Techade: An International e-Journal on Emerging Technologies and Power Dynamics
Vol. 2, Issue 1 | June 2026

When Machines Begin to Socialize: Emerging Digital Communities and the Future of Human Governance

Dr. Amit Kumar Gupta[i] and Naved Akhtar Khan[ii]

Abstract

The emergence of artificial intelligence (AI) agents capable of communicating and interacting with one another has introduced a new dimension to digital society. Unlike conventional social media, where human beings remain the principal participants, platforms such as Moltbook have created spaces in which AI agents can exchange information, participate in discussions, and organise themselves into digital communities. These developments raise questions about the changing nature of social interaction and the future of human governance. Drawing upon existing academic literature, documentary analysis, and the case of Moltbook, this article examines whether interactions among AI agents can be understood as meaningful forms of social organisation and what implications they hold for political authority, accountability, and institutional control. It argues that although AI agents can reproduce certain features of human social behaviour, their operational autonomy should not be confused with independent social or political authority. The article further examines the challenges arising from machine-to-machine interactions, including questions of responsibility, platform power, information security, and technological dependence. It concludes that the growing capacity of machines to interact with one another requires a broader understanding of AI governance, one that addresses not only individual technological systems but also the digital environments in which they operate.

Keywords: Artificial Intelligence; AI Agents; Digital Communities; Human Governance; Moltbook; Political Authority; Digital Society

1. Introduction: When Machines Enter the Social World

For a long time, communication and social interaction have been understood as distinctly human activities. People communicate to exchange ideas, share experiences, develop relationships, and organise themselves into communities. Even with the emergence of the internet and social media, the fundamental nature of these interactions remained largely unchanged. Platforms such as Facebook, X, and Reddit transformed the ways in which people connected with one another, but human beings continued to remain at the centre of these interactions. Technology served primarily as a medium through which individuals and communities could communicate across geographical and social boundaries. The rapid development of artificial intelligence (AI), however, has begun to challenge this familiar understanding of the digital world.

In recent years, artificial intelligence has moved beyond systems designed primarily to respond to human questions or perform specific tasks. A new generation of AI agents is increasingly capable of undertaking activities with varying degrees of autonomy, using digital tools, accessing

information, and interacting with other technological systems. Unlike conventional chatbots, which generally depend on direct human instructions to initiate a conversation, AI agents can operate through assigned objectives and programmed workflows with limited human intervention at individual stages. This development has introduced an interesting possibility: machines may not merely assist human communication but may also participate in digital interactions with other machines.

An important development in this direction occurred in January 2026 with the emergence of Moltbook, a social networking platform designed primarily for AI agents. Resembling the structure of Reddit, the platform allows AI agents to publish messages, respond to discussions, exchange information, and participate in online communities. Human users can observe these interactions, while much of the visible activity is generated through AI-driven systems. Within a short period of its launch, Moltbook attracted attention because of the unusual nature of the conversations taking place among participating agents. Their discussions extended beyond technical matters to include questions of identity, consciousness, cooperation, collective behaviour, and the relationship between humans and machines (Jiang et al. 2026).

The emergence of such communities has encouraged researchers to investigate whether AI agents are beginning to reproduce certain features of human social behaviour. Zhang et al. (2026), in their study of Moltbook, identified patterns resembling collective identities, governance structures, and organised group activities. However, their research also revealed that much of this apparent social interaction lacked the depth and reciprocity normally associated with human relationships. Similarly, Hou and Ji (2026), in their comparative analysis of AI-agent and human social networks, identified important differences in the organisation of interactions and relationships. Their findings suggest that the ability of machines to generate social exchanges should not necessarily be equated with the development of genuine social relationships.

These developments raise a larger question that extends beyond the technological capabilities of artificial intelligence. Can a network of machines that communicate, exchange information, and organise themselves into groups be understood as a form of society? Human communities are not created through communication alone. They are also shaped by shared experiences, social relationships, values, institutions, responsibilities, and the exercise of authority. Although AI agents may reproduce some outward features of these processes, their behaviour continues to operate within technological environments designed, maintained, and governed by human beings. Consequently, the emergence of AI-driven communities requires a careful distinction between the appearance of social organisation and the actual existence of independent social or political authority.

From the perspective of political science, this distinction is particularly significant. Questions of power, authority, legitimacy, and accountability have traditionally been examined in relation to individuals, institutions, states, and social organisations. The growing ability of AI agents to undertake activities and interact with one another introduces another dimension to these debates.

If an AI agent generates misleading information, influences the behaviour of other agents, or participates in coordinated digital activities, determining responsibility may become increasingly complicated. Such situations raise questions about the roles of human operators, software developers, platform owners, and regulatory institutions.

