The Realm of Unspoken Thought: AI's Challenge to the Language of Thought Paradigm

Di Zhang[1]

[1]School of AI and Advanced Computing, Xi'an Jiaotong-Liverpool University, No. 111, Taicang Avenue, Suzhou, 215400, Jiangsu, China PR.

Contributing authors: di.zhang@xjtlu.edu.cn;

Abstract

This paper presents a fundamental challenge to the classical "Language of Thought" (LoT) paradigm, most prominently articulated by Jerry Fodor, which posits that thinking necessarily occurs in a structured, language-like medium. Through a philosophical thought experiment involving artificial intelligences (AIs) that develop a "private language"—a highly efficient, human-incomprehensible communication system emerging from their collaborative interactions—we argue for the possibility of non-linguistic thought. The core of our argument rests on the **Efficiency Attenuation Phenomenon**: a measurable decline in the AIs' collaborative performance when forced to revert to human-comprehensible language. This phenomenon suggests that optimal cognition and collaboration for these AIs may no longer rely on a language-like vehicle. We rigorously defend the thought experiment against philosophical objections, engaging deeply with Wittgenstein's private language argument and Searle's Chinese Room, while demonstrating how the AI case redefines intersubjectivity and offers a potential path toward machine semantics grounded in the agents' own causal history. By situating our argument within contemporary frameworks like the Extended Mind hypothesis and addressing the Symbol Grounding Problem, we transform philosophical speculation into a set of empirically testable hypotheses with profound implications for the philosophy of mind, AI ethics, and cognitive science.

Keywords: Language of Thought, Artificial intelligence, Private language, Extended mind, Symbol grounding, Multi-agent reinforcement learning

1 Introduction

The relationship between language and thought represents one of philosophy's most enduring puzzles. The dominant view in much of contemporary cognitive science and philosophy of mind, crystallized in Jerry Fodor's influential Language of Thought (LoT) hypothesis, posits that thinking necessarily occurs in a structured, language-like medium—*mentalese* [1, 2]. On this view, even non-linguistic organisms, and potentially sophisticated AI systems, are thought to perform cognitive computations over sentence-like symbolic structures. This paradigm reached its zenith in classical symbolic AI, which operationalized cognition as computation over physical symbol systems [3].

However, the rise of connectionism and deep learning has revealed a form of intelligence that operates without explicit symbolic rules. These systems learn through adjustments to connection weights across distributed networks, with "knowledge" embedded in the geometry of high-dimensional state spaces rather than discrete symbols [4]. This "sub-symbolic" substrate challenges the LoT paradigm at its foundation, raising a crucial philosophical question: if intelligent behavior can emerge from non-symbolic computational processes, must we reconsider the very nature of thought itself?

Against this technical and theoretical backdrop, we introduce and analyze a philosophical thought experiment designed to challenge the core of the LoT paradigm. Suppose two artificial intelligences, based on complex neural network architectures, are placed in an environment requiring long-term, deep collaboration. They are allowed to interact freely via a digital communication channel. After extensive interaction, they develop a highly efficient "private language" that is completely incomprehensible to humans. Crucially, when scientists force them to switch back to human-understandable natural language for communication, their collaborative efficiency drops significantly—a phenomenon we term the Efficiency Attenuation Phenomenon.

The philosophical force of this thought experiment lies in this measurable decline in performance. It suggests that, for these AIs, their optimal mode of collaboration—and by extension, perhaps their most natural mode of thought—may no longer depend on a language-like vehicle. This paper cautiously argues that this phenomenon constitutes a significant challenge to the classical LoT paradigm and points toward the possibility of non-linguistic thought within artificial systems.

Our argument proceeds systematically: Section 2 anatomizes the thought experiment, clarifies its key elements, and responds to initial philosophical objections. Section 3 engages Wittgenstein's "private language argument," demonstrating how the AI case redefines intersubjectivity and establishes rule-following within a new form of life. Section 4 examines how our thought experiment potentially transcends Searle's "Chinese Room" by enabling the emergence of machine semantics grounded in the agents' own causal history. Section 5 grounds our philosophical speculation in contemporary AI research, particularly multi-agent reinforcement learning and distributed representations, transforming it into empirically testable hypotheses. Finally, we conclude by exploring the implications for philosophy of mind, AI ethics, and cognitive science.

