. In European Conference on Multi-Agent Systems (pp. 167-185). Cham: Springer Nature Switzerland.

[29] Zeng, Y., Wu, Y., Zhang, X., Wang, H., & Wu, Q. (2024). Autodefense: Multi-agent llm defense against jailbreak attacks. arXiv preprint arXiv:2403.04783.

[30] Hosseini, H. (2024, March). The fairness fair: Bringing human perception into collective decision-making. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 20, pp. 22624-22631).