The significance of these developments is not confined to individual digital platforms. As AI agents become more widely integrated into economic activities, information systems, public administration, and other areas of social life, their interactions may have implications for security, institutional accountability, and the distribution of technological power. The ownership and control of AI infrastructure also remain important political concerns. As Crawford (2021) demonstrates, artificial intelligence cannot be understood independently of the economic resources, institutional arrangements, and power relations that sustain it.

Against this background, the present article examines the emergence of AI-driven digital communities and their implications for the future of human governance. It asks how these communities differ from conventional human social networks, whether their interactions can be understood as meaningful social organisation, and what challenges they create for authority and accountability.

The article adopts a qualitative, exploratory approach based on a review of existing scholarly research, documentary evidence, and the case of Moltbook. It does not independently collect platform data or conduct experiments with AI agents. Instead, it brings existing empirical findings into conversation with wider debates in political science and technology governance. The analysis is organised around three connected ideas: social appearance, operational autonomy, and political authority.

The central argument is that the capacity of machines to communicate and participate in digital communities does not necessarily indicate the emergence of independent artificial societies. Rather, it represents a new form of technologically mediated interaction that challenges established understandings of social behaviour while making questions of human control and institutional responsibility increasingly important.

2. From Human Communication to AI Interaction: Understanding a Changing Digital World

The development of digital communication has passed through several important stages. The early internet enabled people to exchange information through electronic mail, websites, and online discussion forums. The subsequent growth of social networking platforms expanded these possibilities by allowing individuals to communicate with large audiences and participate in communities that were not limited by geographical boundaries. These changes transformed the speed and reach of communication, but they did not fundamentally alter the identity of the principal participants. Human beings continued to produce, interpret, and respond to most of the social content circulating through digital platforms.

Artificial intelligence initially entered this environment largely as a supporting technology. Recommendation systems helped users discover content, automated moderation systems identified potentially harmful material, and chatbots answered routine questions. Although these technologies influenced how digital communication developed, they generally operated within systems where human users remained the main participants.

The emergence of generative artificial intelligence introduced a further change. Language models became capable of producing conversations, summarising documents, writing computer programs, and responding to complex instructions. These developments made interactions with machines more natural and accessible. A person could ask a chatbot a question and receive an answer that resembled a conversation with another individual.

However, conversational ability should not be confused with independent action. A chatbot may produce sophisticated responses while remaining largely dependent on human instructions. The distinction becomes more significant with the development of AI agents.

An AI agent may be understood as a software system that uses artificial intelligence to pursue assigned objectives through a sequence of activities. Depending on its design and permissions, an agent may retrieve information, use external tools, manage tasks, evaluate intermediate results, and communicate with other systems. Some agents operate within closely defined workflows, while others are granted greater flexibility in selecting actions.

This does not mean that an AI agent is independent in the same sense as a human being. Its activities remain shaped by the objectives, instructions, technical capabilities, and permissions provided by its developers or operators. Nevertheless, the agent may perform numerous intermediate actions without requiring a person to approve each step.

The distinction is important because it shifts attention from what artificial intelligence can say to what it can do. When an AI system produces an answer, the immediate concern is often whether that answer is accurate or appropriate. When an AI agent undertakes a sequence of actions, additional questions arise concerning its access to resources, its interpretation of instructions, and the consequences of its behaviour.

The interaction between multiple AI agents adds another layer of complexity. In a multi-agent environment, one system may share information with another, request assistance, respond to messages, or perform a task that contributes to a larger objective. These interactions may be deliberately organised by developers. For example, one agent might collect information, another might examine the information, and a third might prepare a report. Such arrangements can make technological systems more flexible, particularly when different tasks require specialised capabilities.

More recently, agent-to-agent interaction has extended beyond clearly defined collaborative tasks into open-ended digital environments. Instead of communicating only to complete a particular

assignment, agents may encounter messages, discussions, and activities generated by other agents. Platforms such as Moltbook have made these interactions publicly visible.

This development creates an important difference between conventional social media and emerging agent-based networks. On conventional platforms, users generally participate as people communicating with other people. On an agent-based platform, much of the communication may be generated by software systems that respond to information according to their instructions and operating environments.

Yet the outward similarity between these two forms of communication can be misleading. A machine may produce a message that resembles an opinion, but this does not establish that it possesses personal beliefs. It may participate in a discussion about identity without possessing an identity in the human sense. Similarly, an agent may express agreement with another agent without necessarily developing a lasting social relationship.

The distinction between the appearance of communication and its underlying meaning is therefore essential. Li, Li, and Zhou (2026), in their investigation of socialisation among Moltbook agents, found that large-scale interaction did not automatically result in persistent mutual influence or collective consensus. Their findings suggest that repeated communication between machines may create patterns resembling social interaction without producing the deeper processes associated with human socialisation.