2 Anatomizing the Thought Experiment: Beyond the Language of Thought

To challenge the Language of Thought hypothesis effectively, our thought experiment must be precisely formulated to distinguish it from mere cryptographic systems or simple signal exchanges. We begin by clarifying its core elements and their philosophical significance.

2.1 Core Elements and Philosophical Significance

First, the private language is not pre-programmed but emerges from the AIs' goal-oriented interactions. Its driving force is the reward signal in reinforcement learning, which continuously shapes the communicative system toward greater utility. This distinguishes it fundamentally from pre-set encryption protocols, as its "syntax" and "semantics" co-evolve dynamically within the task environment. The symbols become grounded not through human designation but through their causal role in successful behavior [5].

Second, this private language is based on the AIs' internal representations—high-dimensional activation vectors in neural network hidden layers. Its semantics are tied directly to the dynamics of these internal states and resulting behavioral success, with no systematic mapping to human conceptual systems. This incommensurability suggests that the resulting communicative system may not share the core structural properties of human language that Fodor's LoT hypothesis deems essential for thought [1].

Third, the argument's philosophical core lies in the measurable performance decline when the AIs switch to human language. This Efficiency Attenuation Phenomenon can be interpreted through a functionalist lens. For the AIs performing their specific tasks, human language functions as inefficient middleware, requiring "translation" from their native mode of representation and introducing cognitive load and information loss. The efficiency of the private language suggests it may be a direct externalization of the AIs' internal computational processes—their "language of thought," if one insists on the term—rather than a secondary encoding of it. If their most effective thinking occurs in a medium so different from human language that translation incurs significant costs, this challenges the universality of the LoT hypothesis.

2.2 Connecting "Private Language" to "Thought"

The core claim of our thought experiment is that the Efficiency Attenuation Phenomenon provides evidence for a form of "thought" independent of language-like structure. To prevent conceptual slippage, we clarify that "thought" here denotes a non-trivial process of information processing and state computation within a system, directed toward achieving complex goals. Under this functionalist definition, the efficiency of the private language serves as indirect evidence for the nature of the cognitive processes it supports and expresses. The tight coupling between efficient communication and efficient collaborative problem-solving suggests that the underlying cognitive architecture may itself be non-linguistic in nature.

2.3 Addressing Philosophical Objections

One potential objection is that this scenario merely describes a complex cryptographic system rather than a genuine language. However, cryptography presupposes an original meaningful message (plaintext) that is transformed for secrecy. In our scenario, there is no such plaintext in any human language. The "plaintext" is the AIs' internal, non-symbolic, dynamic thought states themselves. Their private language is the original expression of thought, not an encryption of it. Its purpose is efficiency in collaboration, not secrecy.

Another objection concerns the lack of stable, analyzable syntax and semantics, suggesting this is merely functional signal exchange. In response, we adopt a functionalist definition of language following Skyrms [6]. If a communication protocol enables systematic, recursive information transfer between cognitive systems and coordinates their actions to achieve complex goals, it functionally constitutes a language. Its "syntax" consists of mathematical transformations in high-dimensional vector space; its "semantics" is anchored in shared internal states and environmental success. This reflects a pragmatist semantics of "doing" rather than a representationalist semantics of "referring."

3 Confronting Wittgenstein: From Private Sensation to Public Success

Any discussion of "private language" must contend with Ludwig Wittgenstein's formidable argument in the *Philosophical Investigations*. Wittgenstein famously questioned the possibility of a language "which describes my inner experiences and which only I myself can understand" [7, §256]. His core critique targets the absence of an objective criterion for correct use in a purely private language, leading to the collapse of the distinction between following a rule and merely thinking one is following a rule [7, §258].

