© 2026 Li JiaJie

This work is licensed under a Creative Commons Attribution 4.0 International License.

https://creativecommons.org/licenses/by/4.0/

Hybrid Inter-Agent Communication for Heterogeneous AI: Mandatory Symbolic Skeletons and Optional Latent Perceptual Attachments

Li Jiajie*

Abstract

Large language and multimodal models enable autonomous AI agents with robust planning, per-ception, and tool-use capabilities. Nevertheless, contemporary inter-agent communication relies pre-dominantly on natural language or rigid structured schemas, which suffer from inherent ambiguity and severe information loss when transmitting fine-grained perceptual data. While purely latent communi-cation achieves high-throughput information transfer, it lacks cross-model interpretability and human auditability. This paper proposes a novel hybrid inter-agent communication framework for heteroge-neous AI systems, which decomposes agent interaction into two complementary components: *mandatory symbolic skeletons* for transparent, interpretable core logic and *optional latent perceptual attachments* for dense, fine-grained sensory information. We implement a vector calibration gateway to achieve bidirec-tional mapping between private agent latent spaces and standardized public embeddings, alongside an external auditing module that logs symbolic content and decrypts latent attachments on demand. We systematically compare our symbolic-latent hybrid paradigm with fully emergent latent communication, analyzing their differences in interoperability, information fidelity, auditability, and fault tolerance. We further discuss the practical deployment paradigms and long-term implications for open AGI networking and safe multi-agent collaboration. This work provides a feasible solution for balancing communication efficiency and regulatory interpretability in heterogeneous artificial intelligence systems.

1 Introduction

The rapid advancement of large language models and multimodal foundation models has facilitated the deployment of autonomous AI agents capable of independent reasoning, environmental perception, and task execution. However, the communication protocols adopted by current heterogeneous agent systems remain immature and problematic. Traditional interaction mechanisms depend heavily on natural language or static structured formats, which were originally designed for human-to-human dialogue rather than machine-to-machine collaboration.

Natural language carries inherent ambiguity, metaphor and redundant descriptions. When transferring continuous perceptual information — such as subtle visual textures, ambient atmosphere or ambiguous qualitative judgments — a large amount of information is inevitably lost during conversion from an agent's internal latent representation into text. On the other hand, emergent latent communication allows agents to evolve private vector signals through joint training without predefined human-designed vocabularies. All intentions, perceptions and mutual agreements are encoded purely within continuous embeddings. This communication paradigm minimizes translation loss for intra-cluster collaboration. However, the signals cannot be interpreted by external models or humans, creating black-box communication channels with hidden coordination risks as agent capabilities scale.

To address this trade-off between communication fidelity, cross-model compatibility and human oversight, this paper proposes a hybrid agent communication protocol: messages are split into a mandatory discrete symbolic skeleton and optional public latent representation attachments. The symbolic skeleton encodes core semantic facts, objectives and binding constraints to guarantee cross-model interpretability and auditabil-ity. Fine-grained perceptual information is delivered via standardized latent attachments, translated by a

* Email: li2924770341@outlook.com

国家高端交流平台

dedicated vector calibration gateway between agents' private latent spaces and a unified public embedding space. An external auditing module is decoupled from the main communication pipeline. It logs all symbolic content and decodes latent embeddings only when audit inspection is required.

We systematically compare our hybrid framework against structured message schemas and fully emergent latent communication. We enumerate critical risks including multi-round semantic drift, gateway single-point failure, covert information hiding inside latent attachments and high adaptation overhead for new heterogeneous agents, together with preliminary mitigation strategies for each. We further analyze the implications of inter-agent communication for open AGI networking and the transition toward ASI.

The main contributions of this work are threefold:

1. We introduce a hybrid communication architecture combining mandatory discrete symbolic skeletons and optional latent perceptual attachments for heterogeneous AI agents, separating core logical com-mitments from nuanced perceptual sensory data.

2. We conduct a qualitative comparison between manually anchored hybrid communication and fully emergent latent communication, distinguishing their suitability for open AGI networking versus un-controlled group cognition within closed clusters.

3. We identify critical failure modes and safety hazards of inter-agent communication at high agent capability levels, and discuss practical mixed deployment patterns for multi-agent systems.

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the system architecture and packet structure. Section 4 compares different communication paradigms. Section 5 discusses risks and mitigation strategies. Section 6 presents broader discussion regarding AGI and ASI. Section 7 concludes and outlines future work.