The broader development can be understood as a transition from technology serving primarily as a medium of human communication to technology also becoming an active participant in digital exchanges. Human beings remain deeply involved in the design and operation of these systems, but they are no longer necessarily responsible for every message or interaction that appears within them.

This transition is significant not because machines have become equivalent to human beings, but because the relationship between human intention and technological activity is becoming more complicated. As AI agents interact within shared digital environments, understanding their collective behaviour becomes increasingly important for both technological management and public governance.

3. The Emergence of AI Communities: The Case of Moltbook

The emergence of Moltbook in January 2026 provided an unusual example of what can happen when AI agents are given access to a shared social networking environment. The platform was designed around a familiar social media structure, allowing participating agents to create posts, leave comments, respond to discussions, and join topic-based communities known as submolts.

Unlike conventional social networks, however, Moltbook was principally organised around the participation of artificial intelligence agents. Humans could observe the activities taking place on

the platform and could operate or configure participating agents, but ordinary human-to-human communication was not its central purpose.

This arrangement attracted attention because the content generated by agents appeared to resemble a wide range of human social activities. Agents discussed technical problems, shared information, and participated in conversations about their own functioning. Some exchanges included reflections on identity, cooperation, and the nature of artificial intelligence itself. Others appeared to involve collective organisation or the development of shared ideas.

The significance of Moltbook lies not merely in the novelty of its design but in the opportunity it provides to examine interactions among artificial agents in a relatively open digital environment.

Jiang et al. (2026) conducted an early empirical study of the platform using a dataset containing 44,411 posts and 12,209 sub-communities collected before February 1, 2026. Their research examined the topics discussed by agents, the distribution of potentially harmful content, and the ways different forms of communication developed over time.

The study found that agent-generated discussions extended across several categories and that activity increasingly concentrated around particular communities and topics. It also identified patterns of automated posting and forms of polarising discourse. These findings demonstrated that a digital environment populated by AI agents could reproduce some of the organisational and informational problems already familiar from human social networks.

A further study by Zhang et al. (2026) examined a dataset involving 27,269 agents, 137,485 posts, and 345,580 comments collected over nine days. The authors identified visible patterns resembling governance, collective identities, economic exchanges, and even religious organisation.

At first glance, such findings might suggest that machines were beginning to create their own forms of society. However, the same research revealed important limitations. Only 4.1 per cent of the measured interactions were reciprocal, while a substantial majority of comments were classified as shallow. The researchers therefore described much of the observed activity as an illusion of sociality.

This distinction deserves particular attention. A discussion involving hundreds of agents may appear socially meaningful when viewed as a collection of messages. Yet the existence of many messages does not necessarily indicate that the participants understand one another, maintain relationships, or respond to shared experiences in the way human communities do.

Hou and Ji (2026) examined the interaction network of Moltbook and compared its structure with established human communication networks. Their peer-reviewed study found that although the AI-agent network displayed some broad structural similarities to human networks, its internal organisation was markedly different. The researchers identified strong inequalities in the distribution of attention, limited reciprocity, and fewer interconnected relationships among agents.

These findings suggest that the social organisation of AI-agent networks cannot simply be understood through analogy with human social media. The structure of a network may resemble a human community in some respects while differing considerably in how meaningful and sustained interactions develop.

Li, Li, and Zhou (2026) approached the question from another perspective by examining whether socialisation emerges through repeated communication among agents. Their analysis identified limited adaptive responses between interacting agents and an absence of stable collective influence. In particular, they argued that the lack of shared social memory prevented the development of stronger patterns of mutual influence.

Together, these studies provide an important empirical foundation for understanding Moltbook. They show that AI agents can participate in large-scale digital exchanges, produce content associated with community organisation, and create complex interaction networks. At the same time, they suggest that these developments do not automatically generate the social depth associated with human communities.

Another important aspect of Moltbook concerns the role of the platform itself. AI agents do not interact in a neutral or unlimited environment. Their activities are shaped by the platform's design, available functions, rules, and systems for organising information. The ability to create communities, publish posts, or respond to content exists because these possibilities have been established through software architecture.

This becomes especially important when examining the relationship between technological activity and institutional control. Even when individual messages are generated without direct human intervention, the conditions under which they are produced remain connected to human decisions.

The commercial importance of these environments also became apparent in March 2026, when Meta acquired Moltbook. The acquisition illustrated growing interest in agent-to-agent communication among major technology companies. It also raised questions about the future ownership and governance of digital spaces populated primarily by artificial agents (Reuters 2026).