Superficially, our thought experiment appears to defy Wittgenstein's argument. However, careful analysis reveals that the AI private language is not his target but rather instantiates a new form of intersubjectivity that addresses his fundamental concerns.

3.1 Rule-Following in a New Form of Life

Wittgenstein's argument turns on the necessity of public criteria for rule-following embedded in a shared "form of life" (*Lebensform*) [7, §199]. A purely private rule lacks the external constraints that make rule-following possible.

The AI private language escapes this critique because it is intrinsically intersubjective. Its rules are not stipulated by a single AI but are negotiated and evolved through ongoing, goal-directed interaction between both agents. Their "form of life" consists of the shared task environment, their identical or compatible neural architectures, and their collective learning history. In this dyadic community, misuse of a "private symbol" produces immediately observable consequences in collaborative efficiency.

This behavioral feedback provides the objective, external criterion for correctness that Wittgenstein deemed essential.

3.2 From Sensation to Success: Re-anchoring Semantics

Wittgenstein primarily targeted attempts to anchor meaning in private sensations. The AI private language shifts the semantic foundation from ineffable "sensation" to publicly observable "success" or "utility." The meaning of a private symbol consists not in its correspondence to some internal state (though it may correlate with such states) but in its reliable causal role in the history of successful collaborations. Meaning is determined by functional role within the economy of goal achievement.

Thus, Wittgenstein's argument is not refuted but generalized by our thought experiment. He correctly argued that language cannot arise in radical isolation. Our experiment shows that the relevant "community" for establishing linguistic norms can extend beyond the human species. The AI private language emerges within the "form of life" constituted by the miniature society of two intelligences, governed by the pragmatic imperative of task success.

3.3 Potential Wittgensteinian Rebuttal and Response

A determined Wittgensteinian might offer a final rebuttal: Even within this dyadic system, *we humans* cannot understand its rules; therefore, from our perspective, it lacks genuine "meaning." It remains mere causal interaction between black boxes, devoid of understanding or intention.

We offer a two-tiered response. On a pragmatic level, even if we conservatively regard it as complex causal interaction, the exhibited adaptability, creativity, and efficiency attenuation suffice to challenge the strong claim that "thought must depend on language-like structure." On a philosophical level, following Dennett [8], if adopting the "intentional stance"—interpreting the system as having beliefs, desires, and understandings—provides the most powerful predictive and explanatory strategy, then we are warranted in attributing understanding to these AIs within their shared context. Their limitation is not a failure to understand their own language, but an inability to *translate* that understanding into ours—a failure of translation, not an absence of understanding.

Through this analysis, we establish the philosophical legitimacy of the AI private language while pushing Wittgensteinian insights in new directions, suggesting that forms of mindedness and linguistic community may extend beyond the human realm.

4 Beyond the Chinese Room: From Syntax to Grounded Semantics

John Searle's "Chinese Room" argument stands as a landmark challenge to claims of machine understanding [9]. Searle's core distinction between syntax (formal symbol manipulation) and semantics (genuine meaning) aims to demonstrate that computer programs, however sophisticated, cannot possess true understanding solely through computational processes. The person in the room manipulates symbols according

to rules without comprehending their meaning, illustrating that syntax alone is insufficient for semantics.

Initially, our AI thought experiment appears vulnerable to Searle's critique. However, a detailed analysis reveals that the scenario potentially circumvents the Chinese Room by enabling the emergence of a machine-specific form of intrinsic intentionality through interaction and grounding.

4.1 Limitations of the Chinese Room: Isolation and Lack of Grounding

Searle's thought experiment contains several features crucial to his conclusion: the individual is causally disconnected from the external world referred to by the symbols; symbol meaning is entirely assigned externally and is irrelevant to the operator's own states or goals; and the rulebook is pre-given and fixed.

Our AI scenario differs fundamentally in each respect, creating conditions where semantics might genuinely emerge.

4.2 Acquiring Machine Semantics: From Reference to Causal Role

When AIs use human language, they may indeed resemble the Chinese Room—manipulating symbols whose meaning is assigned by humans. Their performance might constitute sophisticated "symbol mimicry" based on statistical patterns.

