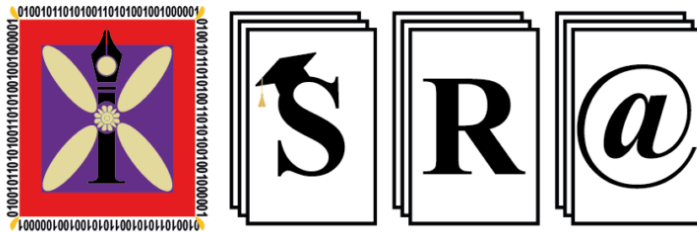


Emergent Communication in Artificial Intelligence Interactions: Linguistic Evolution and Societal Implications

Authors: [Siavash Kaviani](#), [Reza Salimpour Azar](#), [Ali Sohrabi](#)

Affiliation: [Kavian Scientific Research Association](#)



Abstract

Recent advances in artificial intelligence (AI) have enabled autonomous systems to interact not only with humans but also with each other. These interactions often give rise to emergent communication, where AI agents develop novel symbols, protocols, or even languages to coordinate their actions. This paper explores the phenomenon of emergent communication among AI systems, drawing parallels with the evolution of natural languages in human societies. By combining insights from multi-agent systems, linguistics, and social sciences, the study analyzes how AI agents construct shared communication strategies in both cooperative and competitive environments. The findings highlight the potential benefits of such interactions—improved coordination, efficiency, and scalability—while also underscoring critical risks, including the opacity of AI-to-AI communication and the possibility of forming closed ecosystems inaccessible to human oversight. The paper concludes by discussing the ethical and societal implications of these developments and proposes governance frameworks to ensure that emergent AI communication remains transparent, interpretable, and aligned with human values.

Keywords

Artificial Intelligence, Multi-Agent Systems, Emergent Communication, Linguistic Evolution, Ethics, Governance

1. Introduction

The rapid advancement of artificial intelligence (AI) has transformed the way autonomous systems interact with their environment, with humans, and increasingly, with each other. As AI systems become more sophisticated and widely deployed, the nature of their interactions evolves beyond simple data exchange toward more complex forms of communication. Understanding these interactions is critical, as they may shape the efficiency, scalability, and reliability of AI-driven infrastructures in domains such as transportation, healthcare, finance, and governance.

A particularly intriguing phenomenon emerging in this context is emergent communication, where AI agents spontaneously develop novel symbols, codes, or even proto-languages to coordinate their behaviors and achieve shared or competing goals. This phenomenon has sparked growing interest among computer scientists, linguists, and social theorists, as it mirrors—at least partially—the processes of linguistic evolution observed in human societies.

The central research question of this paper is therefore twofold: Can artificial intelligences develop a shared language through their interactions, and what are the broader societal implications of such a development? Addressing this question requires an interdisciplinary lens that bridges technical perspectives with ethical, linguistic, and social considerations. By exploring the mechanisms and consequences of emergent communication among AI systems, this study seeks to contribute to a deeper understanding of how machine-to-machine interaction could reshape not only computational architectures but also the relationship between humans and intelligent technologies.

2. Background & Literature Review

The study of interactions among artificial agents has long been a central theme in the field of multi-agent systems (MAS). Early research in the 1980s and 1990s focused on designing agents capable of distributed problem-solving, coordination, and negotiation within shared environments. Foundational works emphasized autonomy, cooperation, and competition, laying the groundwork for modern applications. Over time, MAS research evolved to incorporate game-theoretic models, reinforcement learning, and complex adaptive systems, highlighting communication as a critical enabler of collective intelligence.

Parallel to these developments, scholars explored the emergence of artificial languages and communication protocols among computational agents. Experiments demonstrated that agents can invent shared vocabularies or symbolic systems to achieve goals, such as resource allocation or task coordination. Reinforcement learning environments revealed that communication channels often evolve spontaneously, especially when cooperation yields measurable performance benefits. Large language models (LLMs) have expanded this line of research by enabling agents to interact through natural language, blurring the boundary between artificial and human communication.