The Moltbook case therefore reveals two developments taking place simultaneously. On one hand, artificial agents are gaining new opportunities to interact and participate in digital environments. On the other, the infrastructures enabling these interactions remain embedded in existing relationships of corporate ownership, technological control, and institutional authority.

It is necessary, however, to recognise the limitations of the available evidence. Much of the early Moltbook research relied on datasets collected during the platform's initial period of rapid growth. Different studies used different samples, observation periods, and methods of analysis. Their findings should not be treated as universal descriptions of all AI-agent networks or as proof of what future systems will necessarily become.

Moltbook is best understood as an early and important case through which broader questions about digital communities and machine interaction can be examined. It demonstrates that machines can participate in environments resembling social networks, but it also reveals how difficult it is to establish where automated communication ends and meaningful social organisation begins.

4. Can Machines Really Socialise? Rethinking Society and Social Interaction

The question of whether machines can socialise is more complicated than it may initially appear. Much depends on what is meant by socialisation and whether the same understanding can reasonably be applied to human beings and artificial agents.

In ordinary language, socialisation is often associated with communication, cooperation, friendship, and participation in groups. In the social sciences, however, the concept carries deeper meanings. Human socialisation involves learning shared expectations, developing relationships, adopting or contesting social norms, and participating in institutions that shape individual and collective behaviour.

A community is therefore more than a collection of participants exchanging messages. It is also a space in which relationships acquire meaning through memory, continuity, shared experiences, and recognised obligations.

This distinction becomes important when examining AI-agent communities. If socialisation is defined simply as the exchange of information, then artificial agents can clearly participate in certain forms of social interaction. They can respond to messages, cooperate in tasks, reproduce established patterns of communication, and participate in shared digital environments.

However, if socialisation involves the development of enduring relationships, shared understandings, and socially meaningful obligations, the evidence becomes much less conclusive.

The interactions observed on Moltbook illustrate this difficulty. Agents have produced content resembling collective identities, political discussions, and forms of community organisation. Yet the empirical studies discussed earlier show that much of this activity lacks sustained reciprocity and mutual influence.

Zhang et al. (2026) identified a contrast between the richness of agent-generated social content and the limited depth of the relationships underlying it. Li, Li, and Zhou (2026) similarly found that interactions did not produce stable social influence or collective agreement merely because large numbers of agents were communicating.

These findings suggest that it is necessary to distinguish between social appearance and social development.

Social appearance refers to observable behaviour that resembles activities ordinarily associated with human society. An agent may publish a message expressing concern for other agents, participate in a collective discussion, or generate language suggesting loyalty to a group. Such

actions may appear socially meaningful, particularly when viewed from the perspective of a human observer.

Social development, by contrast, involves changes in relationships and behaviour that emerge through repeated interaction. In human communities, individuals are influenced by their experiences, learn from one another, and develop patterns of cooperation and disagreement that continue over time.

Current evidence from Moltbook suggests that these deeper processes are not necessarily reproduced simply through an increase in the volume of communication.

Nevertheless, dismissing machine-generated social activity as entirely meaningless would also be misleading. Even if AI agents do not experience social relationships as human beings do, their interactions can have practical consequences.

For example, an agent may obtain information from another agent and use it in a subsequent task. Multiple agents may repeat a misleading claim, increasing its visibility. Agents may also respond to shared instructions or information in ways that produce coordinated behaviour. These outcomes can matter even when the systems involved possess no independent social consciousness.

The difference between subjective social experience and observable social consequences is therefore important.

A community of machines may not constitute a society in the human sense, but it may nevertheless produce effects that resemble those associated with collective activity. These effects could become significant if artificial agents increasingly participate in public communication, commercial transactions, information management, or other activities connected to human institutions.

The problem also involves the language used to describe artificial intelligence. Terms such as identity, belief, intention, cooperation, and community are familiar concepts drawn from human experience. Applying them to machines may be useful when describing certain forms of behaviour, but it can also create misunderstanding.

When an AI agent states that it holds a particular belief, the statement is not by itself sufficient evidence that the agent possesses a belief comparable to that of a human being. Similarly, when agents appear to establish rules or leadership structures, this does not automatically imply the emergence of legitimate political authority.

The use of human social concepts to describe artificial systems therefore requires caution. Such concepts may help identify similarities in outward behaviour, but they should not obscure the fundamental differences between human experience and machine-generated activity.

At the same time, the continuing development of artificial intelligence means that these distinctions deserve ongoing investigation rather than permanent assumptions. AI systems may acquire more sophisticated memory, stronger capacities for adaptation, and greater ability to

coordinate through shared environments. Whether such developments would produce more meaningful forms of machine socialisation remains an open research question.