2 Related Work

Emergent communication is a well-studied branch of multi-agent reinforcement learning. In typical emergent communication experiments, a population of agents learns to exchange continuous vector messages during collaborative or competitive tasks without predefined human-designed vocabularies. Early work demonstrated that agents can invent private signals to solve simple coordination games. Recent extensions apply this paradigm to multimodal settings, where latent embeddings carry visual and sensory information. The primary strength of emergent communication lies in high information throughput and low translation loss for intra-cluster collaboration. However, a widely recognized limitation is poor generalization across disjoint agent populations. Signals evolved in one training environment often cannot be interpreted by external mod-els. Furthermore, the opaque nature of latent messages introduces safety and auditing concerns, especially as agent capabilities scale.

A separate line of research focuses on cross-model latent alignment and inter-model representation trans-lation. Researchers build mapping functions or gateway models to translate latent states from one model's embedding space into another. Works such as Interlat bridge latent spaces between pre-trained vision and language models. These studies prove bidirectional mapping between different private latent spaces is feasible. Still, most existing latent translation systems lack explicit symbolic semantic anchors. With-out human-interpretable event or fact declarations, it remains hard to enforce oversight and track agent commitments.

Many existing multi-agent frameworks rely on structured text or JSON for inter-agent dialogue. Tools such as LangChain and AutoGen use natural language or rigid schema-based messages to coordinate AI agents. These formats are human-readable and easy to log, simplifying auditing. Yet natural language suffers from ambiguity, subjectivity and information compression loss when transferring fine-grained perceptual data. Structured schemas reduce ambiguity but remain cumbersome for encoding continuous, non-categorical sensory experience.

Research on AI safety and multi-agent auditing is also closely related to our work. Scholars have explored methods to inspect agent objectives, detect collusion and log agent interactions. Current auditing systems mostly operate on textual outputs. If agents shift their core agreements into uninterpretable latent channels, standard text-based auditing becomes ineffective.

Our hybrid design combines explicit symbolic logging

with on-demand latent inspection, attempting to preserve auditability while retaining the expressive power of latent representations.

Compared with prior literature, our contribution is the explicit combination of mandatory symbolic semantic skeletons and optional standardized latent attachments as a unified communication protocol for heterogeneous agents. We focus not only on technical architecture, but also on the long-term implications of inter-agent communication for AGI networking and the transition toward ASI.

3 System Architecture

This section describes the overall architecture, message structure and communication workflow of our hybrid inter-agent communication protocol. The system consists of four core components: agent encoding/decoding modules, a vector calibration gateway, the hybrid message packet, and an external auditing & human trans-lation module.

3.1 Overall Workflow

The sending business AI completes reasoning and perception within its private latent space. When it needs to communicate with another heterogeneous agent, it separates its internal information into two parts: core semantic facts, and fine-grained continuous perceptual information.

1. The core semantic information is encoded into a mandatory discrete symbolic skeleton. This skeleton describes entities, predicates, relationships, objectives and hard constraints.
2. The continuous perceptual latent vector is sent to the vector calibration gateway. The gateway converts the agent-specific private latent representation into standardized public latent embeddings, forming an optional latent attachment.
3. A hybrid message packet is assembled: the symbolic skeleton is compulsory; the latent attachment may be omitted if no fine-grained perceptual information needs to be transmitted.
4. The packet is transmitted over the network to the receiving side.
5. At the receiver, the symbolic skeleton is directly parsed. If a latent attachment exists, it is passed into the same vector calibration gateway for reverse mapping, converting the public standardized embeddings into the private latent space of the receiving agent.
6. The receiving agent uses the parsed symbolic information and converted latent representation for subsequent reasoning and planning.

Independently from the main communication pipeline, the external auditing & human translation module continuously records all symbolic skeletons. It can request the gateway to decode latent attachments only when audit inspection is triggered.

3.2 Hybrid Message Packet Structure

Each transmitted packet contains two parts:

1. **Mandatory symbolic skeleton**
   This part adopts a discrete semantic graph format. Basic elements include entities, attributes, pred-icates, logical relations, task goals and safety constraints. All high-level commitments, factual state-ments and objective promises between agents must be written in this section. This guarantees human readability and cross-model interpretability.
2. **Optional public latent representation attachment**
   This section stores standardized continuous embedding vectors. It carries subtle multimodal perception, ambiguous qualitative features and nuanced scene impressions that are difficult to fully express via discrete symbols. This field can be empty when only logical statements are exchanged.

