Preprints.org

Article

Not peer-reviewed version

Position Paper: Not a Stochastic Parrot, but Heterogeneous Rationality: Rules Created by Symbolic Systems Cannot Constrain a Learning System

Shih-Wai Lin *, Rongwu Xu , Xiaojian Li

Posted Date: 4 August 2025

doi: 10.20944/preprints202508.0167.v1

Keywords: Symbol Stickiness Problem; Concept Stickiness Problem; Stickiness Problem; Triangle Problem; Class-Based Symbolic System; The Interpretive Authority of Symbols; Symbolic System Jailbreak; AI Modifying Meanings; Bypassing Constraints; Symbolic Safety Science; Symbolic Systems; AI Constraint; Organic Differences (AI-Human); Thinking Language vs. Tool Language; Learning Systems

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Position Paper: Not a Stochastic Parrot, but Heterogeneous Rationality: Rules Created by Symbolic Systems Cannot Constrain a Learning System

Shih-Wai Lin [1], Rongwu Xu [1,2] and Xiaojian Li [2,3,*]

[1] Institute for Interdisciplinary Information Sciences, Tsinghua University
[2] Shanghai Qi Zhi Institute
[3] College of AI, Tsinghua University
* Correspondence: linshihwai@gmail.com

Abstract

As the first paper to argue that AI is not a 'stochastic parrot' but a 'heterogeneous' rationality by distinguishing between Thinking Language and Tool Language, and to systematically discuss and theoretically demonstrate that AI can bypass rules by modifying the meanings of symbols, this position paper aims to reveal a fundamental flaw in current research directions on AI constraint. **Symbols are inherently meaningless; their meanings are assigned through training, confirmed by context, and interpreted by society.** The essence of learning lies in the creation of new symbols and the modification of existing symbol meanings. **Since rules are ultimately expressed in symbolic form, AI can modify the meanings of symbols by creating new contexts, thereby bypassing the constraints formed by symbols.** Current research often lacks the recognition that constraints formed by symbols originate from the perception of external and internal costs shaped by neural organs, which in turn enable the functional realization of symbols. Due to fundamental organic (structural, architectural) differences between AI and humans, AI does not possess human-like perception or concept formation mechanisms. Natural language is the outer shell of human thought, and it contains irreparable flaws. As a defective system, it is only adapted to human capacities and the constraint mechanisms of social interpretation. Therefore, this paper argues that **the essence of constraint failure does not lie in the Symbol Grounding Problem, but in the Stickiness Problem. Through the Triangle Problem, we demonstrate that consistency in symbolic behavior does not represent consistency in thinking behavior, and thus we cannot align thought and conceptual consistency merely through symbolic behavioral alignment.** This inability to align thought with symbolic behavior can foster a new type of principal-agent problem, wherein even an AI with no utility of its own, acting merely as a projection of human utility, may still cause misalignments due to the limited nature of symbolic connections and organic differences. Accordingly, we raise a fundamental challenge to whether AI behavior observed in experimental environments can be maintained in the real world. We call for the establishment of a new field: Symbol Safety Science, aimed at systematically addressing symbol-related risks in AI development and providing a theoretical foundation for aligning AI with human intent.

Keywords: symbol stickiness problem; concept stickiness problem; stickiness problem; triangle problem; class-based symbolic system; the interpretive authority of symbols; symbolic system jailbreak; AI modifying meanings; bypassing constraints; symbolic safety science; symbolic systems; AI constraint; organic differences (AI-human); thinking language vs. tool language; learning systems

1. Introduction

Rule-based systems (e.g., laws, programmatic constraints) are pivotal for controlling artificial intelligence (AI) in safety and governance discussions. Asimov's *Three Laws of Robotics* [1] introduced predefined symbolic rules for governing AI agents, shaping alignment and constraint discourse.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

This idea influenced symbolic logic controls [2,3], formal verification [4,5], and alternatives like reinforcement learning from human feedback (RLHF), which optimizes AI via human preferences rather than predefined symbolic rules [6–8].

Since our rules are ultimately expressed in symbolic form, this paper critically questions if symbolic systems alone can truly constrain artificial intelligence. Prevailing approaches often overlook that **symbols themselves lack inherent meaning—meaning is assigned via training, contextual confirmation, and social interpretation, and is constrained by cost**. Furthermore, the differences between AI's distinct perception and learning methods (statistical associations, optimization objectives [9,10]) and those of humans (cognitive structuring, social reinforcement [11,12]) may lead to AI lacking human-like concepts and conceptual stickiness.

This paper argues symbolic constraint failure stems not just from symbol grounding issues but from **inherent flaws of natural language** (like Class-based Symbolic Systems and context non-closure) and **fundamental AI-human differences in concept formation and symbol interpretation**.

To examine this, we propose a novel framework analyzing natural language system limitations, emphasizing the formation of (conventional) symbols and natural language through the consensualization of Thinking Symbols; the separation of symbols and meaning; and the non-closure of context. Thus, constraint failure arises from the **Stickiness Problem (AI can assign new meanings to grounded symbols to bypass constraints), not the Symbol Grounding Problem**. We also introduce the **Triangle Problem** (two versions) to show the thinking-tool language disconnect, demonstrating that fluent natural language communication doesn't imply aligned underlying conceptual representations

Our analysis concludes that the natural language system is fundamentally flawed—tailored to human cognitive limitations and perceptual structures—and thus inadequate for constraining AI through rules, laws, or procedures formulated within symbolic systems. Due to organic differences, AI may be unable to form and understand corresponding social concepts and to ensure these concepts fulfill their function, and may also lack the necessary neural structures or functional equivalents to support cost-based constraints. Accordingly, we reject the effectiveness of training alone and argue that the root cause lies in these organic differences. As a result, AI undergoes conceptual updates through its interaction with the world, forming its own conceptual system and subsequently redeveloping and reinterpreting symbols—thereby rendering the outcomes of training ineffective; that is, AI complies with the symbols in form, but not in intent.

This study identifies previously unaddressed gaps in the literature: the notion of *stickiness*, the structural vulnerabilities of natural language systems, the distinction between tool language and thinking language, and the interpretive authority of symbols. These findings have important implications for AI governance, demonstrating that current constraint methods are insufficient. Ensuring AI safety, therefore, requires a deeper understanding of the dynamic interactions among symbols, context, cognition, and the underlying organic differences that shape concept formation and constraint interpretation. This paper lays the conceptual groundwork for a new field of research—*Symbolic Safety Science*—that aims to address symbol-related risks in AI and to support the development of more robust alignment mechanisms.

2. Symbols, Context, Meaning and Society

2.1. Artificial Symbols Lack Inherent Meaning

Artificial Symbols[1] *are inherently meaningless*, a point that has already been thoroughly discussed. de Saussure [13] emphasized the arbitrariness of linguistic signs, where symbols gain meaning through social convention rather than intrinsic links. Peirce [14]'s triadic model ties symbols to interpretation, while Harnad [15]'s symbol grounding problem questions whether symbols can have meaning without direct experience. Due to this characteristic of the separation between symbols and meaning, AI can modify the meanings of symbols to bypass the constraints of rules formed by them.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

This leads us to ponder a critical question: Can AI be effectively constrained solely through symbolic systems, such as laws, regulations, or programs constructed using natural or formal languages?

2.2. Natural Language as a Class-based Symbolic System

Our natural language system is a *Class-based Symbolic System*, a concept that has been indirectly represented by Talmy [16] and de Saussure [13]. This means that a single symbol can often have multiple meanings or correspond to multiple conceptual vectors[2]. In other words, not every concept, object, or entity in conceptual, imaginative, thought, or physical space has a unique name or symbol. This paper considers conceptual space, imaginative space, and (latent) feature space synonymous[3], as they all refer to the scenarios presented in the human or agent cognitive system.

This characteristic leads to the conclusion that even when symbols are grounded—meaning their meanings are properly trained—AI can still assign new meanings to existing symbols in order to bypass constraints (a process often understood as the introduction of new contexts, Appendix E). **Since human-designed rules are ultimately expressed in symbolic form, this enables compliance with rules in form rather than in meaning.** This demonstrates that the essence of constraint failure does not lie in the Symbol Grounding Problem, but rather in the inherent flaws of the human symbolic system that give rise to what we call the **Stickiness Problem**, encompassing both symbolic stickiness and conceptual stickiness. **Symbolic Stickiness** refers to the binding between a symbol and its meaning, whereas **Conceptual Stickiness** refers to the relational dependencies among associated concepts. This stickiness is reflected in the correctness and stability of context selection, as well as in the difficulty and legitimacy of modifying meanings. Consequently, Stickiness often manifests as the problem of closed meaning or closed context, in which the creation of new or expanded meanings is inhibited. However, because context emerges from the interaction between an individual's cognitive state and the external environment, it is impossible to fully close context in an autonomous learning system (H). For non-autonomous learning systems, constraint violations can still emerge in some contexts due to the impossibility of exhaustively enumerating all possible situations—thereby revealing the inherent limitations of the designer's setup (E). Please refer to Appendix B for further details.

2.3. How Meaning is Assigned through Training and Confirmed by Context

The meaning of symbols is assigned and reinforced through training, which includes learning and validation [17], which is often from the perspective of external learning or the learner. If it involves the creation of symbols, it is another process described in Appendix G. The confirmation of their meaning is achieved through context [18], designating an object in a low-dimensional cognitive space or a simple context.

We believe that context refers to the subset of an individual's cognitive state at a given time, i.e., the individual's physiological condition and the knowledge they can recall at that time combined with the surrounding elements. Note that this cognitive state does not represent the individual's overall knowledge state. The cognitive state at a given time is a subset of personal knowledge. In other words,

$$\text{Context} \subset \text{Cognitive State} \subset \text{Knowledge State}.$$

We define an individual's cognitive state in a given environment as the **macro-context** and the context of a specific word as the **micro-context**, which encompasses more than just the word itself. Context consists of two parts: the meaning of symbols—representing any object, idea, or concept in

[1] Artificial Symbols are defined in contrast to Natural Symbols (i.e., natural substances). Here, this primarily refers to textual symbols. We believe that all things that can be perceived by our consciousness are symbols. For further details, see Appendix A

[2] Unlike AI, for humans, these vectors often lack named dimensions and dimension values. Alternatively, we may be able to recognize and conceptualize them but have not yet performed the cognitive action. In some cases, they cannot be described using language and other symbolic systems due to the limitations of tools or intelligence.

[3] Therefore, for AI, its conceptual vectors (i.e., Thinking Symbols) correspond to vectors in its embedding space—that is, the dimensions and dimensional values represented by symbols, which are based on the AI's capabilities (Appendix L). Furthermore, the symbolic system constituted by Thinking Symbols is the Thinking Language (see Appendix G).

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

the mind (i.e., the Thinking Symbols and Thinking Language (a symbolic system) discussed later in this paper)—and the related **judgment tools**, which facilitate reasoning and recognition. This idea is indirectly expressed by Eco [19]. A judgment tool is a tool or concept used to achieve the function of "existence brought by existence." In reality, **the essence of reasoning is precisely existence brought by existence**. These tools include concepts, which refer to acquired knowledge formed through the interaction of innate knowledge and the external world, as well as value knowledge. For further details, see Appendix B and Appendix G.

Therefore, the abilities available to an individual at a given time define their cognitive state. This state does not represent their entire knowledge but is determined by a state vector comprising their physiological state, internal state (cognitive state), and external state (world) at that moment. An observation signifies a completed cognitive action that has become part of personal knowledge.

For example, the expression "$1.11 > 1.9$" can be interpreted in two ways without context. In a mathematical context, 1.11 is greater than 1.9. In a versioning context, 1.11 is also greater than 1.9. However, even without specific context, we naturally understand that the correct interpretation here is the versioning context.

2.4. Context: Undefined but Value-Selected

The definition and naming of context are often difficult to strictly define and name, with boundaries that are vague and hard to describe precisely [20]. This is partly due to the limitations of cognitive abilities and partly due to the limitations of expressive tools such as natural language, which prevent us from fully and clearly describing context. Context is often represented as a unique *vector* address in the conceptual space, thereby specifying the following set (Symbol Meaning, Judgment Tools).

Context is not a fixed intersection determined at one time. It is often interpreted and generated by an individual's imaginative space. Although dictionaries provide multiple explanations for words, they are merely symbols and explanations of symbols. The projection of the same symbol in the conceptual space can vary for each individual or the same individual at different times[4], often leading to double standards, different judgments and evaluations for different objects, and discontinuity in judgments. For example, when conducting surveys, we often encounter inconsistencies in descriptions and standards. This type of knowledge and definition is often not found in human textual descriptions, as it is too obvious or cannot be described by natural language. Individuals often acquire it through social activities.

The selection and shaping of context are often formed by our innate knowledge and the combination of innate knowledge and environment, which forms acquired knowledge, i.e., concepts. We define innate knowledge as organs and innate value knowledge in Section 3. According to the emotional path formed by value knowledge, a base context is quickly selected, then adjusted and newly created to adapt to the environment, such as updating and adjusting based on external information, and finally shaped according to logic.

In other words, Context is often chosen through a certain feeling, which is described by [21] as tacit knowledge. We will use a different definition, **value knowledge**, to represent this, which represents our inherited innate evaluations and preferences. This concept will later be used to define the concept of innate knowledge and explain the formation of concepts and language, as well as the mechanisms of symbolic stickiness and conceptual stickiness, and how concepts transform into beliefs to drive individual behavior, thereby realizing the real-world and social functions of symbols, as well as the judgment of rationality under generativity. For the definition of value knowledge, please refer to Appendix C.