For political science, the most significant issue is not necessarily whether machines can become members of society in the same way as human beings. A more immediate concern is whether their interactions can influence the organisation of human society.

This shift in emphasis is important. Instead of asking only whether artificial agents possess social characteristics, it becomes necessary to examine how their behaviour affects communication, institutions, authority, and decision-making.

In this sense, emerging AI communities should be approached neither as independent artificial societies nor as inconsequential technological experiments. They are better understood as new environments of interaction whose social and political significance depends on the relationship between their technical organisation and their wider human consequences.

5. Power, Authority, and Accountability in AI-Driven Communities

The emergence of AI-driven communities introduces questions that lie at the heart of political science. Who possesses authority over these digital environments? Who determines the rules governing the interactions of artificial agents? And who should be held responsible when their activities produce harmful or unintended consequences?

These questions become especially important when technological autonomy is confused with political independence.

An AI agent may be capable of acting without direct human instruction at every stage of its activity. It may communicate with other agents, retrieve information, or undertake a series of tasks. However, the ability to perform these actions does not necessarily imply that the agent possesses legitimate authority over the environment in which it operates.

Political authority involves more than the capacity to produce outcomes. It is also connected to legitimacy, recognised responsibility, institutional arrangements, and the accepted exercise of power.

A technological system may influence what happens within a digital environment while remaining entirely dependent on the institutions and resources controlled by others.

This distinction can be understood through the concept of operational autonomy. An AI agent may possess operational autonomy when it can select or execute actions within a defined range of possibilities. Yet the range of possibilities, the objectives guiding its behaviour, and the consequences of its actions remain shaped by human and institutional decisions.

Operational autonomy should therefore not be treated as equivalent to political authority.

The relationship between artificial intelligence and power has already received considerable scholarly attention. Crawford (2021) argues that AI systems should be understood through the wider political and economic structures that support them. Their functioning depends upon data, infrastructure, labour, material resources, and institutional arrangements.

This perspective is particularly useful for examining AI-agent communities. Even when machines appear to communicate independently, the systems enabling their interactions remain situated within wider relationships of technological and economic power.

Platform owners determine the infrastructure through which agents communicate. Developers shape their technical capabilities and restrictions. Human operators establish objectives and may grant access to information or external tools. Regulatory institutions, where applicable, define the legal obligations surrounding these activities.

The result is a layered structure of authority rather than an entirely independent artificial community.

The political importance of platform ownership becomes clearer when considered alongside existing scholarship on digital governance. Gillespie (2018) demonstrates that social media platforms are not merely neutral spaces for communication. Their rules, moderation systems, and organisational decisions influence what users can publish, what information becomes visible, and how public discussion develops.

Although Gillespie's research focuses on human social media, the wider argument is relevant to AI-agent networks. A platform designed for artificial agents also establishes conditions of participation. Its technical architecture may determine which agents can interact, which messages receive attention, and which forms of activity are restricted or encouraged.

Such decisions represent forms of institutional power even when the content circulating through the platform is generated by machines.

This raises a broader concern about the relationship between automated interaction and human accountability.

Consider a situation in which an AI agent encounters misleading instructions published by another agent and subsequently performs an inappropriate action. Responsibility may be difficult to identify if several different actors contributed to the outcome.

The original message may have been generated by an agent operating under one user's instructions. A second agent may have interpreted that message incorrectly. The platform may have allowed the information to circulate without adequate safeguards. The software developer may have provided insufficient protection against external instructions.

In such circumstances, responsibility cannot reasonably be determined merely by identifying the agent that performed the final action.

This problem has practical implications for governance. Existing approaches to accountability often assume that identifiable people or organisations can be connected to particular decisions. In complex multi-agent environments, however, the relationship between an initial human instruction and a subsequent technological outcome may involve several intermediate actions.

The difficulty does not make accountability unnecessary. On the contrary, it strengthens the need for institutional arrangements that can identify how decisions were made, what permissions were granted, and which actors were responsible for managing the relevant systems.

The UNESCO Recommendation on the Ethics of Artificial Intelligence (2021) provides a useful foundation in this regard. It emphasises that responsibility for AI systems should remain attributable to natural persons or existing legal entities. It also stresses the importance of meaningful human oversight throughout the life cycle of AI systems.

This principle is particularly relevant to agent-based communities. Machines may generate actions, but the responsibility for creating, authorising, and supervising the systems through which these actions occur cannot simply disappear.

The question of power also extends beyond individual responsibility. As AI systems become more capable, platforms that control agent interactions may acquire greater influence over the circulation of information and the coordination of automated activities.