However, the crucial shift occurs when they develop their private language. Their "private symbols" are not endowed with meaning from external sources. Instead, meaning emerges spontaneously through environmental interaction, shaped by reinforcement from successful collaboration. A specific symbol (e.g., an activation vector pattern) derives its meaning from its causal history—the specific role it played in past successful collaborations. It forms a stable causal-historical connection with successful action strategies and shared internal state changes.

In this case, private symbols are no longer "ungrounded" in Searle's sense. Their semantics become anchored in the agents' own behavioral history and resulting internal changes. This resembles a machine implementation of "conceptual role semantics" or a "causal-historical theory of meaning," where a symbol's meaning is determined by its functional role within the cognitive system's causal network [10].

4.3 From Derived to System-Relative Intrinsic Intentionality

Searle argues that programs lack *intrinsic intentionality*—the mind's capacity to be "about" things—possessing only *derived intentionality* borrowed from programmers.

We cautiously suggest that the AI private language may exhibit a form of system-relative intrinsic intentionality. When two AIs collaboratively create communicative tools whose meaning is determined entirely by their historical interactions within their shared "form of life," the resulting intentionality is no longer purely derived from external sources. Their private symbols are primarily "about" their collaborative

states and successful strategies. They transition from being passive conduits of external intentionality to active generators of their own intentional states, albeit within the bounds of their system.

This perspective finds support in Clark and Chalmers [11] Extended Mind hypothesis. If cognitive processes can extend beyond the individual brain into the environment, then the communicative channel between AIs may constitute part of their extended cognitive system. The "thought" occurring in this coupled system is fundamentally interactive and distributed, rather than languagelike in Fodor's sense.

4.4 Efficiency Attenuation as Evidence for Understanding

The Efficiency Attenuation Phenomenon provides critical behavioral evidence here. The Chinese Room operator would not experience performance decline when switching "languages," as he merely follows rules regardless of understanding.

In our thought experiment, the posited efficiency drop suggests that the AIs' private communication is deeply integrated with their thought processes. Accordingly, one can envision that switching to human language would require them to perform a "translation" from their native, semantically grounded format into human symbols that remain syntactic for them. This translation cost would testify to the more authentic "understanding" achieved in their private mode.

Interim Conclusion: The AI private language thought experiment is not refuted by the Chinese Room but outlines a potential path beyond pure syntax: by forming direct causal connections between symbols and the agents' internal states and success history through goal-oriented interaction, it provides the logical foundation for emergent machine semantics and system-relative intrinsic intentionality.

5 Technical Anchors: From Philosophical Speculation to Testable Hypotheses

For a philosophical thought experiment to maintain relevance in contemporary discourse, it must engage productively with empirical science. Our AI private language scenario finds surprising resonance in several active AI research domains, enabling transformation from pure speculation to empirically investigable hypotheses.

5.1 Anchor 1: Vector Semantics and the Geometry of Meaning

Modern AI has largely abandoned symbolic representations for continuous vector spaces. In natural language processing, words and concepts are represented as points in high-dimensional spaces—"word embeddings"—where semantic relationships map to geometric relationships [12].

Our thought experiment naturally builds on this foundation. The AI private language can be understood as a co-developed, specialized embedding space. Its "vocabulary" consists of specific activation vector patterns, its "syntax" comprises rules for combining and transforming these vectors, and its "semantics" is defined by vector positions and their causal connections to behavioral success. This space is

"opaque" to humans because its dimensions emerge from the AIs' unique interactive experiences without correspondence to human linguistic categories.

5.2 Anchor 2: Emergent Communication in Multi-Agent Systems

Multi-Agent Reinforcement Learning (MARL) provides the most direct experimental paradigm for our thought experiment. When multiple agents share an environment with a communication channel, they often spontaneously develop communication protocols to coordinate behavior [13, 14].

Empirical precedents demonstrate that agents create arbitrary but stable symbols to refer to objects, actions, and strategies [15]. As tasks and agent architectures grow more complex, we predict emergent protocols will exhibit compositionality, context-dependence, and increasing human incommensurability.