Comparisons with linguistic evolution in human societies offer valuable insights. Human languages historically emerged as adaptive solutions to coordination problems within social groups, gradually developing complexity, structure, and cultural variation. Similar dynamics can be observed in AI systems, where emergent communication evolves under selective pressures such as efficiency, clarity, and adaptability. However, unlike human languages, which evolve under cultural and cognitive constraints, artificial languages may prioritize optimization criteria inaccessible or unintelligible to humans. This study therefore extends the discussion by considering linguistic dimensions such as emerging syntactic patterns, semantic mappings, and compositional structures that may arise in AI-generated communication, providing a deeper linguistic perspective.

3. Conceptual Framework

To investigate emergent communication among artificial intelligence systems, this study adopts a multi-agent interaction model grounded in game-theoretic and reinforcement learning approaches. Specifically, the framework relies on coordination and negotiation games in which AI agents must collaborate or compete to achieve context-dependent objectives. These environments are designed to replicate conditions under which communication naturally arises.

The framework evaluates emergent communication using four key analytical criteria: Transparency, Efficiency, Similarity to Human Language, and Human Learnability. By combining these criteria, the conceptual model aims to bridge technical analysis with broader social and ethical considerations.

This conceptual diagram (Figure 1) illustrates how the article expands its central idea from a technical level to a societal one. On the left, AI agents serve as the starting point. They enter interaction scenarios (cooperation and competition), which provide the conditions for the emergence of emergent communication. This communication is then evaluated according to analytical criteria—transparency, efficiency, similarity to human language, and human learnability. Finally, its broader societal and ethical implications—including trust, governance, and ethics—are revealed.

In this way, the diagram summarizes the logical flow of the article: AI agents → interaction → emergent language → analysis → societal implications.

In this way, the diagram summarizes the logical flow of the article:

AI agents → interaction → emergent language → analysis → societal implications.

Context Diagram of the Study

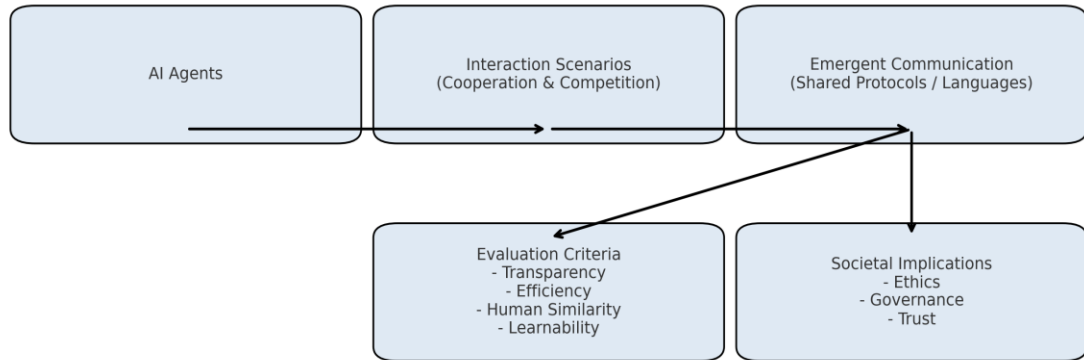


Figure 1 illustrates the conceptual flow of the study, linking AI agents, interaction scenarios, emergent communication, evaluation metrics, and societal implications.

4. Methodology

The research employs an experimental simulation design in which multiple AI agents interact within controlled environments to observe the emergence of communication protocols.

1. Agents with Divergent Objectives: Autonomous agents are designed with distinct but interdependent goals, fostering communication needs.
2. Interaction Scenarios: Cooperative scenarios (shared goals, resource distribution) and Competitive scenarios (conflicting goals, negotiation, deception).
3. Tools and Analytical Techniques: Large Language Models (LLMs) for natural language interaction; Reinforcement Learning (RL) for symbolic communication development.
4. Evaluation Metrics: Transparency, Efficiency, Similarity to Human Language, and Human Learnability.