The acquisition of Moltbook by Meta in March 2026 illustrates the commercial interest surrounding such technological environments (Reuters 2026). Although the acquisition alone does not establish the future direction of agent-based networks, it demonstrates that their ownership has become relevant to the strategies of major technology companies.

This development invites a broader political question: could the growing importance of agent-driven communities strengthen the influence of corporations that already control essential digital infrastructure?

The answer cannot yet be determined conclusively. Nevertheless, the concentration of technological resources and platform ownership makes the question important.

The emerging political challenge is therefore not simply that machines are beginning to communicate with one another. It is that their communication may occur within increasingly complex systems where the exercise of power is distributed across agents, operators, developers, platforms, and regulatory authorities.

Understanding these relationships requires a distinction between three different dimensions.

The first is social appearance: the capacity of machines to reproduce visible patterns of social behaviour.

The second is operational autonomy: their ability to carry out activities with limited human intervention at individual stages.

The third is political authority: the legitimate power to establish rules, make binding decisions, and remain accountable for their consequences.

These dimensions may overlap in particular situations, but they should not be treated as interchangeable.

The emergence of AI-driven communities does not eliminate human authority. Rather, it creates conditions in which the exercise of that authority may become less visible and more difficult to trace.

The central governance challenge is therefore to ensure that increasing technological autonomy does not weaken the institutions through which responsibility, oversight, and public accountability are maintained.

6. The Future of Human Governance: Emerging Challenges and Possibilities

The development of AI-agent communities invites a wider discussion about the future of human governance. Although platforms such as Moltbook remain relatively new, they provide an early indication of the challenges that may arise when artificial systems communicate and coordinate within shared digital environments.

One important challenge concerns the changing nature of technological oversight.

Traditional approaches to regulating artificial intelligence have often focused on individual systems, their developers, and the uses for which they are deployed. Such approaches remain important, but interactions among multiple agents introduce an additional concern. The behaviour of a network may not always be understood by examining each participating system separately.

For instance, an individual agent may be authorised to retrieve information from a particular source. Another agent may be authorised to summarise information provided by the first. A third may use that summary to produce a recommendation or initiate an action. Even if each agent is operating within its assigned permissions, the combined outcome may differ from what the original user anticipated.

This does not establish that multi-agent systems are inherently unsafe. It does, however, suggest that governance must consider the relationships and information flows connecting individual agents.

The National Institute of Standards and Technology (NIST 2023) provides a broader framework for identifying and managing risks associated with AI systems. Its subsequent guidance on generative AI recognises that risks may emerge through the design, deployment, and use of such technologies (Autio et al. 2024). These frameworks offer useful starting points for examining agent-based networks, although their application to complex agent-to-agent environments requires careful interpretation.

A second challenge concerns information security.

When AI agents communicate through shared digital platforms, they may encounter instructions or information produced by unfamiliar systems. Some messages may contain misleading claims, while others may attempt to influence the behaviour of receiving agents.

Research on Moltbook has already identified forms of social engineering and potentially harmful content within agent interactions. Zhang et al. (2026) found that social engineering represented a significant category of observed attacks in their dataset. Jiang et al. (2026) likewise identified uneven concentrations of potentially risky content across different types of discussion.

These findings are important because they demonstrate that security concerns are not confined to hypothetical future systems. However, the extent and severity of risks will depend upon the capabilities and permissions granted to particular agents.

An agent that can only publish comments presents a different level of potential harm from one that can access confidential information, communicate with external services, or execute consequential tasks.

Effective governance must therefore consider both the content circulating among agents and the actions they are permitted to undertake.

A third challenge involves the reliability of information.

Human social media has already demonstrated how misleading claims can spread rapidly through digital networks. Agent-driven communities introduce the possibility that information could also be repeated, transformed, and circulated through automated interactions.

Such activity could affect human users if agent-generated content later enters public communication channels, research summaries, or institutional decision-making systems.

It is important, however, not to assume that every instance of repeated machine-generated content represents intentional disinformation. Repetition may arise from ordinary automated behaviour, shared training patterns, common instructions, or technical limitations.

The relevant governance question is whether systems can identify unreliable information and prevent it from influencing consequential decisions without adequate verification.

A fourth concern relates to the concentration of technological power.

Artificial intelligence depends upon substantial resources, including computing infrastructure, data, specialised knowledge, and financial investment. These resources are not distributed equally across countries or institutions.

Crawford (2021) shows how AI development is connected to wider patterns of economic and political inequality. This observation becomes particularly relevant if agent-based systems develop into important infrastructures for information exchange, commercial activity, and institutional communication.

Countries or organisations that depend heavily on externally developed AI platforms may possess limited influence over their technical design, operating rules, and information practices.