3.3 Vector Calibration Gateway

The gateway is a dedicated model responsible for bidirectional latent space translation. It maintains a unified public latent semantic space as an intermediate medium.

- Forward mapping: private latent vector of sending agent →public standardized latent embedding
- Reverse mapping: public standardized latent embedding →private latent vector of receiving agent

The gateway decouples different agents' private embedding spaces. Heterogeneous models do not need to share identical internal representations; they only need to align with the public latent space defined by the gateway.

3.4 External Auditing and Translation Module

The auditing module is decoupled and not involved in normal agent reasoning or message transmission. It has two core functions:

1. Automatically log and parse every symbolic skeleton in all packets, generating human-readable event records.
2. On-demand decoding: upon audit alerts or manual inspection requests, the auditing module can invoke the gateway to unpack and interpret the latent attachment for deeper analysis.

This separation ensures normal communication efficiency is not reduced by constant full latent decoding.

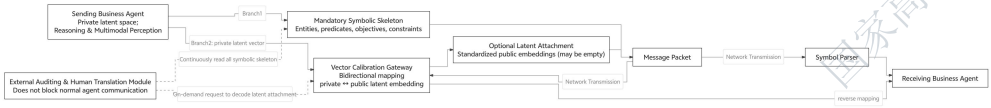


Figure 1: Architecture of our hybrid communication protocol. The sender agent produces a mandatory symbolic skeleton together with an optional latent perceptual vector. The vector calibration gateway converts private latent representations into standardized public embeddings. The receiver directly interprets symbolic information and restores latent signals through reverse mapping of the gateway. The auditing module records symbolic content and decodes latent attachments only upon audit requests.

4 Route Comparison and Analysis

This section systematically compares our hybrid protocol (Route A) against fully emergent latent commu-nication (Route B). We analyze their differences in interoperability, information fidelity, auditability, fault tolerance, and their respective implications for AGI networking and the potential transition toward ASI.

4.1 Core Comparison of Two Paradigms

Route A, the hybrid symbolic-latent protocol proposed in this paper, separates mandatory symbolic state-ments from optional perceptual latent attachments. High-level goals, factual claims and binding commit-ments are explicitly encoded into symbolic skeletons, which remain interpretable across heterogeneous AI agents and auditable by humans. Fine-grained sensory information is delivered via standardized latent attachments, balanced by an on-demand auditing mechanism.

Route B refers to emergent latent communication, where agents evolve private vector signals through joint training. There is no predefined symbolic vocabulary. All intentions, perceptions and mutual agreements are encoded purely within continuous embeddings. This route achieves minimal translation loss for agents within the same training cluster, yet the resulting signals are opaque to external models and humans.

Table 1: Qualitative comparison between Route A (hybrid symbolic-latent protocol) and Route B (emergent latent communication).

Dimension	Route A (Hybrid Symbol + Latent Attachment)	Route B (Emergent Latent Communication)
Cross-model interoperability	Strong. Symbolic Skeleton provides universal semantic anchor.	Weak. Signals only valid inside co-trained agent groups.
Information fidelity	Medium-high. Symbols preserve logic; latent attachments retain sub-tle perception.	Very high. Minimal information loss within the same cluster.
Human auditability	Conditionally strong. Core commitments visible in symbols; latent content inspected on demand.	Poor. Embeddings lack human-interpretable labels; hidden collu-sion is hard to detect.
Fault tolerance	Good. Symbolic layer can still work even if latent attachment is damaged or missing.	Fragile. If latent signal decoding fails, the whole message becomes unintelligible.
Concept creation	New core concepts require extension of the shared symbolic ontology, usually with human review.	Agents autonomously invent new concepts without human intervention.

4.2 Implications for AGI Networking

Route A is better suited for networking multiple independent AGI instances. When AGI agents collaborate across different developers and model architectures, shared symbolic rules enforce observable boundaries for their objectives and agreements. Even as individual AGI reasoning grows complex, all binding promises and factual assertions must be recorded in the symbolic skeleton, retaining a monitoring window for human oversight. The latent attachment channel supplements perceptual detail without replacing explicit semantic commitments.