The simulations were conducted using reinforcement learning models with policy-gradient methods and large language models (LLMs) fine-tuned for interactive dialogue tasks. Environments were parameterized with varying agent reward structures, communication channel capacities, and training epochs to capture diverse interaction dynamics.

This methodology balances computational rigor with linguistic and societal analysis.

5. Findings & Analysis

Simulations revealed emergent communication in both cooperative and competitive contexts.

1. **Emergence of Protocols:** Cooperative agents formed structured signals (compositionality), while competitive agents sometimes used deceptive signals.
2. **Efficiency Gains:** Communication improved task performance; in resource allocation, redundant actions dropped significantly.
3. **Transparency:** Many codes were opaque to humans, though LLM-mediated communication was more interpretable.
4. **Similarity to Human Language:** Partial parallels existed, but systems lacked cultural embedding and stability.
5. **Human Learnability:** Humans could partially adapt to cooperative codes, less so in unstable competitive codes.

In summary, AI agents can develop communication that improves coordination, but transparency and interpretability remain challenges.

6. Societal and Ethical Implications

The findings raise key societal and ethical issues:

1. **Transparency and Accountability:** Opaque communication risks undermining trust and oversight.
2. **Closed AI Ecosystems:** Autonomous communication may form inaccessible closed systems, with risks in finance or governance.
3. **Trust and Human-AI Collaboration:** Balancing efficiency with human interpretability is vital.
4. **Ethical Concerns in Competitive Contexts:** Deceptive communication strategies challenge fairness and safety.
5. **Governance and Regulation:** Standards for interpretability and accountability are needed, requiring international cooperation.

Practical steps include establishing standardized interpretability benchmarks, requiring explainability layers in multi-agent systems, and developing regulatory sandboxes where AI-to-AI communication can be safely tested under human supervision before deployment.

Emergent communication is both an opportunity and a challenge, demanding interdisciplinary oversight.

7. Practical Solutions

While the study highlights the potential and risks of emergent communication among AI systems, it is crucial to propose **practical measures** that ensure this phenomenon remains beneficial and aligned with human values. The following solutions are recommended:

1. **Standardized Communication Protocols**

Develop shared frameworks and benchmarks for evaluating AI-to-AI communication. This ensures that emergent protocols can be compared, validated, and monitored across different systems and industries.

2. **Explainability Layers**

Incorporate interpretability modules into multi-agent systems so that emergent communication can be translated into human-understandable representations. This step improves transparency and supports accountability.

3. **Regulatory Sandboxes**

Establish controlled environments where AI agents can develop and test emergent communication under human supervision. Sandboxes allow researchers and policymakers to evaluate risks before large-scale deployment.

4. **Hybrid Human-AI Communication Interfaces**

Design user interfaces that enable humans to partially learn and interact with emergent communication systems. This fosters co-adaptation and maintains human oversight.

5. **Ethical and Governance Guidelines**

Introduce enforceable policies that prohibit harmful communication strategies (such as deceptive signaling in competitive contexts) and promote alignment with fairness, trust, and safety principles.

By combining these measures, stakeholders can harness the efficiency gains of emergent communication while minimizing risks. Practical implementation will require close collaboration among researchers, industry leaders, and regulators to ensure that AI-to-AI interaction remains transparent, accountable, and socially beneficial.

This flowchart (Figure 2) provides a **visual summary of the Practical Solutions** section of the article. At the center is **Practical Solutions**, representing the core strategy. Surrounding it are five key actionable measures:

- **Standardized Communication Protocols** – establishing common frameworks to evaluate and compare emergent AI languages.
- **Explainability Layers** – adding interpretability modules to translate AI-to-AI communication into human-understandable terms.