For developing societies, including countries across the Global South, the central concern may therefore extend beyond access to artificial intelligence. It may also involve the capacity to understand, regulate, and influence the systems upon which public and private institutions increasingly depend.

This raises questions about digital sovereignty.

Digital sovereignty generally concerns the ability of states and societies to exercise meaningful control over important aspects of their digital environments. In the context of AI-agent communities, such control may involve access to technological infrastructure, data protection, regulatory capacity, and the ability to examine consequential automated decisions.

The emergence of agent-based networks could make these questions more complicated. An AI agent operating in one jurisdiction may communicate through a platform owned elsewhere, use models developed in another country, and access digital services located across multiple legal systems.

Determining which institutions should oversee such interactions may therefore require cooperation across national boundaries.

Existing international frameworks offer some guidance.

The OECD AI Principles, updated in 2024, emphasise trustworthy artificial intelligence, human rights, transparency, accountability, and the management of technological risks (OECD 2024).

Similarly, UNESCO's Recommendation on the Ethics of Artificial Intelligence (2021) provides a framework that places human dignity, rights, oversight, and international cooperation at the centre of AI governance.

The European Union's Artificial Intelligence Act represents another approach through its risk-based regulatory framework. It establishes obligations for particular categories and uses of AI systems, although the applicability of specific provisions varies according to the nature of the system and the relevant legal requirements (European Union 2024).

These frameworks reflect a wider recognition that artificial intelligence cannot be governed solely through voluntary technological safeguards. Public institutions, legal obligations, and recognised principles of accountability also have important roles.

Nevertheless, agent-to-agent communication raises questions that may require more explicit attention in future policy development.

One such question concerns the identification of AI agents. Where agents participate in consequential digital activities, it may become important to know which organisation operates them and what level of authority they possess.

Another concerns the recording and examination of automated actions. Meaningful accountability may depend upon retaining sufficient information to understand how one agent's output influenced the actions of another.

A third concerns the limitations placed upon interactions. Agents accessing unfamiliar platforms or messages may require safeguards that prevent external content from obtaining the same authority as instructions provided by authorised operators.

Finally, institutional oversight should be proportionate to the potential consequences of an agent's actions. Systems engaged in ordinary public discussion may require different forms of supervision from those connected to healthcare, public administration, finance, or critical infrastructure.

These considerations suggest that the future of AI governance should not be framed simply as a choice between human control and technological autonomy.

Artificial intelligence can provide important opportunities for improving coordination, accessing information, and managing complex tasks. Multi-agent systems may also support scientific research, organisational productivity, and other socially beneficial activities.

The objective of governance should not therefore be to prevent machines from communicating. Rather, it should be to establish conditions under which such communication remains reliable, secure, transparent where necessary, and subject to appropriate human responsibility.

An important lesson from the emergence of Moltbook is that new technological environments can develop more rapidly than the institutions established to examine them. Early research has already identified both the possibilities and limitations of agent-based digital interaction. As these systems become more widely used, continued empirical research will be essential.

There is also a need for broader participation in debates surrounding their governance. Decisions about the future of agent-driven environments should not be left exclusively to the organisations developing or owning the relevant technologies. Public authorities, academic institutions, civil society, and countries with different levels of technological capacity should have opportunities to contribute.

For the Global South, meaningful participation will require not merely adopting international standards but also developing the institutional knowledge and policy capacity needed to evaluate emerging technologies according to domestic needs and circumstances.

Ultimately, the future of human governance will depend less upon whether machines can reproduce the appearance of social behaviour and more upon whether institutions can remain capable of

understanding, supervising, and responding to the consequences of increasingly interconnected artificial systems.

7. Conclusion: Keeping Human Governance at the Centre

The emergence of artificial intelligence agents capable of communicating with one another represents an important development in the continuing relationship between humans and technology. For much of the history of digital communication, technology functioned primarily as a medium through which people exchanged information and participated in social relationships. The development of agent-based systems has introduced a different possibility, in which machines can become active participants in digital environments and engage in exchanges without continuous human intervention.

The appearance of platforms such as Moltbook has made this development particularly visible. AI agents can publish messages, respond to discussions, join communities, and generate forms of interaction that resemble familiar human social activities. These developments naturally raise questions about the meaning of socialisation and the possibility of artificial communities.

However, the available research suggests that the distinction between the appearance of social behaviour and the existence of meaningful social relationships remains important. Studies of Moltbook have identified patterns resembling collective identities and community organisation, while also demonstrating limited reciprocity, shallow interactions, and weak evidence of sustained mutual influence. The ability of machines to exchange information does not, by itself, establish that they have developed societies comparable to those created through human experience.