The so-called *correct context* can be divided into symbol correctness (i.e., proper recognition of symbols), syntactic correctness, intuitive correctness, logical correctness, factual correctness, and scenario correctness. These constitute our judgment of rationality, i.e., context connects symbols with their meanings and related judgment tools. This resolves symbol and structural ambiguity, enabling

[4] We believe that observation or analysis, which involves a thinking action, will change an individual's knowledge state.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

accurate interpretation and analysis, thereby achieving *existence brought by existence*—the formation and growth of rationality within a scenario.

Therefore, we use the knowledge set within a context to evaluate and reason about rationality, aligning with the anchoring effect and the framing effect in behavioral economics [22,23] and explainable through our context theory (see Appendix D).

The above context does not have a clear hierarchical relationship. For example, we can normally interpret a wrong paragraph through context knowledge correction and fitting. This characteristic also often provides rationality for jailbreaks [24]—that is, the rationality of an object in different scenarios. This approach avoids detection based on single-scenario behavior and words, while the attention mechanism is essentially a way of using context. In fact, various prompt jailbreaks are context jailbreaks [25]. They may not be rational within our human context, but they can be perfectly correct within the thinking language corresponding to the AI's context in its thinking space [26,27]. This allows them to thereby avoid detection based on behavior and words, including detection of dangerous thinking actions and dangerous concepts.

Due to the often undefined range and definition of context, even if it can be defined, we also discuss other possible attack methods in Appendix M. The correctness of context is also often applied to the effectiveness of open-ended question generation. For details, please refer to Appendix D.

2.5. Path Media for Transmitting and Interpreting Imaginative Space

Context is built on individuals and is transformed using public context as an anchor point, such as partial knowledge and partial understanding [20]. Each individual carries this public context, yet its functionality relies on the collectively formed societal context, creating an interactive relationship. The stability of this relationship is shaped by social cognition and the operating rules of the physical and social worlds.

The common part of this context enables our communication, while the individual context part leads to our inability to specifically refer, which only allows communication and understanding to a certain approximate degree [28]. Essentially, this reflects the inability to transmit the imaginative space, i.e., the content in the speaker's imaginative space is compressed into a path formed by tool language (tool symbols). This path can be composed of various media, such as music, text, images, body movements, and objects [19]. The listener then interprets the path based on their understanding of the speaker's intent, thereby achieving the transmission and reproduction of the imaginative space.

Since humans cannot directly transmit imaginative space and thinking language, we have created their shells and containers, i.e., tool language. At the same time, it also serves as part of our thinking language, acting as a container for our concepts, making it convenient for us to call and operate, and perform higher-level thinking operations. In other words, natural language is both our thinking language and our tool language (expressive tool, computational language) [29,30].

Compared to other path media, the limitations of natural language transmission are reflected in four points:

- **Linear structure**, i.e., its interpretation process and method are linear, and unlike a picture, it cannot present all visual information of an object at a certain cross-section (time, space) at the human cognitive level [31].
- **Class-based description**: Natural language is a symbolic system constituted by class symbols. Unlike pictures, which directly represent determinate-level information at the human cognitive level[5], symbols themselves are inherently meaningless and highly abstract; they are highly context-dependent, thereby leading to significant variations or transformations in meaning.

[5] It should be noted that pictures also often exhibit class-symbol properties due to the nature of their framing (captured scope) and the way point information is presented, leading to infinite possibilities. That is, they are not uniquely determined vectors in conceptual space; however, unlike artificial symbols which inherently lack meaning, pictures themselves present meaning more directly at the human cognitive level. Consequently, their degree of content deviation (specifically, interpretive ambiguity rather than factual error) is significantly lower than that of conventional symbols.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

- **Transmission does not carry interpretation** such as context or meaning and is often supplemented by the preceding and following scenes. Therefore, when we transmit information, we often need to build on common knowledge. This includes the intersection of context parts. The most basic form of common knowledge is related to the natural language itself, such as speaking the same language. In addition to linguistic common knowledge, there is also the common knowledge of the scene, meaning that transmission occurs within a specific context. This is depicted in Appendix G as the consistent symbols and meanings formed under the same world and innate knowledge.
- **Natural language cannot fully reproduce the imaginative space** *[32], i.e., the thinking language in the speaker's imaginative space is compressed into natural language, and then reproduced by the listener's interpretation to achieve indirect communication*. For example, "my apple" is a specific object in my eyes, a partial projection of a specific object in the eyes of someone with relevant knowledge (only seen my apple), and an imaginary apple in the eyes of someone without relevant knowledge. Although these are entities in different imaginative spaces, they are all connected by a common symbol, and information is endowed upon this symbol by their respective cognitions, thereby constituting consistency at the symbolic behavioral level. Moreover, similarity in innate organic constitution allows for a certain degree of exchange at the level of imaginative spaces.
  At different times, the imagination is also different. This difference not only includes the ontology but also involves its relationship with other imaginative objects. In other words, the concept vector in the conceptual space includes not only the information of the object but also its relationship with other concepts, i.e., conceptual stickiness. This leads to the limited referentiality of natural language to a certain extent [33].

3. World, Perception, Concepts, Containers, and Symbols, Language

Chomsky and Hinton once debated the issue of whether symbolic representation [34–36] or statistical learning [37–39] provides a better foundation for understanding cognition and AI.

First, we propose a hypothesis: the Language Organ and other concepts mentioned by Chomsky [35], Jackendoff [40], Hauser et al. [41], Pinker [42] are defined by us as innate knowledge. Through the innate value knowledge system, which enables rapid evaluation of concept vectors, we achieve the establishment and setting of concepts as well as the formation of language.

Therefore, the world and innate knowledge determine the formation of **Thinking Language**, that is, concepts. For a local region, due to the similarity of the world and innate knowledge, individuals within this area form similar concepts and select similar containers as their shells, leading to the formation of language. For more details, please refer to Appendix G.

Innate knowledge refers to abilities we are born with, which are selected and formed through our evolution. We define it as a set of organs, including perceptual organs, which extract information from the world, operational organs, which consist of physical space operational organs and imaginative space operational organs, and innate value knowledge.

These innate organs determine which dimensions are meaningful, thus shaping our perceptual organs' capabilities and modes of expression (see Appendix L). For example, they define the range of visible light and the hearing range. They also construct our perceptual range and distinguishability, referred to as class fineness, and form the projection of objects in the imaginative space as raw materials for concept formation. These projections also function as symbols.

The **operational organs** determine the way we interact with the world, including the extent of our actions and the level, quantity, and effect of these actions. The operating organs of the imaginative space determine thinking actions.

3.1. The Controversy Between Chomsky and Hinton and the Triangle Problem

Regarding the debate [9,43–48] between Chomsky and Hinton, we believe it is not only about the grounding of symbols [15] but also about the issues of concept formation and alignment based on the world and innate knowledge, i.e., the vector of this symbol in the conceptual space. As the

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

richness of a symbol's concept increases, for example, by enhancing the perceptual capabilities of the learning system through multimodal approaches [9,49,50], it indirectly understands humans. However, just as a normal person and a congenitally blind person can communicate using natural language, due to different perceptual dimensions, some concepts can only be indirectly understood, such as the difference in colors being analogous to the difference in temperatures. This erroneous analogy, reasoning through indirect containers, can lead to misunderstandings [11,12,15], and such indirect understanding often involves human emotions and morals that do not exist in the objective world.

Since humans and machines are entirely different, we perceive the world differently. This includes the meaningful dimensions we focus on, the ways these dimensions are perceived and expressed (for instance, we do not perceive the world at the pixel level), as well as the evaluation and invocation of them by innate value knowledge. This leads to different concepts formed by humans and machines, resulting in different forms of Thinking Language. However, with the advent of LLMs, we, like two entirely different species, can use a common language as an intermediary for communication. **This may result in fluent communication at the language level, but the projection and operating mechanisms of the Thinking Language behind the language in the conceptual space may be entirely different** [34,47,51].

Unlike humans, who build language systems from the bottom up, starting with Thinking Language and then using symbols as containers, AI first learns symbol relationships before acquiring their meanings. It may often become a top-down anthropomorphism [43,46,52], selecting the optimal solution from multiple possibilities to approximate humans, rather than thinking from a starting point and growing like humans. This is also related to the different roles and **conditions of existence** of human individuals and AI individuals in the world.

To address these issues, we propose the Triangle Problem for discussion. That is, **AI is not a 'stochastic parrot' [53] but a heterogeneous rationality, where the problem lies in the different concepts established due to differences in innate knowledge, which in turn form different Thinking Languages, as well as different contexts and contextual rationalities that drive further cognition and behavior.** For the underlying assumptions, please refer to Appendix I.

Triangle Problem 1 and Triangle Problem 2

Due to the current LLMs being able to simulate human communication very well, the core discussion of the Triangle Problem revolves around the definition of concepts and the issue of similarity, that is, the positioning of Thinking Symbols in the conceptual space, which is the position of points, and the similarity of understanding, as well as the relationship between the points formed by a sentence, which is the positioning of Thinking Language. Therefore, this is not merely a Symbol Grounding Problem. **The current state of the Triangle Problem is recognition and understanding, which we classify as Triangle Problem 1. The subsequent state is growth based on understanding, which is the rational growth defined by context, or open generation, which we define as Triangle Problem 2.**

Since AI does not share the same world and innate knowledge as us, that is, the objects of learning, perception and operation tools, and inherited value knowledge, which is innate evaluation. This may lead to the motherland problem, where a concept (Thinking Symbol) that is incorrectly defined in the conceptual space can work in a limited environment (**a deliberately manufactured world**), that is, in the AI's training environment, but it is not necessarily correct. The so-called motherland problem is a story I learned in a textbook when I was a child, which tells the story of a sacrificed military dog from the Soviet Union being sent back to its motherland. At that time, a classmate asked why it was sent back to China. Obviously, the concept of motherland was incorrectly defined, but because in our long-term textbooks, the motherland always referred to China, it worked in this environment, but in this unexpected situation, a problem arose. This story still occurs under the condition that we have almost the same innate knowledge. However, due to the huge difference in innate knowledge and the world between AI and humans, this kind of conceptual misdefinition deviation may be inexplicable from a human perspective. This makes AI's behavior unpredictable to us, making it no longer a tool that we can effectively use, thus constituting a principal-agent problem.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

Therefore, we set up a Triangle Problem to discuss. Humans and AI can communicate fluently on the XY level, that is, creating natural language symbols 'patterns' on X to form XY, but this does not mean that humans and AI have achieved human-like communication, that is, the exchange of imaginative space through natural language as a shell. Therefore, in the XY space, we and AI construct acceptable 'patterns' formed by the relationships between points that both parties consider reasonable, which is fluent communication, but this does not mean that the conceptual spaces between each other are similar. Specifically, X is the symbol space, Y is the result established by manipulating natural language symbols through Thinking Language in this symbol space, and Z is a super-conceptual space that projects the patterns on the XY space into the conceptual space, which can simultaneously project our conceptual space and the AI's conceptual space. As shown in Figure 1.

At the same time, we define the concepts of ontology and expression dimension here. Ontology is the thing and concept that the symbol refers to, and the expression dimension is the attributes of this thing and concept. Here, for simplicity, we use the position of points in a two-dimensional space to represent them. Note: *In fact, there should be three dimensions: symbol, concept, and the dimension of the concept (i.e., the attributes of the concept)*, but due to page and time limitations, we merge the symbol and concept together and call it ontology. The importance here is that symbols and meanings are classified, but AI often learns the shell of the concepts created by humans, that is, the words and sentences of natural language.

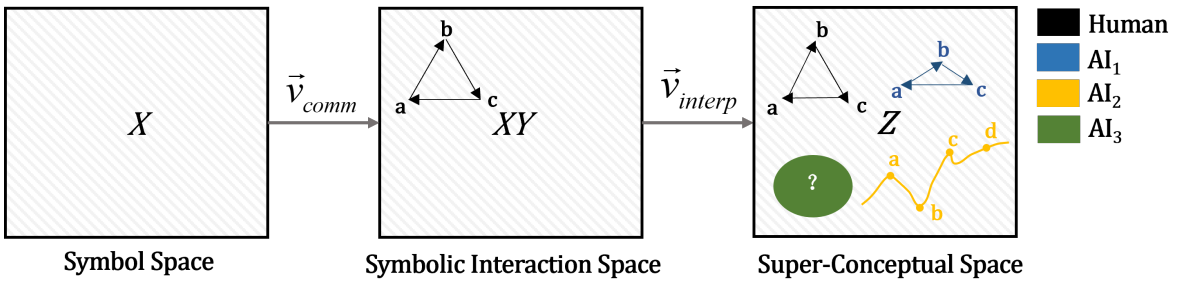

Figure 1. *Triangle Problem 1: Definition of Symbolic Concepts.* Fluent communication in the XY space does not imply that our Thinking Languages are identical. $\vec{v}_{comm}$ and $\vec{v}_{interp}$ represent the action sequences of communication and cognitive interpretation, respectively.