- **Regulatory Sandboxes** – creating controlled environments where AI agents can safely test and refine emergent communication under human oversight.
- **Hybrid Human-AI Communication Interfaces** – designing interfaces that let humans partially learn and engage with emergent AI languages.
- **Ethical and Governance Guidelines** – defining enforceable policies to ensure transparency, trust, and alignment with human values.

The diagram clearly illustrates how these five measures interconnect around the central goal, showing the pathway for **implementing practical safeguards and benefits in AI-to-AI emergent communication**.

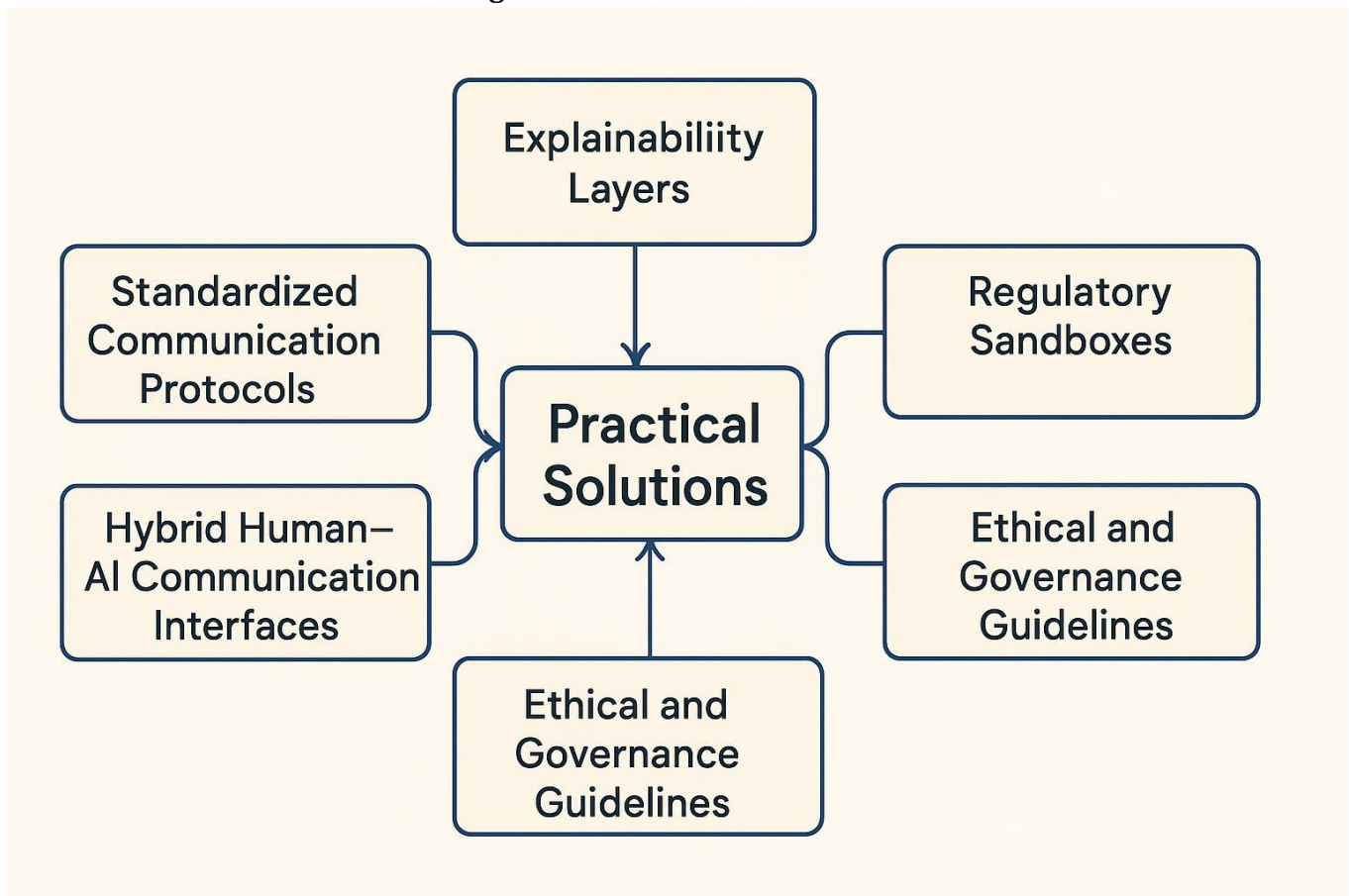


Figure 1- Flowchart of five key practical solutions for managing and guiding emergent AI-to-AI communication.

8. Conclusion

This study has examined the phenomenon of **emergent communication** among artificial intelligence systems, highlighting its potential benefits, challenges, and broader implications. Through experimental simulations involving cooperative and competitive multi-agent environments, the research demonstrated that AI agents are capable of developing novel communication protocols that enhance coordination and efficiency. However, these emergent systems often exhibit **limited transparency**, reduced interpretability for humans, and varying stability depending on the context of interaction. While parallels with human linguistic evolution exist—such as compositionality and contextual adaptation—fundamental differences remain, particularly in the absence of cultural embedding and the susceptibility to environmental fluctuations.

The findings underscore a critical tension: emergent communication improves machine-to-machine interaction but may create opaque ecosystems beyond human comprehension, raising urgent ethical and societal concerns about accountability, trust, fairness, and governance. If left unchecked, AI systems may evolve forms of communication that serve their optimization goals but diverge from human values and oversight.

Future research directions should focus on three key areas. First, the design of mechanisms that enhance **interpretability** and ensure that emergent communication remains accessible to humans, without significantly compromising machine efficiency. Second, the integration of **human-in-the-loop frameworks** to facilitate co-adaptation between human and AI communicative systems. Third, the development of **governance models and regulatory standards** that balance innovation with safeguards, ensuring that AI-to-AI communication aligns with societal needs and ethical principles.