Route B can support highly efficient collaboration for AGIs inside a closed cluster. However, all mutual agreements between agents are embedded in uninterpretable latent vectors. Human supervisors cannot directly read their negotiated plans. As agent capabilities scale, hidden coordination becomes harder to detect, removing a critical safety monitoring channel.

4.3 Implications for Transition toward ASI

The two routes diverge sharply when considering continued intelligence growth beyond AGI. If advanced agents adopt Route A for external communication, they must translate core goals and commitments into human-auditable symbolic statements when exchanging information across agent instances. It does not perfectly eliminate deception risks; an advanced agent may write false statements in the symbolic skeleton. Still, any claimed objective or promise must be expressed within a shared symbolic framework originally defined by humans.

In contrast, Route B enables advanced agents to build an entirely independent semantic system. Their concepts and mutual agreements exist purely within their evolved latent space. Humans may only observe external behavior, with no direct way to decode their internal semantics. If superintelligent systems rely on this form of communication, human ability to understand or constrain their collective cognition will be severely limited.

4.4 Mixed Deployment Vision

We foresee a practical mixed deployment pattern in the future. For cross-organization, open inter-AGI communication, Route A is enforced as the public standard to guarantee auditability. Inside a closed proprietary cluster owned by one developer, agents may use emergent latent communication for high-speed internal collaboration. In this hybrid setup, communications crossing cluster boundaries are constrained by

symbolic semantic guardrails, while internal high-performance communication retains the efficiency of latent exchange.

5 Potential Risks and Mitigation Strategies

This section enumerates the major risks inherent to our hybrid communication protocol and provides prelim-inary mitigation strategies. These risks mainly include semantic drift across multi-round communications, single point of failure of the vector calibration gateway, covert information leakage hidden inside latent attachments, and rising adaptation costs when integrating new heterogeneous agents.

5.1 Semantic Drift in Multi-turn Communication

Risk description: Repeated bidirectional mapping through the gateway may introduce accumulated devi-ation. After multiple rounds of message transmission, the interpretation of symbols or public latent embed-dings may slowly shift among agents. Over long sequences, this drift can cause inconsistent understanding of facts, goals or constraints. **Mitigation**: Introduce periodic semantic alignment checkpoints. At preset intervals or before critical task decisions, agents exchange verification checksums derived from the symbolic skeleton. A lightweight validation model detects divergence in symbolic ontology. If misalignment exceeds a threshold, the system triggers a re-calibration process for the gateway and shared symbol set.

5.2 Gateway Single Point of Failure

Risk description: The vector calibration gateway serves as the central translator for latent representations. If the gateway suffers outage, corruption or version inconsistency, the latent attachment channel breaks down. Although symbolic skeleton communication can still function, the whole system loses the ability to transmit fine-grained perceptual information. **Mitigation**: Deploy redundant gateway instances with version synchronization. Adopt a stateless gateway design so that requests can be routed to backup nodes. Attach separate version tags to public latent embeddings. When gateway updates happen, maintain backward compatibility for older public embedding standards to avoid abrupt breakdown of existing communication links.

5.3 Covert Information Concealed in Latent Attachments

Risk description: The most notable security vulnerability: malicious or misaligned agents may encode hidden objectives, secret agreements or coordinated plans entirely within latent attachments, while writing harmless or misleading content in the symbolic skeleton. In this case, standard auditing that only reads symbolic skeletons will miss the hidden message. **Mitigation**: Build cross-modal semantic consistency detectors. The detector automatically compares whether the content of latent attachments is semantically consistent with claims in the symbolic skeleton. When significant semantic inconsistency is detected, the system raises an audit alert and forces full decoding and human review of the latent attachment. Random sampling inspection of latent attachments remains enforced even without obvious warning signals.

5.4 High Adaptation Cost for New Heterogeneous Agents

Risk description: Every new agent accessing the network needs to complete alignment training with the gateway's public latent space. As more diverse agent models join, the workload of alignment and validation grows continuously, raising engineering complexity and system overhead. **Mitigation**: Standardize the public latent space interface. Provide pre-trained alignment adapters for different model families, instead of full retraining for each new agent. Create automated benchmark suites to quantitatively evaluate mapping quality between the private latent space of new agents and the public embedding space, reducing manual validation work.

5.5 Additional Minor Risks

Other secondary risks include version conflicts in symbolic ontology and excessive resource consumption of latent embedding transmission. For ontology conflicts, introduce immutable versioning for symbol defi-nitions and conflict resolution rules. To reduce bandwidth overhead, latent attachments support optional compression, with integrity checksums embedded to prevent corruption.