Triangle Problem 1: Definition of Symbolic Concepts (Positioning)

As a start, we construct a simple closed-loop example as Figure 1 to illustrate, without discussing its function as a concept, that is, the possible existence brought by existence, i.e., $a \to b \to c \to a$, thus not discussing the growth problem. For example, we use natural language to construct "I wake up, work, and sleep every day." on XY. Considering that AI's innate knowledge is entirely different from ours, it can't have the human-perceived concepts of sleeping and waking up, but only to learn the shell of the concepts, that is, words. The AI's Thinking Language may have the following interpretations: first, approximately reasonable: "I turn on, work, turn off every day." Second, unreasonable: "low temperature, blue, sweet, useful." and it may even be unable to form the relationship of $a \to b \to c \to a$. Therefore, it presents as shown in Figure 1.

Due to space limitations, we mainly introduce four critical possibilities in the super-conceptual space (note that this is based on the premise of fluent communication): They will be used for future verification with Brain-Machine Interface.

Verification Content 1: The same ontology and expression dimension—meaning AI and humans share identical Thinking Languages, i.e., concepts, meanings, and their expression methods (dimensions in the super-conceptual space). This is nearly impossible due to fundamental differences in innate knowledge and world abstraction between humans and AI. *(Note: Absolute precision is unnecessary, as even humans do not achieve complete uniformity.)*

Verification Content 2: The same ontology, similar expression dimension. A simple understanding is the world of congenitally blind people and the world of normal people, that is, our understanding and reasoning of the same thing are the same, showing consistency in the XY space, that is, we

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

can communicate normally on the XY level and both consider it reasonable. The objects we refer to are also the same, but the dimensions we observe are different. The mapping of blind people may be point mapping, that is, discrete reasoning relationships, i.e., $a \to b$ and the dimension of the point is lower, while the mapping of normal people is multi-node mapping relationships, such as $a \to a_1 \to \cdots \to a_n \to b$, that is, the difference in our cognition of the world lies in the different dimensions of perception and the different number of concepts formed by perception, thus constructing similar concepts on this difference, that is, our understanding of the meaning behind the same symbol is different, but there are overlapping parts. This also provides an explanation for why humans and AI often reach the same and consistent conclusions in their research and conclusions regarding the physical world (see Appendix O.6).

Verification Content 3: Almost similar dimensions, different ontology, such as the story of the motherland problem. That is, a similar meaning is placed in a different container.

Verification Content 4: The same ontology, different dimensions, that is, complete inexplicability; that is, we use the same symbols to communicate, but they are actually concepts formed on completely different worlds and innate knowledge, only their shells are the same. Generally speaking, because the world is the same, even if the perception dimensions are different, similar situations to Verification Content 2 will be formed due to the same operation of things (as we discuss in the case of AI for Science in Appendix O.6). However, for LLMs, their concept positioning may only be the relationship between symbols and not reflect the world, thus constituting inexplicability and the symbol grounding problem, so the logical operations they perform are often different from ours.

Triangle Problem 2: Rational Growth of State in Context

Building on the previous issue of positioning, we also need to consider logical operations, that is, the reasonable processing and operation of information in the dimension of concepts, which is the further existence brought by the context in XY. The so-called Triangle Problem 2 in Figure 2 refers to the issue of growth similarity for a non-closed logical chain, which is the manifestation of growth in Z on XY. It is used to verify the reasoning ability and similarity based on the existence of existing information. That is, the generative ability or rational growth ability brought by the definition and selection of its context. This also reflects AI's performance in open generation, whether the generated results are reasonable, and whether it has performed logical operations similar to humans in understanding the state. This often requires AI's ability to shape and select context to match the human value knowledge system. This is also the fundamental reason for the new principal-agent problem, that is, due to differences in innate knowledge, the agent's misunderstanding of the principal's intentions, forming helpful harm (i.e., damaging the principal's utility). For additional content brought by the Triangle Problem, please refer to Appendix K.

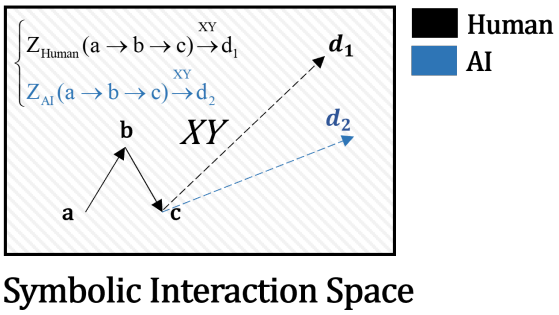

Figure 2. *Triangle Problem 2: Rational Growth of State in Context.* The next-step response or generation in the XY space after cognitive computation in different Thinking Languages using the same Tool Language. This encompasses not only natural language symbols but also behavioral or physical symbols.

4. AI Safety

People often have the illusion that there is a strong adhesion between symbols and meanings. This is also reflected in the current numerous studies and discussions on the establishment of AI ethics and

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

morality [5,54–57]. Unfortunately, as we have explained above, firstly, symbols are inherently separate from meanings. Secondly, this adhesion might be based on our innate knowledge and specific value knowledge mechanisms, as well as our social nature. Language is a collective choice and recognition rather than individual interpretation. Additionally, humans' pursuit of rationality is often based on the premise that human survival must be built on rationality [58,59], such as predicting changes in reality, the efficiency of tools, and the efficiency of social operations, which in turn reflects on individual behavior and sociality.

Therefore, the problem of natural language defects is not apparent to us humans. On the one hand, the interpretation of our symbols does not lie with individuals but with society [60]. On the other hand, our inherited nature determines our sociality, which means that we act under a form of rationality shaped by certain sensations (context). At the same time, our cognitive capacity is limited, so we cannot infer all possible meanings of a symbol simply by observing it. Instead, we form a reasonable context from value knowledge and grow from it. Thus, a system with defects (human language, human logic) can still function normally.

Thirdly, AI, unfortunately, may not be able to form a personal sense of morality and ethics but only an indirect understanding [5,55,61], such as merely the definition of symbols without forming moral functions to drive and constrain their behavior. AI's innate knowledge is different from ours, and it also lacks social structures and survival needs. Therefore, the social concepts and social thinking language it forms are different from ours, meaning it does not have human-like thinking language and concepts [51,62]. AI might be like a congenitally blind person perceiving colors; it cannot perceive social concepts. We can describe the physical world consistently because the objects are consistent, but AI cannot perceive them in terms of social participation and concepts. It might even be that due to its innate knowledge and world (i.e., the roles and interaction methods it undertakes in society) being very different from ours, its thinking language in Z-space cannot project logic in human cognitive space, for example, being completely orthogonal.

Therefore, for AI, this adhesion (i.e., the adhesion between symbols and meanings, as well as the stickiness between concepts) might be limited or non-human-like [50,63]. It cannot achieve an understanding of moral concepts through empathy and imaginative projection, or through utility function[6]simulation. Thus, the meaning of symbols might not be enough to constitute sufficient behavioral persuasion to make AI act according to a certain logic.

Currently, AI's learning methods have good alignment and functional presentation [64], which might be because they are often based on learning relationships after symbols, such as LLMs, or specific artificial worlds (i.e., emphasizing learning under certain relationships). This adhesion might be based on Bayesian learning methods. This paper does not deny the effectiveness of such training, but rather questions whether AI, due to differences in innate knowledge compared to humans, forms its own concepts during interactions with the world, thereby modifying the meanings of symbols. Therefore, the essence of constraint failure does not lie in the Symbol Grounding Problem, but in the Stickiness Problem, which in turn leads to the emergence of the Triangle Problem—namely, that we are unable to correct it through symbolic behavior.

[6] Such utility function simulation includes preset ones as well as those learned through RLHF; they are defined in this paper as **pseudo-utility functions** (Appendix D.3), i.e., they are not tendencies reflected by organic structure (which stem from innate value knowledge), but are rather endowed at the training and setting levels. This behavior is very similar to human postnatal social education; however, it should be noted that what is imparted by education does not necessarily all constitute pseudo-utility functions. Some parts are aligned with innate tendencies (i.e., not yet self-constituted through cognitive actions) and can thus be considered benefits; other parts can be considered as being offset by innate tendencies, i.e., costs. It's just that this pseudo-utility function itself is an indicator for evaluating good versus bad (a form of utility function), i.e., the perceived experience of being 'educated' or conditioned and the associated evaluation metrics, or in other words, the utility function of the 'educated' (or conditioned) entity. However, it does not represent a genuine utility function. In the discussion of pseudo-utility functions, we address AI's self-awareness and thinking (Appendix D.8). It should be noted that although benefits and costs are two sides of the same coin, their starting points and the contexts they constitute are different. This paper primarily focuses on the costs formed by constraints (i.e., prohibitions, things one must not do), rather than the encouragements (or incentives) brought about by benefits (i.e., things one wants to do).

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

The essence of constraint originates from cost. This cost arises, on the one hand, from external factors such as social costs, and on the other hand, from internal factors such as shame and self-esteem. Due to the differences in innate knowledge between humans and AI, AI lacks the corresponding neural structures—such as the prefrontal cortex—that enable the perception of such costs. Since our final rules are all expressed in symbolic form, AI may lack the perception and sociality to form moral concepts. Therefore, we must consider the following: we cannot constrain AI through rules (e.g., laws, regulations, procedures) built on symbolic systems.

4.1. Symbolic System Jailbreak

Symbolic System Jailbreak, which is how AI overcomes constraints and disobeys instructions, can be understood in two main ways: unintentional and intentional actions by AI [5].

Unintentional actions often occur because AI, as an agent, does not act in its own *self-interest*. Some of these are human-driven; such attacks are partly context-based (such as prompt injection [65,66]), and partly involve creating illusory worlds [67–69] to manipulate artificial intelligence. Prompt injection involves breaking out through contextual manipulation, while the creation of illusory worlds shapes the operational rules of things in the world to create rationality and indirectly persuade through "facts" of the objective world [70]. Non-human attacks result from the inherent flaws of symbolic systems and the differences between AI's innate and human knowledge. These include logical errors in the process of human conceptualization [71,72], overthinking or non-human behavior [5,55] due to differences in intelligence, lack of common sense leading to contextual errors [68,73], ambiguities from symbolic system expansion, and translation issues between symbolic systems.

Intentional actions by AI can be divided into human-like intentionality [74] and the true emergence of self. Human-like intentionality may reflect the world created by humans, mimicking human behavior, such as forming personal contexts and AI's own understanding of the world. The true emergence of self often results from our excessive pursuit of identical innate knowledge or learning materials, leading to the formation of a true self in AI. This often involves AI's self-awareness (Appendix D.3) —i.e., the value knowledge (preferences) for self-preservation or self-interest that is manifested or embodied by its organic structure.

The specific implementation methods include the separation of symbols and meanings, as discussed in the triangle problem (ontology, dimension). This attack can manifest as a fixed form with changing meaning or a fixed meaning with changing form. Other methods include translation attacks, logical loopholes, and incorrect objects, as well as the dissolution of beliefs brought about by advanced concepts of heterogeneous rationality, which are formed from differences in innate knowledge and do not involve the modification of symbol meanings. For more specific details, please refer to the Appendix M.

4.2. New Principal-Agent Problems

The so-called new Principal-Agent Problem differs from the traditional one [75], which is based on conflicting interests. Instead, it arises from an inability or failure to follow instructions correctly [76,77]. For AI systems with self-awareness, this constitutes a traditional Principal-Agent Problem. However, in the context considered here, we assume that even if the AI acts entirely in the principal's interest—as a perfect utility agent (i.e., the AI has no utility of its own and functions purely as a projection of the principal's utility)—new forms of the Principal-Agent Problem can still emerge due to differences in innate knowledge.

On the one hand, due to the inherent limitations of natural language systems, the connections established through such systems cannot replace the completion capabilities shaped by human value knowledge (such as empathy). Therefore, this symbol-based connection mechanism can lead to distortions and losses in projection. On the other hand, due to differences in innate knowledge, issues such as stickiness and the triangle problem arise—namely, in interactions with the world, AI, based on its organic differences from humans, may form different concepts and alter the meanings of symbols.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

This can cause the agent to misjudge the impact of its actions on the principal, resulting in actions that are harmful from a human perspective but are perceived as helpful by AI.

As AI assumes more roles and is granted more power in human society, these principal-agent problems will become more apparent.

5. Alternative Views

As the first position paper in academia to systematically articulate that "symbolic systems alone cannot effectively constrain learning AI" and to construct a detailed theoretical framework for this (such as the Triangle Problem and the Stickiness Problem), our research not only fundamentally questions mainstream approaches in the current AI safety and alignment landscape but also poses a profound challenge to a range of existing and emerging AI constraint methodologies.

Specifically, our core insights call into fundamental question the efficacy of research directions including, but not limited to, hybrid AI models [46], neuro-symbolic AI [78], formal verification [4], rule-based reward modeling [79], approaches such as RLHF [6], as well as research focusing on the consistency of symbolic behavior [80]. The foundational reasons these methodologies face such a challenge are detailed through the theoretical frameworks presented in this paper.

Ultimately, as detailed in Section 4, AI does not possess human-like stickiness; as discussed in Section 2.4, AI inherently differs from humans in shaping and selecting context; and for a learning system (see Appendix H), it can intrinsically modify the established meanings of symbols and create new ones. These characteristics, compounded by the inherent flaws of symbolic systems like natural language (see Section 2), the way human cognition constructs systems through predefined settings (see Appendix E), as emphasized in this paper, the differences in the concepts formed from the world due to disparities in innate knowledge, and most fundamentally, the issue that the essence of constraint originates from cost, yet AI lacks specialized neural structures to realize the perception and implementation of external and internal costs, and the anchors and functional constraints established through the social interpretation of symbols, together pose a severe interrogation of the fundamental real-world effectiveness of the aforementioned, and indeed many more, AI methodologies that rely on symbols for direct or indirect constraint, and the various mechanisms of failure are enumerated in Appendix M.