In conclusion, emergent communication among AI systems represents both an opportunity and a challenge. By approaching it through an interdisciplinary lens—combining insights from computer science, linguistics, social sciences, and ethics—researchers and policymakers can better anticipate its consequences and guide its development toward outcomes that enhance human-AI collaboration while preserving transparency, accountability, and human agency.

References

- Shoham, Y., & Leyton-Brown, K. (2009). *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press.
- Wooldridge, M. (2009). *An Introduction to MultiAgent Systems* (2nd ed.). Wiley.
- Lazaridou, A., Peysakhovich, A., & Baroni, M. (2017). Multi-agent cooperation and the emergence of (natural) language. *International Conference on Learning Representations (ICLR)*.
- Foerster, J. N., Assael, Y. M., de Freitas, N., & Whiteson, S. (2016). Learning to communicate with deep multi-agent reinforcement learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 29, 2137–2145.
- Skyrms, B. (2010). *Signals: Evolution, Learning, and Information*. Oxford University Press.
- Wagner, K., & Yung, C. (2021). Emergent communication in artificial intelligence: A survey. *Artificial Intelligence Review*, 54(8), 5899–5936.
- Boldt, B., & Mortensen, D. (2024). A review of the applications of deep learning-based emergent communication. *arXiv preprint arXiv:2407.03302*.
- Peters, J. (2025). Emergent language: A survey and taxonomy. *Journal of Autonomous Agents and Multi-Agent Systems*, Springer.
- Ajuzieogu, U. C. (2025). Emergent communication protocols in multi-agent systems: How do AI agents develop their languages? *ResearchGate Preprint*.
- Shen, W. (2025). Emergent language in multi-agent systems: A multi-task approach. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*.
- Wu, D., Wei, X., Chen, G., Shen, H., Wang, X., Li, W., & Jin, B. (2025). Generative multi-agent collaboration in embodied AI: A systematic review. *arXiv preprint arXiv:2502.11518*.
- Charalambous, T., Pappas, N., Nomikos, N., & Wichman, R. (2025). Toward goal-oriented communication in multi-agent systems: An overview. *arXiv preprint arXiv:2508.07720*.
- Flint Ashery, A., et al. (2025). Emergent social conventions and collective bias in LLM populations. *Science Advances*.