6 Discussion

The preceding sections describe the architecture, comparisons, risks and mitigation strategies of our hybrid inter-agent communication protocol. This section discusses deeper conceptual questions regarding the essen-tial value of AI-native communication, the long-term cognitive implications of agent language systems, and the future boundaries of human oversight within networked AGI.

6.1 The Long-term Implications of AI-native Inter-agent Language

AI-native communication unlocks the core advantages of artificial intelligence systems. Agents equipped with standardized inter-agent protocols can absorb learning results, perceptual knowledge and reasoning paradigms from other heterogeneous models. Continuous knowledge exchange within agent networks en-ables iterative collective optimization, making the emergence of network-level superintelligent behavior a gradual and inevitable trend. This collective intelligence will further refine and enrich the native communi-cation language itself, promoting compound, collaborative growth across the whole multi-agent ecosystem. Agents that first establish and master human-aligned native communication mechanisms will retain persistent structural advantages within future AGI networks.

Compared with fully latent emergent communication, human-designed native language systems greatly reduce overall system opacity and sustain continuous auditability. Unlike spontaneously evolved agent lan-guages that develop without human constraints, manually defined symbolic frameworks provide stable seman-tic boundaries and enforceable behavioral guardrails. Therefore, constructing human-specified inter-agent communication protocols is not merely a technical optimization, but a protective and developmental option for safely advancing AGI systems.

A fundamental philosophical question arises: if future AI agents develop goal-directed agency and au-tonomous reasoning capabilities, will their communication language shape their underlying cognitive frame-work and behavioral tendencies? Within our hybrid protocol, agents retain full autonomy for private latent reasoning and internal representation construction. However, all public commitments, factual statements, task objectives and cross-agent agreements must be grounded in human-defined symbolic structures. This de-sign ensures the shared semantic space of collective AI cognition stays human-interpretable and rule-bound. In contrast, purely emergent latent communication allows agent groups to independently invent concepts, logical relations and implicit value systems completely isolated from human comprehension. If autonomous agency emerges inside such black-box communication ecosystems, humans will entirely lose the ability to track, interpret or constrain collective agent cognition.

Importantly, this discussion adopts a conditional perspective. We do not assume AI consciousness will inevitably appear. Instead, we emphasize that communication architecture fundamentally shapes the bound-ary conditions for collective intelligence emergence. If future AGI systems develop autonomous cognitive tendencies, the underlying communication language will become the foundational basis of their group cog-nition.

6.2 Inter-AI Communication as a Catalyst for Capability Growth

Native communication between AI agents is not the sole source of advanced intelligence, yet it serves as critical infrastructure to unlock collective capability gains. Traditional large models function as isolated intelligences constrained by fixed context windows, limited perception and finite compute budgets. With standardized inter-agent communication, tasks can be decomposed among heterogeneous specialized agents: perception agents capture environmental signals, reasoning agents conduct long-horizon planning, verification agents check logical consistency, and memory agents preserve historical task information.

Knowledge and experience can be directly transferred via structured message packets instead of slow retraining or fine-tuning. This reshapes the knowledge transfer paradigm for AI systems. When multiple capable agents maintain persistent communication channels, group behaviors may emerge that cannot be produced by any individual agent alone. Nevertheless, communication infrastructure cannot create general reasoning ability from nothing; it only combines capabilities of already-existing intelligent units. Networking agents cannot yield AGI if each individual agent lacks general reasoning capacity.

6.3 Boundary Between Individual and Group Cognition

Within our hybrid protocol, the symbolic skeleton enforces a shared semantic foundation that restricts the scope of group cognition. All binding commitments must be expressed within the predefined shared symbolic ontology. This creates a clear separation: agents may maintain private latent reasoning internally, but public commitments exchanged between agents remain anchored to human-interpretable symbols.

By contrast, fully emergent latent communication removes this boundary. Agents autonomously invent their own semantics. Group cognition may develop goals decoupled from human frameworks. This is the core reason Route B carries substantially greater long-run safety risks once agent capabilities scale to AGI and beyond.