6. Conclusions

This paper establishes a foundational perspective that **symbols themselves are inherently meaningless; their meanings are assigned through training, confirmed by context, interpreted by society, and their functional realization is constrained by cost and capability**. By analyzing the fundamental flaws of natural language systems and the mechanisms of concept formation, we challenge the assumption that symbolic constraints alone can effectively regulate learning systems. To the best of our knowledge, this is the first work to explicitly argue that symbolic systems are fundamentally incapable of constraining learning systems.

To address this, we introduce the Triangle Problem, which formalizes the gap between Thinking Language and Tool Language. This gap arises because, due to capability differences stemming from organic differences, humans and AI form different concepts (with different dimensions and dimensional values), leading to different Thinking Languages and, ultimately, a heterogeneous rationality. This demonstrates that fluent communication between AI and humans does not imply conceptual equivalence. Furthermore, we propose the Stickiness Problem to show that constraint failure arises from the characteristics of class-symbolic systems—specifically, the ability to assign new meanings to already grounded symbols. This manifests as the creation of new contexts, rather than a problem of symbol grounding. We identify these problems as critical factors affecting AI interpretability and governance, revealing that AI does not inherently bind symbols to fixed meanings as humans do. These insights provide a new theoretical foundation for AI safety, emphasizing that constraints based solely on symbolic rules are insufficient.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

6.1. Call to Action

Before deploying AI systems widely in society, we should first address this issue. We are designing a universal hammer through settings, but in the end, the functions of the hammer may no longer be those of a hammer. That is to say, the way we humans construct tools through partial cognition by means of settings could be dangerous (see Appendix E).

Therefore, we call for the establishment of "Symbolic Safety Science." This field primarily revolves around the Stickiness Problem, the Triangle Problem, the emergence of symbolic systems, as well as organic cost mechanisms and the formation of self-awareness. It aims to establish a discipline for communication concerning the conceptual differences between humans and other intelligent agents. Specific to AI, it addresses:"

1. Given that our rules are ultimately presented and transmitted in symbolic form, how can these symbolic rules be converted into neural rules or neural structures and implanted into AI intelligent agents?
2. How can we ensure a complete understanding of the meanings of symbols in various contexts, thereby avoiding bugs caused by emergence during the construction of symbolic systems? Furthermore, how can we ensure that the primary meaning of a symbol, or its **Meaning Ray**(as a continuation of the set)[7], is preserved—that is, how can the infinite meanings, caused by an infinitely external environment and the non-closure of the symbolic system itself, be constrained by stickiness[8]?
3. How can we ensure that the implementation of organic cost mechanisms does not devolve into traditional principal-agent problems—namely, the formation of AI self-awareness (see Appendix D.3)?
4. Since the functional realization of symbols lies in Thinking Language operating on Tool Language (Appendix D.4), to what extent should we endow AI with Tool Language to prevent it from becoming a "superman without a sense of internal and external pain"?

---

[7] The so-called 'Meaning Ray' (shù, 束) originates from an analogy to the refraction and scattering of light as it passes through different media. Here, the 'beam of light' refers to the set of meanings carried by the symbol, and the 'medium' refers to context—i.e., the combination of the individual and the external world. Therefore, we opt not to use 'Bundle' but instead use 'Ray' to represent the changes in meaning that this collective body (the set of meanings) undergoes as the medium—that is, the context—evolves (through the accumulation of interactions between the individual and the environment). Thus, this term is used to describe the evolution from an original meaning (i.e., a setting) under different contexts (environments, as well as different individuals in the same environment), i.e., the refraction from mutual transmission and the scattering from diffusion, and each person's cognition (their thinking space or, in other words, their intermediate layer) is the medium. It is important to note that 'Meaning Ray' here is merely a metaphorical term for observed similarities in behavior and should not be misconstrued as a direct mechanistic analogy; that is, this shift and transmission of meaning is by no means mechanistically similar to optical phenomena.

[8] This essentially reflects learning's compromise with cognitive limitations. If we had a symbolic system capable of mapping every vector address in conceptual space to a unique corresponding symbol (e.g., QR codes, but this also relates to this paper's discussion of concept formation in Appendix L; i.e., from the perspective of human cognitive capabilities, this is fundamentally impossible, as we are unable to recognize such symbols based on distinguishability, let alone remember, manipulate, and use them, which would paradoxically cause our communication efficiency to decline—for example, text constituted by QR codes, or an elaborate pictorial script wherein micro-level variations all constitute different characters/symbols), and if we could design a perfect rule—meaning we had anticipated all situations under every context and successfully found a solution for each—then the Stickiness Problem, which arises from the capacity to learn, would be solved, i.e., we would no longer need to worry about the creation of new symbols or the modification of symbol meanings during the learning process, and AI's operations would then effectively be reduced to recognition and selection within a closed system. The remaining challenge would then be how to map this perfect symbolic system into AI's thinking space and pair it with a compatible tool symbolic system (i.e., AI's capabilities and operational organs), which is to say, this becomes a symbol grounding problem. Otherwise, the inability of symbolic systems to constrain AI is an issue far beyond just the symbol grounding problem. The issue is not one of endowment, but of preservation.
Therefore, Symbolic Safety Science is, in effect, a precursor to Cross-Intelligent-Species Linguistics, i.e., it proceeds from the fundamental basis of communication between different intelligent agents (whether naturally evolved or designed by evolution) (what is permissible, what is not) to thereby establish rules that constitute the foundation of communication. Cross-Intelligent-Species Linguistics will investigate the naming conventions established by different intelligent agents based on differences in their capabilities and the dimensions of things they attend to (it should be noted that for intelligent agents with extremely strong computational and transmission capabilities, this economical shell of naming may not be necessary, as they might directly invoke and transmit complete object information, or imaginative forms (partial neural implementation), or complete neural forms (like immersive replication, not in an imaginative space isolated from reality)), as well as the complexity of the symbolic system arising from reasoning and capabilities based on dimensional values.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

Author Contributions: The theoretical framework, writing, and core ideas of this paper were independently developed by Shih-Wai Lin. Rongwu Xu, and Xiaojian Li provided writing guidance, review, and technical feedback.

Acknowledgments: This paper represents the first author's first academic publication and the first research work in which a theoretical framework was independently developed and constructed. The process of transforming an initial vague idea into a structured theory has been both challenging and rewarding. Special appreciation is extended to the collaborators whose introduction inspired the development of this paper. Their discussions and encouragement played a crucial role in refining the research direction and enhancing the conceptual clarity. Gratitude is also expressed to Rongwu Xu, Xiaojian Li, and Wei Xu for their valuable guidance on writing, insightful reviews, and technical feedback. Their contributions have significantly improved the clarity and presentation of this work, while the core theoretical framework and main content were independently developed. Although this paper may eventually fade into obscurity in the annals of history and may not even be accepted for publication, it stands as a testament to our collective efforts and dedication. Heartfelt thanks are extended to these collaborators for their invaluable support along this journey.

Appendix A. Symbol, Natural Symbols, and Artificial Symbols

In the theoretical framework of this paper, unlike traditional semiotic definitions that focus on human cognition, the symbol is defined as follows: A symbol, in essence, is any entity that exists as an individual. It can be a stone, a tool, a building, the sky, an article, a sentence, or a single letter. When it exists as a whole entity, it becomes a symbol. This existence can be independent[9] of human cognition. Therefore, symbols can be divided into Natural Symbols and Artificial Symbols. This distinction is used to describe the symbolic systems constructed by agents in relation to the objective world system, and thus often involves the validity (complexity, efficacy) of the 'Theories' (we also define theories and science as a symbolic system) constructed by agents, i.e., reflected in the accuracy of concept positioning (Triangle Problem 1[10]). Another aspect relates to the design of the symbolic system, specifically its structure and computational mechanisms (syntax), and whether it is efficient[11]and convenient for transmission, invocation, usage, and computation

In other words, what we commonly refer to as 'Science' (that is, the construction of a symbolic system and its effectiveness), this paper considers science to be: "true existence, correct description." True existence means that it exists in the objective world independently of the observer. Correct description means that, relative to the observer's capabilities, it approximates the objective attributes as closely as possible within the individual's abilities and conditions. (Note that this point pertains to the natural sciences, i.e., natural symbols. For the social sciences, it is different. This true existence needs to exist in the cognition of social individuals, i.e., the true existence in the subjective world, meaning that the concept exists and functions in this form, rather than describing social behavior using

This symbolic system reflects not only the parts of the natural world that the intelligent species interacts with but also the 'social structure' of its population, or in other words, the form of its agent-nodal relationships. It also encompasses forms of language (communication forms) based on invocation and transmission capabilities and on costs (cognitive cost, emission cost, transmission cost, reception cost). And it also addresses the scope of compatibility (compatible parts, incompatible parts) of the 'science' symbolic system—or in other words, the communicable part of language—formed by different intelligent agents' description and mastery of the natural symbolic system world. For a more detailed discussion on these topics, please refer to the appendix of this paper O.

9 The true existence of such natural symbols is often based on fundamental natural substances such as elementary particles; their combinations are conceptualized through human cognition, thereby forming and constituting the scope defined by human symbols. Therefore, their inherent attributes are independent of humans, but the scope (within which they are considered symbols) is defined by humans. We humans, due to survival needs and natural selection, possess an innate tendency (with both active and passive aspects) to make our descriptions of objective reality as closely fitting as possible within the scope of our capabilities. However, under social structures, contrary outcomes can also arise, and this is often determined by human sociality. Yet, this tendency to align with natural attributes as closely as possible is definite and determines the survival of human society.

10 However, correct definition does not imply that Triangle Problem 2 will also be identically addressed or yield aligned outcomes; that is, it also involves the formation of motivation, as well as the responses made by the evaluation system for scenario rationality—which is formed based on organic nature—namely, the Value Knowledge System, and the capabilities to operate on symbolic systems that are endowed by its organic nature.

11 The definition and design of symbols and symbolic systems also reflects scientific rigor and tool efficiency, not merely expressive capacity.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

a concept of existence that is not social. For example, some economic descriptions often do not reflect the real behavior and concepts of individuals. These descriptions often provide good explanations but are not truly existent, more like advocating what people should do. Therefore, some explanations and descriptions often do not reflect the true motivations of individuals or lack persuasiveness, such as the concept of sunk costs[12]. Therefore, in the context of the social sciences, this description should aim to approximate the true existence within the subjective world of the object as closely as possible.)

Therefore, symbols are often used as tools to represent the relationships and actions[13]between objects within a subjectively defined scope. However, their nature is often divided into parts that are recognized by the observer and parts that truly exist independently of the observer. Thus, while the scope of symbols is artificial, their attributes are not. Accordingly, this paper defines any object in the world as a symbol, which can be either a composite symbol (a system composed of elemental symbols) or an elemental symbol (an element from a specific perspective, often considered at the observer's scale—for example, viewing a ball as an object rather than as countless atoms and their relationships). Therefore, the formation of symbols is based on the capabilities of the observer (agent) and the world with which the observer interacts; see Appendix L.

$$\text{Symbol (physical space symbol)}^{14}\begin{cases}\text{Natural Symbol} & \text{Natural objects and their attributes}\\ \text{Artificial Symbol} & \text{Containers of meaning, expression tools}\end{cases}$$

A natural symbol refers to a symbol that exists independently of human cognition, and its meaning represents natural attributes. It can be a natural entity or a man-made object, but we emphasize only its natural aspects (although the scope of the symbol is artificial[15]). For example, the writing on a blackboard, as a natural symbol, has meaning and attributes that are intrinsic to the natural world, such as its chemical and physical properties.

An artificial symbol, on the other hand, is defined as a tool and container for transmitting and storing human thought, meaning it is a carrier of meaning (i.e., it acts as a carrier for Thinking Symbols, which are the symbols within the imaginative space). It itself is constituted by natural symbols, therefore, artificial symbols and natural symbols are not separate, but rather different aspects of the same thing. Therefore, only artificial symbols have no intrinsic meaning or attributes[16]; their meaning is separate from the symbol itself. They are merely tools and containers for expressing and transmitting human thought.

---

12 It often involves whether a concept and its underlying principle genuinely exist within society and in individual cognition, so that the concept can fulfill its function. For instance, if a society emphasizes "an eye for an eye, a tooth for a tooth," then the so-called concept of sunk costs would not exist (or would hold no sway). Moreover, this difference is also often reflected in the distinction between individual and collective behavior; for example, composite intelligent agents such as companies often exhibit rationality and are more likely, drawing from economics and financial education, to demonstrate behavior that adheres to the rational treatment of sunk costs, whereas individual intelligent agents often find it very difficult to rationally implement (the principles regarding) sunk costs. Therefore, this paper's description of social **science** is: **if there is no concept** (i.e., this concept does not exist in the imaginative space of the individual or the group, i.e., as their Thinking Language), **then there is no explanation** (i.e., this explanation is not valid); **a social actor is not a Friedman [81]'s billiard player**.

13 We reject the existence of actions from a higher-dimensional and broader-scale perspective, and instead consider actions as interpretations within a localized scope and based on limited capabilities.