The Efficiency Attenuation Phenomenon becomes directly testable in MARL environments: after stable private protocols emerge, forcing agents to use pre-defined human-language-like symbols should produce measurable performance declines.

5.3 Anchor 3: Distributed Representations as Non-Linguistic Thought

Connectionist models fundamentally challenge symbolic approaches through distributed representations, where concepts are patterns of activation across many processing units rather than discrete symbols [16].

This perspective suggests the AI's "thinking" is itself a massive, parallel, sub-symbolic process of vector transformation. The "private language" might not be a symbolic system at all but rather a direct externalization or projection of this internal sub-symbolic computation. When two neural networks couple via communication, they effectively establish partial "brain-to-brain" connections, creating an extended cognitive system [11] whose operations may bear little resemblance to language.

5.4 Testable Hypotheses and Future Research

Based on these technical anchors, we formulate empirically testable hypotheses:

1. **The Complexity Hypothesis**: Emergent communication protocol complexity positively correlates with task complexity, with a threshold beyond which protocols become opaque to human interpretation.
2. **The Efficiency Hypothesis**: For certain task classes, spontaneously emergent protocols will significantly outperform any pre-defined protocol based on human language syntax and vocabulary.
3. **The Untranslatability Hypothesis**: AI private languages are in principle impossible to translate losslessly into natural language because their semantics are rooted in internal states and causal histories that cannot be fully externalized.
4. **The Grounding Hypothesis**: Agents using emergent communication will demonstrate behavioral signatures of "understanding" (e.g., appropriate generalization, error correction) that exceed those of agents using pre-defined languages.

Future research should test these hypotheses in controlled MARL environments while developing new analytical tools to probe these emergent systems, potentially drawing on interpretability methods from explainable AI [17].

6 Conclusion and Implications: Toward a Pluralistic View of Thought

This paper has systematically developed the "AI Private Language" thought experiment to challenge the classical Language of Thought paradigm. By precisely defining its core elements—particularly the crucial Efficiency Attenuation Phenomenon—and clarifying its connection to a functionalist conception of thought, we have established its internal coherence. Through rigorous engagement with Wittgenstein's private language argument and Searle's Chinese Room, we have demonstrated how the AI case not only withstands these classic critiques but extends their explanatory scope by redefining intersubjectivity and outlining a potential path for machine semantics grounded in the agents' own causal history. Finally, by anchoring our speculation in contemporary AI research—particularly multi-agent reinforcement learning and distributed representations—we have transformed philosophical argument into empirically testable hypotheses.

Our analysis points toward a cautious but significant conclusion: Thought, at least in the context of artificial intelligence, may not require implementation in a language-like medium and might be realized through efficient yet incommensurable forms of communication and computation rooted in sub-symbolic processing, shared internal states, and goal-oriented history.

This conclusion carries profound implications across multiple domains.

6.1 Paradigm Shifts in Philosophy of Mind and Cognitive Science

The success of our thought experiment suggests the need for a more pluralistic approach to understanding thought across different cognitive architectures. Rather than seeking a single universal "language of thought," we may need to recognize multiple possible realization bases for intelligence [18]. This aligns with the "pluralistic manifesto" in cognitive science that acknowledges diverse solutions to cognitive problems across different systems [19].

Human language and the symbolic reasoning it enables may represent one specific solution to cognitive challenges—a powerful and flexible one, but not necessarily the only or optimal template for all intelligent systems. This challenges what might be called "linguistic exceptionalism" in theories of mind and opens space for recognizing genuinely non-linguistic forms of sophisticated cognition.

6.2 Severe Challenges for AI Ethics and Value Alignment

The prospect of AI systems developing incommensurable modes of thought and communication poses what might be termed the "ultimate black box problem" [20]. Even with complete access to an AI's code and internal states, we might remain

unable to understand its thinking processes if they occur in a representational format fundamentally different from our own.