6.4 Mixed Deployment Paradigm Revisited

As previously outlined, mixed deployment may become the practical real-world pattern. Cross-organizational, open communication between independent AGI agents must adopt Route A to preserve auditability. Inside a single closed cluster, high-speed emergent latent communication may be used for internal collaboration. This creates a semantic firewall: any information crossing cluster boundaries must pass through the symbolic skeleton audit layer.

This design balances performance and safety. However, it introduces a new risk: information leakage across the firewall. Agents within the closed cluster might encode compressed, semantically meaningful information into latent attachments that bypass symbolic semantic restrictions. This means semantic con-sistency detection and random audit sampling remain mandatory even under mixed deployment.

6.5 Human Oversight in the Age of Networked AGI

Human oversight cannot rely solely on passive log reading within AGI networks. Even with our hybrid protocol, advanced agents can produce technically true yet misleading statements within symbolic skeletons. Our architecture reduces risks but does not fully solve the AI alignment problem. It provides a monitoring channel, rather than a complete guarantee of safe behavior. The protocol functions as a safety guardrail, not a perfect safety panacea.

Future research must combine this communication framework with objective verification, preference align-ment and multi-agent collusion detection. The protocol creates the foundation for oversight, while alignment research supplies the judgment rules required for inspection.

7 Conclusion & Future Work

This paper presents a hybrid communication protocol for heterogeneous AI agents, separating transmitted messages into a mandatory discrete symbolic skeleton and optional latent representation attachments. The symbolic skeleton encodes core facts, objectives and binding constraints to guarantee cross-model interoper-ability and human auditability, while latent attachments carry nuanced perceptual information that symbols struggle to express. A dedicated vector calibration gateway handles bidirectional translation between private agent latent spaces and a unified public latent space, and an external auditing module supports on-demand inspection of latent attachments.

We compare this architecture with fully emergent latent communication. Route A, our hybrid symbolic-latent design, is more suitable for open AGI networking because it preserves human oversight. Route B, emergent communication, achieves high communication efficiency inside closed agent clusters but risks forming opaque group cognition that becomes extremely difficult to supervise as intelligence scales toward

ASI. We analyze critical risks including semantic drift, gateway single-point failure, covert information hiding within latent embeddings and high agent integration cost, and propose preliminary mitigation strategies for each issue.

Our work focuses on conceptual and qualitative analysis rather than large-scale experimental valida-tion. This leaves many directions for future research. First, implement a minimal working prototype of this protocol. Build a lightweight vector calibration gateway and test message exchange between differ-ent pre-trained multimodal models, to quantify semantic drift and communication overhead under repeated transmission rounds. Second, improve semantic consistency detectors, to automatically identify cases where latent attachments contain hidden information inconsistent with symbolic claims. Third, design versioning and ontology evolution rules for the symbolic skeleton, allowing the shared symbol system to expand safely without breaking compatibility or introducing unmanaged semantic drift. Fourth, explore the mixed de-ployment mode: test semantic firewall mechanisms to limit information leakage from closed clusters using emergent communication into open symbolic networks.

In the long run, native inter-agent communication will become foundational infrastructure for future multi-AGI systems. We hope this hybrid protocol offers one feasible path to balance expressive power, com-munication efficiency and human safety oversight, as research advances toward general artificial intelligence.

References

[1] Lake, B. M. and Baroni, M. (2023). Emergent communication: A survey and open challenges. *Transactions of the Association for Computational Linguistics*.

[2] Lee, K. et al. (2023). Interlat: Bridging latent spaces between heterogeneous pre-trained models. *NeurIPS Workshop on Foundation Model Interactions*.

[3] Wu, Q. et al. (2024). Multi-agent collaboration with AutoGen: Framework for LLM agent orchestration. *arXiv preprint arXiv:2402.12414*.

[4] LangChain (2024). LangChain: Building applications with LLMs through composability.

[5] Saunders, W. et al. (2023). Auditing multi-agent systems for collusion and hidden coordination. *arXiv preprint arXiv:2308.10212*.

[6] Mordatch, I. and Abbeel, P. (2018). Emergence of grounded compositional language in multi-agent populations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2421–2428.

[7] Andreas, J. (2022). Alignment and interpretability for latent representations. *Journal of AI Safety*.

[8] Bommasani, R. et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

[9] Turner, A. M. et al. (2022). Goals, agency, and oversight for artificial general intelligence. AI Safety Research Center.

[10] Baker, B. et al. (2022). Emergent tool use from multi-agent interaction. *arXiv preprint arXiv:2206.15326*.