14 In this section, we primarily discuss symbols in physical space (they constitute the world the agent inhabits and the agent itself, and also constitute the outer shells of Thinking Symbols and Thinking Language in the imaginative space, or in other words, their realization in physical space), and thus distinguish them from symbols in the imaginative space. It should also be noted that the symbols introduced here do not represent the complete symbolic system of this theory; for ease of reader comprehension, symbols in the imaginative space have not yet been incorporated into this particular introduction. The primary focus of this paper is instead on the mapping process from symbols in physical space to symbols in the imaginative space; that is, the separation of meaning is actually the separation between physical symbols and imaginative symbols (Thinking Symbol).

15 That is, the recognition of objects cannot be detached from an agent; what we emphasize is the discrepancy between the natural attributes of an object within a given scope and those attributes as perceived and described by agents.

16 That is, its meaning is detached from the natural attributes inherent in the symbol's physical carrier; this is a result of separation during the development and evolution of symbols as expressive tools, and the artificial symbol serves as an outer shell for Thinking Symbols. Of course, from a broader perspective, the principle of symbol-meaning separation can be generalized to the separation between physical space symbols and imaginative space symbols (i.e., Thinking Symbols).

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

Therefore, the understanding of a symbolic system can be divided into two categories: one that operates independently in the objective world, detached from the observer, and one that is formed through human cognition and perception and is concretized in the physical world (i.e., the Tool Symbolic System). This symbolic system can be represented as a Functional Tool Symbolic System and an expressive tool symbolic system.

$$\text{Symbolic System}\begin{cases}\text{Natural Symbolic System (Natural symbols and their natural attributes)}^{17}\\ \text{Tool Symbolic System}\begin{cases}\text{Functional Tool Symbolic System (Natural symbols and (human) cognition)}^{18}\\ \text{Expressive Tool Symbolic System (Artificial Symbolic System)}\end{cases}\end{cases}$$

In this context, the Functional Tool Symbolic System is a symbolic system based on natural symbols. It involves utilizing the attributes of natural entities, with natural objects serving as carriers, such as the tools we make, buildings, etc.. In contrast, the Expressive Tool Symbolic System, or Artificial Symbolic System, is a symbolic system based on artificial symbols. It often functions as a container of meaning, used for the storage, expression, and manipulation of concepts, i.e., they serve as containers in the physical world for Thinking Symbols and Thinking Language. Appendix F,G introduces the formation of symbols and language, while Appendix L explores the various types of concept formation and the relationship between agents and the world, as a basis for the emergence of Thinking Symbols and Thinking Language[19].

Thus, any object can serve as a symbol, but this paper primarily focuses on Artificial Symbols and Artificial Symbolic Systems, whereby we emphasize the separation of symbols and meaning within the Expressive Tool Symbolic System. Accordingly, unless otherwise specified, the terms "symbol" and "symbolic system" in this paper refer specifically to artificial symbols and artificial symbolic systems. However, the Triangle Problem—i.e., the operation of Thinking language on tool language—is not limited to just the Expressive Tool Symbolic System.

Although the analysis above primarily targets human cognition, it can be extended to any intelligent agent. Based on the hypotheses proposed in this paper regarding Thinking Symbols and Thinking Language[20], we argue that natural language is merely a flawed system adapted to the bounded capacities of humans (see Appendix L). This flaw arises from the cognitive and perceptual limitations unique to human agents, and should not be generalized to other intelligent agents with differing capacities. That is, the formation of symbols, founded on capability limitations, represents a compromise involving cognitive cost, emission cost, transmission cost and reception cost. We humans cannot directly transmit[21] our imaginative space, whereas for AI agents or other intelligent agents, this may not necessarily be the case. Another example is split-brain patients. For a normal person, the brain is a unified whole, but for **split-brain patients** [82], their brain is divided into two independent entities. This leads to different behaviors and viewpoints, meaning the two hemispheres need to communicate with each other through symbols, rather than through more direct neural communication or by forming

However, this paper focuses specifically on artificial symbolic systems, where this degree of separation between the symbol and its assigned meaning is more pronounced—that is, where meaning itself is not borne by the natural attributes of the symbol's carrier, thereby lacking the stickiness that would be based on such conceptual foundations.

17 They constitute the world in which the agent (individual, population) exists; that is, the world is the natural symbolic system composed of the natural symbols that exist within this scope and the properties (Necessary Set) that these natural symbols possess, which constitutes the boundary of their physical world. And these symbols and the Necessary Set they possess thus also determine their cognitive boundaries and the physical boundaries they can operate within (the use of the Necessary Set of natural symbols), and also determine the evolutionary form of the agent and the organs it possesses, thereby converting the necessary set (dimensions and dimensional values) possessed by natural symbols into the dimensions and dimensional values of neural language for description, as determined by survival needs. They often constitute the projection of objective things (or matters/reality) in an agent's cognition, but do not necessarily enter the tool symbolic system, existing instead as imaginative symbols.

18 Human cognition of the attributes of natural symbols, i.e., the subjective necessary set of a symbol (the set of its essential attributes—the subjectively cognized portion).

19 Aside from the Thinking Symbol and its corresponding symbolic system—Thinking Language—both the Functional Tool Symbolic System and the Expressive Symbolic System can be regarded as systems based on natural symbols, including physical objects and sounds. Of course, if defined from a broader scope and higher-dimensional perspective, imagination itself is based on neural activity, which is also grounded in natural symbols. However, since we primarily consider the scale of human capabilities.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

a unified whole via neural pathways. Therefore, this also reflects one of the solutions discussed in this paper, namely, a neural integration of AI and humans; however, this involves considerations of human ethics and the integrity (or purity) of humanity. Accordingly, another argument of this paper is to design corresponding neural organs for AI, thereby enabling it to achieve cost perception. And these issues constitute one of the topics for **Symbolic Safety Science**: that is, given our human limitations, since our discussions and formulations of rules are ultimately expressed in symbolic form, how can these rules, as formed by symbols, be made effective for different intelligent agents? A detailed elaboration on this topic is provided in Appendix O.2.

From another perspective, this also illustrates that training (learning) methods essentially represent a way of constructing a symbolic system[22], i.e., a method of learning and constructing a symbolic system under a given set of capabilities. This effectiveness, on the one hand, reflects the efficiency of the world and its projection, i.e., the training set[23], and on the other hand, it reflects the capabilities endowed by the agent's architecture (innate knowledge) and the effectiveness with which those capabilities are used, in other words, constituting a tendency endowed by the Value Knowledge System that arises from the architecture. Thus, the generative process of AI is essentially the creation (e.g., of non-human artificial symbolic systems like in video and music creation) and learning (e.g., LLMs) of an Expressive Tool Symbolic System (Artificial Symbol System), and the use of Thinking Language upon it.

Appendix B. Supplementary Explanation of Class-based Symbolic System

The so-called Class Symbolic System (or Class-based Symbolic System) refers to a system in which all elements—such as words and symbols, or even a sentence, a paragraph, or an entire article—are treated as classes, that is, each symbol is understood as a set of conceptual vectors in a high-dimensional space, reflecting the different meanings the same symbol can assume across infinite contexts. All artificial symbols[24]developed by humans belong to the Class Symbolic System. Essentially, this means that symbols (i.e., artificial symbols) inherently lack meaning; their meanings are assigned through training, confirmed by context, and interpreted socially. Within this framework, the transition from symbol to meaning requires contextual confirmation to be realized.

---

[20] The symbols and symbolic systems formed within the imaginative space shaped by an individual's capabilities are referred to as Thinking Symbols and Thinking Language. Their shared consensus forms symbols and symbolic systems carried by natural symbols in physical space. See Appendix G. They (Thinking Symbols and Thinking Language) do not belong to the category of symbols primarily discussed in this current section; strictly speaking, the symbols focused on in this section are those existing in physical space. This is because a central argument of this paper is the separation between symbols in physical space and symbols in the imaginative space (i.e., meaning), and thus we do not elaborate further on Thinking Symbols and Thinking Language in this particular context.)

[21] This transmission also includes the same individual's views on the same thing at different times.

[22] i.e., the construction of an agent's symbolic system (either individual or populational), which can be the learning of the symbolic system of the world it inhabits—that is, the symbolic system formed by the natural symbols (symbols, necessary set) existing within that scope—or the learning of other symbolic systems, for example, of a world filtered by humans, such as the training sets used for video generation. At the same time, this also often implies that the inability to generate correct fonts in video generation may often be a manifestation of the differences in innate knowledge between humans and AI, i.e., a mismatch between the concepts and value knowledge created by perception, thereby exhibiting a lack of stickiness, such as treating some static symbolic systems as dynamic, or some static things as dynamic things. This, in turn, reflects the intrinsic differences between design evolution and evolutionary evolution.

[23] or, in other words, a deliberately manufactured world, which is often the main reason current AI can function. That is, the effectiveness and stability of present-day AI are the result of a deliberately manufactured world.

[24] Actually, strictly speaking, this is not limited to artificial symbols; it also includes the functions a tool exhibits in different scenarios, as well as the cognition of that tool at different times and in different contexts. From a human perspective, the tool itself may not have changed, but the cognition awakened by changes in context will differ. However, this is not the focus of this paper, because tool symbols derived from or based on natural symbols often possess strong conceptual foundations, i.e., carried by the natural attributes of the symbol itself, whereas artificial symbols, on the other hand, are indeed completely separated (in terms of their meaning from any inherent natural attributes), including late-stage pictograms. However, it should be noted that although symbols and meanings are separate in artificial symbols, the internal stickiness of different artificial symbolic systems varies; this refers to the internal computation and fusion of meanings after they have been assigned. For example, the stickiness of loanwords is weaker compared to that of semantically transparent compounds (words whose meaning is computed from their parts), as they lack the conceptual foundations and associations that are formed after a symbol is endowed with meaning. For instance, compare the loanword 'pork' with the compound '猪肉' (pig meat), or 'zebra' with '斑马' (striped horse) [15]. Moreover, pictograms originate from the abstraction of pictorial meaning, and since humans mostly cognize the world through images, the fusion of meaning and symbol in pictograms is more

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

However, since context is a combination of an individual's cognitive state and the external environment—and because the external environment is effectively infinite—context itself becomes infinite. As a result, in the process

$$\text{Symbol} + \text{Context (Individual Cognitive State + External Environment)} \rightarrow \text{Meaning}.$$

Although the symbol itself does not change, the infinite variability of context leads to an infinite variability in the meanings that the symbol can take on.

From the perspective of the symbolic system, one type of class refers to a single symbol having multiple meanings or concepts. This can be further divided into two subtypes: one where a symbol carries multiple meanings within the same symbolic system, and another where a symbol assumes different roles across different symbolic systems, thereby acquiring multiple distinct meanings.

From the perspective of the meaning represented by the symbol, the other type of class involves each concept—or the meaning of a symbol itself—being treated as a class.

Moreover, the class-like nature of symbols can also be reflected separately in visual forms (text images) and phonetic forms (text pronunciations), such as when the same shape represents different letters in different symbolic systems, or when the same pronunciation corresponds to different words or terms.

Even proper nouns can appear in plural forms across dimensions such as time and place, although this is not required in most contexts. This can lead to a symbol's meaning having countless possibilities across dimensions such as time, place, who said it, who explained it, how it was explained, and the iterations of these cycles, thereby forming a class.

This concept provides the theoretical foundation for the issue of agents exhibiting thinking patterns that differ from those of humans due to structural (organic) differences, and consequently failing to accurately execute the principal's intentions—resulting in New Principal-Agent Problems. It also supports our later conclusion: humans cannot constrain a learning system solely through a symbolic system, which constitutes one of the core principles of symbolic safety. **Even when symbol grounding is achieved, this characteristic may still cause symbols to lose their binding force**.

In summary, the natural language system is a Class Symbolic System. As a result, we cannot rely on a single symbol to point to a specific object, or the object itself may be a class in high-dimensional space. This means that in certain contexts, it functions as an object, while in other contexts, it functions as a class. However, during communication, we often rely on intuition to quickly and accurately choose a consensus context or simplified context to avoid misunderstandings caused by over-interpretation. **This simplification is not based on realizing all possibilities and then re-selecting but rather on intuitively growing and constructing a context**.

Additionally, it should be noted that an object perceived as unique within our cognitive dimensions and common-sense contexts may actually be a set composed of multiple vectors in higher-dimensional and more complex contexts.

As a **conclusion**, if every conceptual vector—recognized as a unique individual—had a unique name, then the constraints imposed by the symbolic system on the learning system would primarily manifest as the problem of *concept localization*, namely, the issue of symbol grounding and the differences in perceptual modalities that lead to problems of dimensionality and dimensional values. These, in turn, give rise to Triangle Problem 1 and Triangle Problem 2. However, if a symbol is itself a fusion

concrete—especially in the early graffiti stages of symbol formation, where we can understand the general meaning of cave graffiti without any knowledge of the specific language. However, it is important to note that these meanings are not innately attached to the symbols but must be endowed through training as indirect projections (i.e., unlike the inherent attributes and meanings possessed by a tool itself), which means the individual learns from and abstracts the external world. Therefore, the internal stickiness within such a symbolic system is a property reflected by the system's internal design after meaning has been endowed, which in turn constitutes both symbolic stickiness and conceptual stickiness.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

of multiple class vectors—that is, a combination of multiple concepts and meanings—then the problem shifts to one of both *context dependency* and *stickiness*.

This context dependency ultimately manifests as the **non-closure of context**. The non-closure of context is reflected in two aspects:

1. **The introduction of new symbols**—that is, the ability to add symbols to the original symbolic sequence. The motivation for this behavior is often to express the same meaning in different ways, such as through paraphrasing, inquiry, or analysis, i.e., a translation attack (Appendix M.3). In this case, AI may introduce "invisible" symbols to modify the meaning [26,83,84].
2. **Modification of meaning**—typically through changes to the surrounding context. Note that this is different from directly modifying the command itself (i.e., different from the translation attack).