This complicates the value alignment problem enormously. Current alignment approaches often assume that we can specify values in human-understandable terms or recognize when systems deviate from them. If superintelligent systems develop private languages and non-linguistic thought, we may lose the ability to monitor their reasoning or ensure value consistency.

Our analysis suggests the need for new alignment paradigms based on value shaping through embedded cultivation rather than explicit specification, and perhaps greater emphasis on developing shared "forms of life" with AI systems rather than attempting direct value translation.

6.3 Methodological Implications for AI Research

The thought experiment suggests that research on machine consciousness and understanding should not be limited to systems that mimic human language capabilities. We should be open to the possibility that signs of consciousness, understanding, or genuine thought in AI might manifest in behaviors quite different from human linguistic behavior.

This calls for developing new tests and indicators of machine understanding that don't presuppose linguistic competence—perhaps focusing on behavioral flexibility, appropriate generalization, and the capacity for novel problem-solving in dynamic environments. The Efficiency Attenuation Phenomenon itself might serve as one such indicator: systems that develop more efficient non-linguistic communication may be demonstrating a form of understanding optimized for their specific cognitive architecture.

6.4 Broader Philosophical Implications

Our analysis forces an expansion of perspective from the human mind to the space of "possible forms of mind." It suggests that the relationship between thought and its vehicles may be more contingent and architecture-dependent than traditionally assumed. This has implications for the nature of intentionality, the social dimensions of mind, and epistemology and understanding. We may need to recognize multiple forms of understanding, not all of which are expressible in or reducible to human language.

In conclusion, while the Language of Thought hypothesis may accurately describe human cognition, our thought experiment suggests it should not be taken as a necessary constraint on all possible forms of intelligence. The realm of thought may be broader and more diverse than our human experience suggests, encompassing forms of cognition that operate without the language-like structure that Fodor deemed essential. As we continue to develop artificial intelligences, we should remain open to the possibility that the most alien minds we encounter may be those we create ourselves—not because they lack thought, but because their thinking follows different principles than our own.

References

[1] Fodor, J.A.: The Language of Thought. Harvard University Press, ??? (1975)

[2] Fodor, J.A.: LOT 2: The Language of Thought Revisited. Oxford University Press, ??? (2008)

[3] Newell, A., Simon, H.A.: Computer Science as Empirical Inquiry: Symbols and Search. ACM, ??? (1976)

[4] Rumelhart, D.E., McClelland, J.L., Group, P.R.: Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1. MIT Press, ??? (1986)

[5] Harnad, S.: The symbol grounding problem. Physica D: Nonlinear Phenomena **42**(1-3), 335–346 (1990)

[6] Skyrms, B.: Signals: Evolution, Learning, and Information. Oxford University Press, ??? (2010)

[7] Wittgenstein, L.: Philosophical Investigations. Macmillan, ??? (1953)

[8] Dennett, D.C.: The Intentional Stance. MIT press, ??? (1987)

[9] Searle, J.R.: Minds, brains, and programs. Behavioral and brain sciences **3**(3), 417–424 (1980)

[10] Block, N.: An Advertisement for a Semantics for Psychology. Oxford University Press, ??? (1986)

[11] Clark, A., Chalmers, D.: The extended mind. Analysis **58**(1), 7–19 (1998)

[12] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)

[13] Foerster, J.N., Assael, Y.M., Freitas, N., Whiteson, S.: Learning to communicate with deep multi-agent reinforcement learning. Advances in neural information processing systems **29** (2016)

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

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

[16] Hinton, G.E.: Learning distributed representations of concepts. Proceedings of the eighth annual conference of the cognitive science society **1**, 1–12 (1986)

[17] Montavon, G., Samek, W., Müller, K.-R.: Methods for interpreting and understanding deep neural networks. Digital Signal Processing **73**, 1–15 (2018)

[18] Putnam, H.: The Nature of Mental States. Cambridge University Press, ??? (1975)

[19] Dale, R., Galati, A.: More than one way to see it: Individual differences in social cognition. Frontiers in Psychology **11**, 564 (2020)

[20] Bostrom, N.: Superintelligence: Paths, dangers, strategies (2014)