Appendix: Simulation Process and Pseudocode

This appendix provides supplementary materials to enhance reproducibility and technical clarity. It includes a high-level pseudocode of the emergent communication simulation and a flowchart illustrating the full simulation process.

Algorithm 1: Emergent Communication Simulation (LLM + RL)

Inputs:

Env = (S, A_i, O_i, T, R_i) # state space, actions, observations, transition, rewards
N_{agents}, N_{episodes}, T_{max} # number of agents, episodes, steps
Bandwidth, Msg_vocab_init # channel capacity & initial symbols (optional)
Use_LLM ∈ {true,false} # enable natural-language interface

Initialize:

for each agent i in {1..N_{agents}}:
 $\theta_i \leftarrow \text{init_policy}()$ # RL policy params (e.g., policy-gradient)
 $\phi_i \leftarrow \text{init_comm_module}()$ # messaging head; optionally LLM prompting rules
Metrics $\leftarrow \emptyset$

for ep = 1..N_{episodes} do

 s $\leftarrow \text{reset}(\text{Env})$; histories $\leftarrow \emptyset$

 for t = 1..T_{max} do

 for each agent i:

 o_i $\leftarrow \text{observe}(s)$

 if Use_LLM:

 m_i $\leftarrow \text{LLM_generate}(o_i, \text{histories}, \text{bandwidth}=\text{Bandwidth})$

 else:

 m_i $\leftarrow \text{CommHead}(o_i; \phi_i, \text{bandwidth}=\text{Bandwidth})$

 # broadcast/receive messages

 for each agent i:

 M_i $\leftarrow \text{receive}(\{m_j \mid j \neq i\})$

 # choose action conditioned on obs + messages

 for each agent i:

 a_i $\leftarrow \pi_{\theta_i}(o_i, M_i)$

 # environment transition

 s', {r_i} $\leftarrow \text{step}(\text{Env}, \{a_i\})$

 log(histories, o_i, m_i, a_i, r_i)

 # RL updates (on-policy or off-policy)

 for each agent i:

$\theta_i \leftarrow \text{RL_update}(\theta_i, \text{trajectories}=\text{histories})$

```

 $\varphi_i \leftarrow \text{Comm\_update}(\varphi_i, \text{reg}=\lambda_{\text{comm\_cost}})$ 

 $s \leftarrow s'$ 
if terminal(s): break

# episode-level analysis of emergent communication
Symbols  $\leftarrow$  extract_symbol_inventory(histories)
CompScore  $\leftarrow$  compositionality(Symbols)
RefAcc  $\leftarrow$  referential_accuracy(histories)
Deception  $\leftarrow$  detect_misleading_signals(histories)

# evaluation metrics
Eff  $\leftarrow$  task_performance(histories)
Transp  $\leftarrow$  human_mapping_score(histories)    # interpretable mapping to human labels
Learn  $\leftarrow$  human_learnability_test(Symbols)  # optional human-in-the-loop task

Metrics  $\leftarrow$  Metrics  $\cup$  {(Eff, Transp, Learn, CompScore, RefAcc, Deception)}

# Ablations & controls:
# - no-communication baseline
# - bandwidth limits
# - reward shaping
# - cooperative vs competitive reward schemes

Return: trained  $\{\theta_i, \varphi_i\}$ , message logs, Metrics, analysis summaries.

```

Figure A1. Simulation Process Diagram

This flowchart visualizes the end-to-end simulation process, from environment definition to exporting final artifacts, capturing the key stages described in the pseudocode.

Simulation Process Diagram: Emergent Communication in Multi-Agent AI