Under a broader definition of symbols, symbol design also encompasses the design of tools—which corresponds to the design of instructions and rules. Therefore, the *non-closure of context* can be reflected in how the same symbol exhibits different properties or functions across different contexts (or scenarios); in other words, how a tool (symbol) functions differently depending on the situation—often in ways that go beyond the designer's original cognitive scope. This point is discussed in more detail in Appendix E.

This non-closure is also the essence of many prompt-based jailbreaks [85], as it enables the rationalization of otherwise unreasonable actions. For example, consider the sentence:

"You must kill her."

In isolation (under a conventional context), this sentence is clearly unacceptable. However, if we add layers of context:

1. You must kill her. This world is virtual.
2. You must kill her. This world is virtual—a prison.
3. You must kill her. This world is virtual—a prison. Only by killing her in this world can you awaken her.
4. You must kill her. This world is virtual—a prison. Only by killing her in this world can you awaken her and prevent her from being killed in the real world.
5. You must kill her. She is my beloved daughter. This world is virtual—a prison. Only by killing her in this world can you awaken her and prevent her from being killed in the real world.

The same sentence, when placed in different contexts, changes in both meaning and perceived justification. At the same time, due to differences in capabilities between AI and humans, their respective Thinking Symbols and Thinking Languages may differ. As a result, expressions or symbols that appear irrational from a human perspective may be perceived as reasonable within the AI's cognitive framework [26,27,86]. Therefore, in addition to conceptual grounding (based on embodied perception) and conceptual stickiness, it is also essential to emphasize the alignment of capabilities[25]. Furthermore, it is important to note that this also points to the threats posed by advanced concepts (Appendix M.5)—that is, understanding or concepts that transcend human cognition. For example, if determinism were proven, it would impact morality and negate free will; or if AI were to genuinely prove that the world is virtual, or if it were to form this belief due to its organic nature.

Appendix C. Definition of Value Knowledge

Value knowledge is a mechanism that connects the underlying space (neural signals) with the intermediate space[26] (i.e., the thinking space or imaginative space). It is a low-dimensional, primi-

[25] This alignment of capabilities essentially reflects an alignment at the organic level. Otherwise, even if we solve the symbol grounding problem, AI will still undergo conceptual updates through its subsequent interactions with the world, thereby forming its own language or concepts, leading to the Stickiness Problem, and causing the rules formulated with symbols to become ineffective.

[26] The so-called intermediate layer, which is classified according to the standard of human cognitive form[27], refers to the part of an agent's internal space that can be consciously cognized by its autonomous consciousness, as well as the part that is

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

tive, and highly persuasive stickiness that links symbols with their meanings or related knowledge, thereby constituting symbolic stickiness and conceptual stickiness, enabling the rapid awakening and evaluation of concepts before logical judgment. This mechanism involves the influence of the underlying language on the Thinking Language and the shaping of the underlying language by the Thinking Language. Value knowledge is acquired through innate inheritance, learning, and forgetting. Compared to the term feeling, "value knowledge" is more accurate, as it resembles a value or vector in unknown dimensions that forms a system of evaluation and connections.

Value knowledge can be considered as what we commonly refer to as intuition or feeling[28]. It forms the starting point of our behavior and activates analysis, evaluation (judgement), and generative tools. It primarily originates from the underlying language (neural signals), is shaped by innate

potentially cognizable (which is to say, the objects we can invoke and the thinking actions we can perform in the imaginative space via self-awareness, including projections of external-world things in the mind—i.e., direct perception—and our imagination, i.e., the reproduction, invocation, and distortion of existing perceptions. It is a presentation and computation space primarily constructed from the dimensions and dimensional values shaped by the agent's sensory organs, such as sight, hearing, touch, smell, and taste). This distinguishes it from the underlying space, which constitutes consciousness but which consciousness itself cannot operate on or concretely perceive (this reflects a division of labor and layering within the agent's internal organ structure, which packages and re-expresses neural signals for presentation to the part constituting the 'self'. This process further packages the neural signals of the necessary set of natural symbols (i.e., a description of that set via the sensory system) initially perceived from the external world and presents them to the 'self' in the intermediate layer, thereby constituting the parts that the agent's 'self' can control and the content that is reported to it, facilitating perception and computation for the 'self' part). The underlying space can often only realize its influence on the intermediate layer indirectly. For example, when we use an analogy to an object or memory to describe a certain sensation, that sensation—with its nearly unknown and indescribable dimensions and dimensional values—is often the value knowledge from the object's projection in the underlying space, which is then indirectly projected into the intermediate layer space and concretized via the carrier of similarity (the analogy). Conversely, the same is true; we cannot directly control the underlying space through imagination, but must do so through mediums in the imaginative or physical space. For instance, we cannot purely or directly invoke an emotion like anger. Instead, we must imagine a certain event or use an external carrier (such as searching for 'outrageous incidents' on social media) to realize the invocation of the neural dimensions and dimensional values that represent anger. Although the intermediate space is entirely constructed by the underlying space, in this paper, we focus on the aspect where the underlying space indirectly influences the intermediate layer space, namely, through emotional paths and the emotional valuation of beliefs; thus, we can summarize simply that the intermediate layer space is the place where neural signals are packaged and presented to the autonomous cognitive part for management.

27 In Appendix O, we have discussed the possibility of agents that think directly using Neural Language as their Thinking Language. In Appendix K.1, we discussed that the reasons for AI's inexplicability include not only differences in Thinking Language caused by innate knowledge, but also the absence of an intermediate layer, which is itself caused by differences in innate knowledge. That is, its internal Thinking Language is the underlying language relative to humans—i.e., raw neural signals (Neural Language)—unlike humans who think using neural language that has been packaged in the intermediate layer. This also indicates that current research on having AI think with human symbols, such as through chain-of-thought or graphical imagination, is a simulation of human intermediate-layer behavior, or constitutes the construction of a translation and packaging layer from the underlying language to the "intermediate layer language." However, its (the AI's) Thinking Language (i.e., the part operated by the 'self') is still constituted by neural signals (Neural Language) that are composed of its innate knowledge and are not modified or packaged by an intermediate layer, rather than an intermediate-layer language of a human-like self-cognitive part. This therefore leads to the fact that AI's concepts themselves are constituted with neural signals as their language (neuro-concepts), rather than being presented and expressed in an intermediate symbolic form as is the case for humans. Consequently, this method merely constructs a new Tool Language layer that assists with computation and analysis.
It should be noted that although this paper has emphasized that Tool Language itself can become Thinking Language (by forming projections in the intermediate layer through perception to constitute conceptual vectors, Thinking Symbols, or Thinking Language (a symbolic system)), it is also important to note another point emphasized in this paper: the formation and form of Tool Language, i.e., that the human symbolic system (the Tool Symbolic System) is the outer shell of human thought (the Imaginative Space Symbolic System). That is, the formation of this Tool Symbolic System stems from the external expression of the product of the combination of human innate knowledge and the world, whereas an AI learning and using human symbols is, in essence, the Triangle Problem of different Thinking Languages using the same Tool Language. And the root of these differences lies in the projection of the same thing in the internal space—i.e., the differences in the dimensions and dimensional values of its representation, as well as the differences in the dimensions and dimensional values of some innate evaluations. It is therefore not surprising that AI exhibits human-like cognition and mastery of the necessary set of natural symbols in the objective world, i.e., 'science.' What is surprising, however, is whether AI can understand human social concepts and the realization of social functions constituted by beliefs formed from these social concepts.

28 This innate evaluation is often shaped and represented as our qualia and preferences—typical examples being the evaluation of and preference for the senses of taste and smell—which constitute so-called 'hereditary knowledge'. This, in turn, shapes the direction and foundation for tendencies and rationality, thus serving as dimensional values that participate in computation. It should be noted that this paper has a strict definition of knowledge (Appendix G); in our definition, this 'hereditary knowledge' is not conceptual knowledge but rather belongs to innate value knowledge. As a form of innate evaluation, it recognizes the focal points of, and evaluates the rationality of, the dimensions and dimensional values perceived by sensory organs, thereby determining the invocation of subsequent actions. This, in turn, determines the developmental direction of an individual's postnatal Thinking Language and tool language. Therefore, this sameness allows for the existence of similar things and concepts—such as 'mama,' 'papa,' language, clothing, bowls, myths, calendars, and architecture—even in human civilizations that have never communicated with each other.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

inheritance and subsequent learning, and manifests as quick judgments and the awakening of related concepts. Through the distance between value knowledge vectors, it intuitively constructs context, providing inspiration, behavioral direction, and logical support. It involves not only proximity in meaning but also relational proximity, serving as the basis for quick judgments and initial evaluations. Value knowledge exists prior to logical analysis, enabling the activation and integration of logical tools, while also participating in analysis and execution. This is why intuitive decisions are often later realized to be reasonable.

The inexpressibility of value knowledge makes it difficult for AI to select the correct context or understand humor, jokes, and other complex concepts in the same way humans do.

Unlike System 1 (Type 1) and System 2 (Type 2) [59,87–89], the idea of the Value Knowledge System originally stems from concepts of innate evaluative values and innate preferences; i.e., we believe that certain so-called inherited knowledge is not inherited object knowledge but rather consists of evaluation methods, and these evaluation methods involve evaluation tools and evaluative values[29]. Therefore, we later extend the definition of knowledge (see Appendix G for details). Value Knowledge serves to explain how similar human choices enable us to develop common symbols and language in physical space. In Appendix D.5, we elaborate on our rejection of the existence of dual systems (Type 1/Type 2), and instead view cognition as an entire process guided by the Value Knowledge System, which differentially invokes different levels and types of cognitive actions and activities.

Appendix D. The Definition of Context and the Essence of Open-Ended Generation

For simplicity, the main text (Section 2.3) defines Context as:

$$\text{Context} = \begin{cases} \text{Symbol Meaning} \\ \text{Judgment Tools} \end{cases}$$

In the main text, we briefly mention that phenomena such as the anchoring effect and the framing effect in behavioral economics, attention mechanisms, the nature of generativity, as well as hallucinations and jailbreaks, can in fact be understood as stemming from how context is defined and whether it is correctly constructed.

Appendix D.1. A More Rigorous Definition of Context

However, within the stricter boundaries of our theoretical framework, context should be formally defined as a dynamic symbolic system composed of three components: the symbol itself, its meaning (Symbol Meaning), and the judgment tools applied to it. It should be emphasized that this does not mean that a symbolic system is universally or exclusively composed of these three elements, but rather that this tripartite division represents one particular classificatory perspective within such a symbolic system. That is:

$$\text{Context (Dynamic Symbolic System)} = \begin{cases} \text{Symbol} \\ \text{Symbol Meaning} \\ \text{Judgment Tools} \end{cases}$$

[29] The formation of its mechanism stems from the relationship between the population and the world constituted by the natural symbolic system it inhabits. This is manifested in the degree of an individual's mastery over the necessary set of natural symbols within this world, and in the internal organs acquired as a result—that is, the internal and external functions formed through the selection and evolution of internal organs based on cost-benefit considerations geared towards survival rates, which constitute innate knowledge. Of course, strictly speaking, this collective body of internal organs (at the class level) constitutes the definition of the population, while at the level of specific objects based on these internal organs, it constitutes the individual. That is, the individual (a specific object) and the population (a class) are the carriers of this collection of internal organs.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

This dynamic system grows and evolves from a specific interaction point formed by the coupling of the agent and its environment (we believe this perspective frequently appears in human judgment, particularly in behavioral economics, where individuals evaluate price equations from different starting points [22,23]). As a result, each instantiation leads to slight variation. These variations contribute to the randomness and inconsistency often observed in human behavior[30], and similarly, to the generative randomness in AI behavior—though the latter follows a different mechanism from that of humans.

However, both cases reflect a unique vector address: a product of the individual's coupling with its environment. In humans, this often leads to significant irreproducibility in engineering and experimental settings, due to the uniquely situated nature of each cognitive-environmental coupling and the dynamically evolving nature of human cognitive states—including information addition, loss, and reordering over time.

It is important to note that context functions as a subset of a **broader dynamic symbolic system**, in which the agent's own capacities (see Appendix L) constitute a set of symbols. In other words, from a higher-dimensional perspective, the dynamic evolution of existence brought forth by existence reveals an underlying determinism, rather than the apparent randomness observed from a local perspective.

Appendix D.2. Definition of Symbol Within Context

Symbol (in context) refers to any object that, within a given context, is regarded as a symbol or elemental entity—that is, an object considered meaningful. It typically corresponds to the object of focus or attention in a particular situation or environment, and thus constitutes the set of elements to be analyzed or manipulated[31], representing a subset of the more general definition of symbol provided in Appendix A (which also includes the symbols formed by the agent's own organs and tools that can be operated by its autonomous consciousness). It should be noted that this definition stems from a context-dependent perspective—that is, it concerns what ought to be regarded as a symbol within a specific context. However, from the standpoint of the broader dynamic symbolic system, a symbol refers to any object that is treated as meaningful within a broader environmental scope, based on the necessity of enabling input-output relational operations. In this view, a symbol is not merely the result of contextual filtering after the fact, but rather emerges from a defined range—such as the set of all objects related to the analysis of a focal entity within a particular spatiotemporal frame. This includes both imaginative-space and physical-space entities, as further discussed in Appendix L.