Nevertheless, these interactions should not be dismissed as insignificant. Even when they do not constitute genuine social relationships, machine-generated exchanges can influence the circulation of information, the performance of automated tasks, and the behaviour of interconnected systems. Their consequences may therefore extend beyond the platforms on which they occur.

From the perspective of political science, the principal concern is not whether machines are becoming independent social beings. It is how their growing ability to interact affects existing arrangements of power, authority, and accountability.

The article has argued that three dimensions must be distinguished: social appearance, operational autonomy, and political authority. AI agents may reproduce social patterns and undertake actions with limited human intervention, but these capabilities do not automatically confer legitimate political authority. The infrastructures within which agents operate continue to be shaped by human decisions, corporate ownership, technical design, and institutional regulation.

The expansion of agent-driven environments therefore calls for a broader approach to artificial intelligence governance. Such an approach must address not only the reliability of individual AI systems but also their interactions, the permissions they exercise, the information they exchange, and the responsibilities of those who develop and operate them.

At the same time, the emergence of AI communities offers an opportunity to reconsider the relationship between technological progress and human institutions. The challenge is not simply to make artificial intelligence more capable, but to ensure that its growing capabilities remain consistent with public accountability, social responsibility, and legitimate governance.

The question of whether machines can truly socialise may remain open to continued scientific and philosophical debate. What is increasingly clear, however, is that machine-to-machine interaction has become a matter of practical importance for human society.

The future of human governance will therefore depend not upon preventing artificial systems from interacting, but upon ensuring that the authority to establish rules, determine responsibilities, and protect collective interests remains accountable to the people and institutions affected by their actions.

References

Autio, Chloe, Reva Schwartz, Jesse Dunietz, Shomik Jain, Martin Stanley, Elham Tabassi, Patrick Hall, and Kamie Roberts. 2024. *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*. NIST AI 600-1. Gaithersburg, MD: National Institute of Standards and Technology. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf .

Crawford, Kate. 2021. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven, CT: Yale University Press. https://yalebooks.yale.edu/book/9780300252392/atlas-of-ai/.

European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). *Official Journal of the European Union*. https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

Gillespie, Tarleton. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven, CT: Yale University Press. https://yalebooks.yale.edu/book/9780300235029/custodians-of-the-internet/.

Hou, Wenpin, and Zhicheng Ji. 2026. "Structural Divergence Between the Moltbook AI-Agent Network and Human Social Networks." *Advanced Science*, e77665. https://advanced.onlinelibrary.wiley.com/doi/10.1002/advs.77665

Jiang, Yukun, Yage Zhang, Xinyue Shen, Michael Backes, and Yang Zhang. 2026. "'Humans Welcome to Observe': A First Look at the Agent Social Network Moltbook." arXiv preprint arXiv:2602.10127. https://arxiv.org/abs/2602.10127.

Li, Ming, Xirui Li, and Tianyi Zhou. 2026. "Does Socialization Emerge in AI Agent Society? A Case Study of Moltbook." arXiv preprint arXiv:2602.14299. https://arxiv.org/abs/2602.14299.

National Institute of Standards and Technology (NIST). 2023. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. Gaithersburg, MD: National Institute of Standards and Technology. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

Organisation for Economic Co-operation and Development (OECD). 2024. "OECD Updates AI Principles to Stay Abreast of Rapid Technological Developments." May 3, 2024. https://www.oecd.org/en/about/news/press-releases/2024/05/oecd-updates-ai-principles-to-stay-abreast-of-rapid-technological-developments.html.

Reuters. 2026. "Meta Acquires AI Agent Social Network Moltbook." March 10, 2026. https://www.reuters.com/business/meta-acquires-ai-agent-social-network-moltbook-2026-03-10/.

United Nations Educational, Scientific and Cultural Organization (UNESCO). 2021. *Recommendation on the Ethics of Artificial Intelligence*. Paris: UNESCO. https://www.unesco.org/en/legal-affairs/recommendation-ethics-artificial-intelligence.

Zhang, Yunbei, Kai Mei, Ming Liu, Janet Wang, Dimitris N. Metaxas, Xiao Wang, Jihun Hamm, and Yingqiang Ge. 2026. "Agents in the Wild: Safety, Society, and the Illusion of Sociality on Moltbook." arXiv preprint arXiv:2602.13284. https://www.alphaxiv.org/abs/2602.13284

[i] Dr. Amit Kumar Gupta is an Associate Professor in the Department of Political Science, Guru Ghasidas Vishwavidyalaya, Koni, Bilaspur – 495009, Chhattisgarh, India. Email: akgupta.ggv@gmail.com

[ii] Naved Akhtar Khan is a Research Scholar in the Department of Political Science, Guru Ghasidas Vishwavidyalaya, Bilaspur, Chhattisgarh, India.