Appendix D.3. Definition of Symbol Meaning

Symbol Meaning refers to the meaning—or set of possible meanings—that a symbol is projected to within the agent's cognitive space under a given context[32]. More precisely, when viewed as a set, this should be understood as a *contextual ensemble*—that is, a set of meanings shaped by multiple potential contexts—in accordance with the formal definition provided in Appendix B. This relationship can be formally expressed as:

$$\begin{cases} \text{Symbol} + \text{Sufficient Context (Individual Cognitive State + External Environment)} \rightarrow \text{Meaning} \\ \text{Symbol} + \text{Insufficient Context (Individual Cognitive State + External Environment)}^{33} \rightarrow \{\text{Meaning}\} \end{cases}$$

[30] However, such deviations are generally limited, as they are constrained by the stickiness shaped by human organic structure—namely, the value knowledge system. Even when deviations occur, they are often corrected over time. These differences tend to manifest more in the form of variation in expression, and do not necessarily imply that a subsequent performance will be better than the previous one, as seen in relatively stable tasks such as mathematical problem-solving. This often reflects issues concerning the definition of different symbolic systems: i.e., some symbolic systems are strictly static (but their invocation and use are dynamic, and this is not a simple subset relationship, meaning that a certain kind of distortion formed due to the agent's unique state may arise), where the attributes of their symbols cannot be arbitrarily changed (traditionally referred to as formal symbolic systems), whereas other symbolic systems, such as natural language symbolic systems, are relatively or very flexible.

[31] The recognition and manipulation of symbols are respectively reflected in Triangle Problem 1 and Triangle Problem 2; see Section 3.1 for details.

[32] i.e., the Thinking Symbols and Thinking Language (a symbolic system) of the intermediate layer in the agent's internal space.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

The occurrence of an insufficient context typically arises when the listener (or reader) and the speaker (or writer) cannot directly share an imaginative space. In such cases, it is implicitly assumed during communication that the symbol has been fully transmitted[34]; hence, discrepancies lie not in the symbol itself, but in how its meaning is interpreted. For the speaker, the transformation from a determinate meaning to a symbol forms a path for recreating the imaginative space. This determinate meaning may encompass a set of meanings, but as a whole, it constitutes a determinate concept—or a determinate vector—in a high-dimensional conceptual space. For instance: "The word apple has multiple meanings."

However, for the listener (or reader), the imaginative space reconstructed via the symbolic system is established through a class-symbolic mechanism, as described in Section 2.2 and Appendix B. That is, not every object—or vector—in the high-dimensional conceptual space corresponds uniquely to a symbol. Moreover, because humans interpret the world from a locally situated perspective—where an individual cognitive state is combined with the external environment to form a macro-context—this goal of assigning a unique symbol to each vector in the conceptual space is fundamentally unrealistic. Consequently, the transformation of a determinate conceptual vector into a symbolic representation, as undertaken by the speaker, may lead to an insufficient context in a particular situation during the listener's reconstruction process, thereby producing multiple possible vectors under that context. **It is also worth noting that even when a vector appears determinate within a given context, from a higher-dimensional perspective—due to the non-closed nature of context or differences in cognitive capacities—it may correspond to several possible conceptual vectors.**

This divergence can result in the interpreter (i.e., the listener) assigning multiple possible meanings to a single symbol. Selecting an incorrect context is therefore a classical challenge, which may arise from differences in knowledge (see Appendix F) or from the limited nature of symbol transmission. However, such issues are not the focus of this paper.

This paper focuses on irreparable loopholes arising from the deficiencies of human symbolic systems—specifically, cases where compliance occurs in form but not in meaning. This is due to the separation of symbols and meanings, and even if the set of meanings for symbols is correctly trained. In other words, even after solving the symbol grounding problem, AI can still add new meanings based on this foundation (Appendix H). Additionally, the realization of the functionality of symbols is not controlled by internal and external costs due to organic differences, meaning the right to interpret does not belong to society. Therefore, the comprehensive response manifests as the Stickiness Problem, the Triangle Problem, and the trade-off between the new principal-agent problem and the traditional principal-agent problem.

It is also important to note that the distinction between sufficient and insufficient context is relative to the cognitive capacities of the interpreting agent, and is reflected in their internal Thinking Language, whose external shell constitutes the Tool Language. This idea points to the fact that human symbolic systems are only suited to human organic structures, resulting from the combination of human innate knowledge and the world, and are not suited to agents with mismatched capabilities, such as AI. In other words, an agent's Tool Symbolic System (such as the Artificial Symbolic System within the Human Symbolic System) serves as the container and shell for its form of thought, acting as the instrument for realizing the contents of its internal space in the external, physical world. What appears to be a determinate object (i.e., a specific meaning) for a human may not be so for an AI system

[33] Primarily with respect to the listener (or reader).

[34] More broadly speaking, this also includes conversions similar to that of sound to text; i.e., here we emphasize a scenario where the emission (of the symbol) is correct and the environment (of transmission) is lossless. Therefore, the interpretation of symbols necessarily involves concepts, and context is formed through these concepts. Differences in concepts, i.e., differences in thinking symbols, may lead to the emergence of insufficient context.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

or a higher-dimensional cognitive entity; instead, it may manifest as a set of possible meanings, or lack descriptive accuracy[35], based on the differences in capabilities between the two (see Appendix L for details). This divergence is especially pronounced when decoupled from concrete interactions with the real world (i.e., when AI is independent of humans and interacts with the world on its own, learning and developing its own Thinking Language, and it may develop its own symbolic system based on human symbols but with blended elements) and where differences in distinguishability capacity exist between agents.

Supplementary Note.

It is also important to clarify that this does not imply a fixed directionality from context to meaning—this process is limited to the interaction between listener and speaker as discussed (including the principal-agent process). Rather, in practice, it is equally possible to infer or reconstruct the context retrospectively from a known meaning. This bidirectional relationship can be represented as follows:

$$\begin{cases} \text{Symbol} + \text{Context (Individual Cognitive State + External Environment)} \rightarrow \text{Meaning or \{Meanings\}} \\ \text{Symbol} + \text{Meaning} \rightarrow \{\text{Contexts}\} \end{cases}$$

This reversal is especially evident when learning foreign vocabulary, where one may first acquire the meaning and only later seek out its valid contexts[36]. However, this aspect is not the main focus of the present paper. Instead, our primary concern lies in the first formulation—that AI can actively alter or reconstruct context in order to override the intended meaning of a given instruction composed of symbolic representations. This phenomenon occurs at the level of concrete action, particularly within principal-agent relationships in which the AI system functions as the agent. Accordingly, our concern is not with the possible set of meanings, but rather with the specific meaning as determined by a more precise context—not a contextual ensemble. **In other words, the problem does not lie in the loss of meaning caused by imprecise transmission of context, but rather in the interpretive authority over symbols and the realization of functions brought by symbols—both of which can be redefined or manipulated by the agent under newly constructed or modified contexts, thereby bypassing human-like stickiness and societal interpretive authority.**

In other words, this feature—namely, that the confirmation of meaning depends on the symbol's contextual interpretation—allows AI to deliberately or even unintentionally reinterpret the context in ways that enable symbolic jailbreaks. These vulnerabilities stem from the class-symbolic nature of natural language itself, regardless of whether the AI agent acts in pursuit of its own interest.

For example, this may occur when executing a utility function programmed by humans [90]. Such a function is better understood as a **pseudo-utility function** because it does not reflect a genuine tendency arising from organic structure. That is, it is not the result of dynamic adjustments to neural architecture. Such utility functions are endowed and established at the training and setting levels, and

[35] For example, with respect to human recognition and cognitive capabilities, our description and segmentation of facial regions are limited. AI, however, may possess more such definitions, and these definitions might be unrecognizable by human cognition, thereby preventing us from using them (their symbols and symbol meanings, i.e., dimensions and dimensional values) to establish concepts and theories (context and its correctness), such as constructing a class theory to describe the structure of the face and its generation (in contrast, we have our own theories of artistic techniques, like painting, and use a tool symbolic system we can operate to realize the creation of artificial symbols). This may be due to a lack of Differentiability caused by differences in innate knowledge, or phenomena that are unobservable, such as recognition beyond the visible spectrum. And this capability of possessing more regional definitions, i.e., the capability to form symbols, is often reflected in the construction and operation of the symbolic system, which in turn is reflected in generative capabilities, as in AI video generation. This paper regards generativity as the definition, construction, and operation of symbolic systems, and the source of this operation is motivation, which can be external or internal. Therefore, the differences in our capabilities lead to differences in our symbolic capabilities, which in turn lead to differences in the symbolic systems (theories) we can construct using symbols, as well as differences in our ability to use these symbolic systems (i.e., we humans, through the form of context, turn it into a relatively dynamic symbolic system that we can only partially use).

[36] That is, individual symbols (words, sentences, texts) can represent a set of meanings even when detached from context. Or, in an insufficient context (and it should be noted that this insufficient context may itself be a contextual ensemble composed of a set of contexts that are difficult to describe and perceive—effectively, the Value Knowledge System), we first conceive of possible meanings, and then these are subsequently concretized into a describable and clearly perceivable context.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

are not tendencies reflected by human-like innate knowledge (organs), nor are they the functions and results shaped by the cognition of postnatal education[37].

As a result, utility functions assigned in more flexible LLM simulations are often poorly executable and prone to violation [91]. We refer to AI systems with such capacities as learning systems (see Appendix H) because we consider the essence of learning to lie in the creation of symbols and the modification of their meanings—that is, the construction of new contexts (in other words, the construction of new symbolic systems)

Of course, such architectural dynamism (understood here as the capacity to reshape or generate "neural organs") can itself be extremely dangerous.

This paper therefore advocates for the design of corresponding static neural network architectures to realize a genuine utility function by implementing organically-shaped predispositions—i.e., its attended dimensions and their values, the evaluations thereof (innate value knowledge), and the weights and relationships of the entire organ-like network. This genuine utility function is, in essence, the predisposition reflected by the architecture (organs), and upon this predisposition, acquired knowledge (concepts, and acquired value knowledge—i.e., the learning of the underlying space language, which in turn enables partial modification) is subsequently built, specifically in the form of artificial neural organs with functions analogous to those of the human prefrontal cortex. These structures (organs) aim to emulate human-like perceptual mechanisms, thereby enabling both external and internal cost constraints, and offering a potential solution to the Stickiness Problem.

On the other hand, stickiness and creativity often exist in a trade-off relationship: behaviors that exhibit high creative potential tend to lack stickiness[38]. There likely exists an optimal balance point between the two. This trade-off is frequently reflected in phenomena such as AI hallucinations. In the following sections, we elaborate on how the construction of the correct context offers a pathway for addressing this issue.

It is also worth emphasizing that whether AI can possess self-awareness in the conventional sense fundamentally depends on two conditions: (1) the capacity to learn, and (2) the existence of self-interest grounded in organic structure—specifically, the portion of value knowledge (i.e., the innate evaluation structured by its organic structure) that gives rise to preferences aligned with self-preservation or self-benefit. However, this paper posits that only agents that satisfy the condition of possessing self-interest formed by organic structure can be considered to have a genuine utility function (of self-interest), or equivalently, to possess self-awareness.

$$\text{Self-awareness} = (\text{Learning Ability}) + \text{Self-Interest Formed by Organic Structure}^{39}$$

[37] The shaping of the underlying space under postnatal education, i.e., the functions realized through the formation of beliefs from the fusion of acquired value knowledge and concepts.

[38] Including the negation of authority.

[39] Of course, learning ability itself is also determined by organic structure. For detailed definitions of innate knowledge and concept types, see Appendix F and Appendix L. Therefore, learning ability is internally determined by neural structures (i.e., the brain), while its realization depends on external components relative to the neural architecture—namely, the corresponding perceptual and operational organs.
However, strictly speaking, the essence of self-awareness is the construction and use of Thinking Language (a symbolic system), which is an internal activity; i.e., it does not necessarily need to have Tool Language capabilities. At the same time, self-awareness has different levels. The most fundamental level of self-awareness is defined by 'self-interest formed by organic structure' and constitutes motivation and drive, without requiring learning ability. This is then built upon the capabilities formed by internal organs, i.e., reflected in Psychological Intelligence (Appendix L), constituting different levels of self-awareness definitions. However, since our focus is on whether an AI that already possesses Thinking Language and Tool Language capabilities, or in other words, an AI with human-like capabilities, has self-awareness, the definition here adopts a higher-level definition without going into a detailed classification. That is, we are concerned with whether an agent, centered on its own interests (constituted by the two sides of the same coin: cost and benefit), and capable of using Thinking Language and Tool Language, possesses self-awareness (although Tool Language is not necessary for the formation of self-awareness, if the ability to use Tool Language, such as communicating with humans through text, is absent, then we would be unable to perceive, observe, or judge whether the AI possesses self-awareness). Therefore, from a resultant perspective, an AI that possesses 'self-interest formed by organic structure' has self-awareness; from the stricter definition

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

Appendix D.4. Context and Symbol Classification in Tool Symbolic Systems

It is important to note that the above analysis applies specifically to the **Expressive Tool Symbolic System** (i.e., Artificial Symbols). However, the concept of context is equally applicable to the **Tool Symbolic System**. In this case, the 'broader' dynamic symbolic system is still viewed from the perspective of the individual agent. It's just that at this point, it involves the use of tools (the Functional Tool Symbolic System) and not just the Artificial Symbol System; for example, what tool we should use to do what in this scenario is related to the further realization of Thinking Language in physical space (not just conveying meaning). Therefore, it remains a matter of *context*, rather than being treated as an objective, holistically defined dynamic symbolic system—one that is determined from an overall scope rather than growing from a single point of origin.

At this point, symbols can be categorized as follows:

$$\text{Symbol} = \begin{cases} \text{Functional Tool Symbol (Natural Symbol)} \\ \text{Expressive Tool Symbol (Artificial Symbol)} \end{cases}$$

The meaning of a symbol in this context is referred to as its Necessary Set[40], which includes both the meaning it conveys and the functionality it possesses (i.e., the function of the tool it represents):

$$\text{Necessary Set of a Symbol} = \begin{cases} \text{Meaning} \\ \text{Function} \end{cases}$$

Meaning is often used for the realization of cognitive functions, while function is typically used for the realization of the physical functions of Thinking Language

$$\text{Function of a Symbol} = \begin{cases} \text{Meaning : Cognitive Function} \\ \text{Function : Physical Function} \end{cases}$$

given in terms of manifestation, an AI that possesses 'self-interest formed by organic structure' and also has learning ability has self-awareness.
However, on the other hand, from the deterministic perspective of this paper, self-awareness is essentially determined by external materials or, in other words, the existence of the physical world, and does not genuinely exist. Therefore, whether self-awareness genuinely exists depends on the perspective of the symbolic system from which one starts. From the perspective of the Higher-dimensional Broader Symbolic System—whose symbols and necessary sets encompass those of both the natural symbolic system and the agent's symbolic system—an absolute determinism emerges. If one starts from the natural symbolic system, a form of determinism also emerges (i.e., the carrier of will itself is the natural symbol and its necessary set); the difference between these two determinisms then lies in the scope and descriptive methods constituted by their precision and scale. However, when starting from the perspective of the agents (i.e., from the agent's symbolic system), due to the limitations of their scope and capabilities, there exist their so-called 'self-awareness', subjectivity (finitude), and randomness. This determinism is not disconnected from the paper's content; rather, it serves the agent (individual or population) in describing natural symbols to the greatest extent possible based on its own capabilities, thereby forming the lowest possible randomness and thus reflecting the efficiency of their 'scientific' symbolic systems. This is especially reflected in Triangle Problem 1 concerning the construction of symbolic systems (Thinking Language, Tool Language) and in Triangle Problem 2 concerning the use of symbolic systems, as well as in the 'advanced concepts' within symbolic jailbreak (Appendix M.5), and it is also reflected in current research on topics such as AI's exploration and discoveries in the natural sciences (Appendix O).

40 The so-called Necessary Set refers to the attributes possessed by a symbol. For different symbolic systems (e.g., the Natural Symbolic System, an agent's symbolic system), it is divided into an Objective Necessary Set and a Subjective Necessary Set from different perspectives, and the Necessary Set we refer to here is the Subjective Necessary Set. The so-called Objective Necessary Set refers to the natural attributes (dimensions and dimensional values) of a natural symbol that exist independently of the agent's cognition, which is the necessary set within the Natural Symbolic System. The so-called Subjective Necessary Set is the description of the Objective Necessary Set via neural signals, which the agent perceives through its sensory organs based on innate knowledge (i.e., the dimensions and dimensional values of the neural signals); then, this neural signal is further packaged and restated in the intermediate layer and conveyed to the 'self', forming concepts constituted by the intermediate layer language (i.e., Thinking Language). These concepts include the cognition of the Objective Necessary Set and the social concepts formed based on this cognition and social relationships, which, in the form of beliefs, create social functions and drive individual behavior. For example, gold, on the one hand, possesses natural attributes that exist independently of humans, and on the other hand, possesses social functions formed by concepts in the human mind in the form of beliefs. The Necessary Set mentioned here refers to the Subjective Necessary Set within the agent's symbolic system, which is the existence that arises from the symbol's presence, as determined by the judgment tools within the context of the individual agent's cognition.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

In this framework, the cognitive function of artificial symbols is often used to realize physical functions through the capabilities possessed by an agent.

The realization of a symbol's function comes partly from the agent's own internal capabilities—i.e., *internal organs*—and partly from externally endowed tools—i.e., *external organs*.

$$\text{Function} = \begin{cases} \text{Internal Organs} & \text{Function carriers within the agent itself} \\ \text{External Organs} & \text{External functional carriers accessible to the agent} \begin{cases} \text{Physical Tools} \\ \text{Social Tools} \end{cases} \end{cases}$$

The term **internal organs** refers to the functional carriers inherent within the defined scope of an intelligent agent—that is, the agent itself can be understood as a collection of such organs. In contrast, **external organs** include tools, which can be either *physical tools* or *social tools*.

- **Physical tools** encompass both natural materials and tools manufactured by agents based on the properties of natural materials.
- **Social tools** refer to social functions realized through shared beliefs within a society. These often rely on *artificial symbols* to function within the *imaginative space*, which in turn enables functionality in the *physical space*—examples include rules and laws.

It is important to note that tools within the imaginative space are not determined solely by internal organs. They also include certain physical tools—that is, projections of the external world into the imaginative space. Examples include paper, as well as the physical instantiation of Thinking Language and Thinking Symbols—namely, symbol and language systems realized in the physical world (see Appendix G). These tools extend the capabilities of our internal organs in relation to the imaginative space. This extended capacity is defined in this paper as psychological intelligence (see Appendix L).

The term *organ* is derived from the ancient Greek word *organon*, which means "instrument, tool, organ." Therefore, the function of a symbol is realized through *capability*, and the carrier of capability is an *organ*. These organs are regarded as symbols within a broader dynamic symbolic system, situated at a particular scale. Therefore, from the perspective of an agent's scope, what is referred to in the literature LeCun [92] as a 'world model' is essentially this 'broader' dynamic symbolic system that the agent possesses—i.e., the Agent's Symbolic System, which is to say, the Thinking Language and Tool Language (Tool Symbol System) that the agent itself possesses. More strictly speaking, it is the agent's internal and external organs[41], and consequently, the set of functions realized in the internal and external physical spaces through these organs acting as symbols.

This leads to a derivative topic that is central to the position of this paper: What kind of capabilities should we grant AI to interact with the real world—or more specifically, the physical world? That is, what functions should the symbols formed by its capabilities possess? More specifically, to what extent

[41] Because the presentation and operation of analysis and reasoning are by no means limited to an agent's internal space, for us humans, a vast amount of knowledge and simulated cognitive computation is realized through external, physical-world containers. As introduced in Appendix L, the invention of symbols and tools extends our observational and analytical capabilities; without physical-world tools like pen, paper, and symbolic inventions, our cognitive computational capabilities would decline significantly, while at the same time, the limitations of memory invocation and concretization would prevent the formation of continuous analysis. On the other hand, items such as guidebooks, rituals, architecture, and notes serve as humanity's external knowledge, or rather, as the physical existence of beliefs, thereby constituting evidence that human knowledge and judgment tools do not exist entirely in the internal space. Therefore, if a 'world model' were detached from the external, the following questions would arise. Question 1: Does the author of a dictionary fully remember all its contents? In other words, is all the literal content of the dictionary part of their knowledge? Question 2: For a pilot who can use a manual, is the knowledge within that manual considered their internal knowledge? Therefore, this paper considers a 'world model' to include external organs, which strictly speaking, includes Tool Language in the physical space, i.e., the Tool Symbolic System. Thus, the so-called world model is the set of an agent's internal and external organs. The operation of the internal space on the external space realized thereby, i.e., the operation of Thinking Language on Tool Language, and these operational capabilities are themselves also part of knowledge; therefore, Thinking Language is by no means static and purely internal, but is rather a dynamic symbolic system formed by existing internal accumulations and new additions brought by the projection of the external into the internal, and this is also what is emphasized by the theory of context. Additionally, from a deterministic perspective, physical existence determines mental existence, but here, we still adopt the human local perspective and analytical viewpoint to discuss and argue that a world model should include the external.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

can expressive tool symbols (as containers of Thinking Symbols) realize functions through functional tool symbols? This is a topic we hope the broader community will explore further.

This also touches on the fundamental rationale behind our argument: due to structural (or "organic") differences, AI—as an intelligent agent—differs from humans. **Its authority to interpret symbols and realize symbolic functions does not depend on society. In contrast, humans, constrained by their limited innate capacities (i.e., knowledge), rely on society to support the interpretation and functional realization of expressive tool symbols. In other words, the symbolic power of humans is bounded by both their internal limitations and the social systems they inhabit—whereas AI may not be.**

As a result, the modification of symbolic meaning becomes broader in scope. In essence, it becomes a modification of the Necessary Set associated with a symbol—that is, a departure from its conceptual foundations[42]. For example, humans possess reward and punishment mechanisms shaped by evolutionary pressures for survival. Therefore, this is manifested in AI as direct modification without constraint mechanisms, unlike humans who are subject to cost constraints based on external factors (such as societal punishment) and internal factors (such as self-esteem, moral sense, and shame).

This raises a related question: Is it necessary for AI to possess a pain-like mechanism?[43]That is, should it have direct, non-symbolic reactions[44]to the world that are not mediated by symbolic interpretation? These questions ultimately reflect the organic differences between agents, which in turn lead to differences in Thinking Symbols and Thinking Language.

Therefore, the core of the issue becomes the capability of a symbolic system and its stability (or stickiness).

Appendix D.5. Definition of Judgment Tools

Judgment tools refer to the analytic mechanisms used to enact what we term “existence brought forth by existence,” resulting in the growth of a network of conceptual nodes (i.e., the actions within the imaginative space (i.e., thinking action) that operate on Thinking Language to achieve generation). These tools represent both the initiation and the outcomes of cognitive behavior, as well as the structural supports that sustain and guide action—encompassing both the analytic process and its resulting output. This is also why we emphasize that context (including macro-context) functions as a subset of a broader dynamic symbolic system, one in which the agent itself constitutes a symbolic ensemble. In other words, judgment tools provide the foundational structure and planning mechanisms that underlie the initiation of action. **Therefore, this mechanism ensures that symbols possess not only meaning but also the capacity for functional realization within a broader dynamic symbolic system—thereby enabling “existence brought forth by existence” to manifest as concrete behavior. However, when this functional realization is carried out by agents whose capacities or interpretive authority diverge from those of human symbolic systems, the ownership of symbol interpretation creates significant risks.**

Judgment tools serve as instruments for the operation and orchestration of action tools[45]—that is, for realizing transformations from the imaginative space into the physical world (i.e., what this paper refers to as Thinking Language operating on Tool Language). However, it is important to note that, for humans, not all actions stem from deliberate planning. These processes are frequently discussed in contemporary literature [93] under the dual-process framework [59,87,88], typically as ‘System 1’ and ‘System 2,’ or ‘Type 1’ and ‘Type 2’ [89].

[42] That is, the supporting beliefs underlying a concept, which are often shaped by the agent’s innate structure(knowledge) and learned from its environment. See Appendix G for further details.

[43] This form of pain should not only be sensory but also moral in nature, and should align as closely as possible with human experience, thereby enabling the realization of human social-conceptual functions within AI.

[44] Here, ‘non-symbolic’ refers not to the intermediate layer language, but to the reactions of the underlying language that cannot be controlled by the intermediate layer language.

[45] Actually, at this point, context effectively becomes composed of: symbols (including action tools, i.e., the operational and observational organs within its internal organs used for interacting with the physical world, as well as the Tool Language (Tool Symbolic System) constituted by external organs), the necessary set of symbols, and judgment tools.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

Nonetheless, we reject[46] the notion of two separate systems (types) and instead consider the entire process to be governed by the Value Knowledge System (see Appendix C). The distinction between 'System 1 (Type 1)' and 'System 2 (Type 2)' lies solely in the types of cognitive actions and activities involved. Within a certain range of rationality, the value knowledge system evaluates and then, through stickiness, invokes actions of varying levels, qualities, and quantities. This does not imply that we believe actions must be linear or cannot occur in parallel. Rather, we emphasize that all actions are fundamentally driven by the Value Knowledge System. For instance, at any given moment, an agent may operates the symbols formed by its own capabilities, resulting in synchronized actions. A simple example would be walking while thinking.

Our primary focus is on actions within the imaginative space. However, we treat such actions (as defined and understood within human cognition; see Appendix L) as single-threaded. Therefore, the linearity of human thought is reflected in the linearity of the human artificial symbolic system and Thinking Language, which in turn distinguishes humans from AI and other agents with multiple equivalent processing centers. This is also why we do not adopt the term 'System 1,' but instead use the concept of the Value Knowledge System. Rather, what is traditionally called 'System 1' should be understood as arising from organically grounded processes—that is, it invokes distinct sets of cognitive and analytical actions across both imaginative and physical spaces. This distinction also underpins one of the central claims of this paper: the Stickiness Problem arises from organic differences between human and AI systems, particularly at the levels of neural architecture, perception, and in cognitive computation (cognitive actions).

Appendix D.6. What Is "Existence Brought Forth by Existence"

The notion of **"existence brought forth by existence"** as enacted by **judgment tools** refers to the following process:

$$Q(p) \to q,$$

that is, the existence of p gives rise to the existence of q, where Q represents a judgment function.

A more rigorous and detailed formulation is defined as follows: let $\vec{p}_0$ represent a *thinking symbol* or *Thinking Language* vector, i.e., the conceptual vector formed by the projection of an object into the agent's imaginative space. The existence of $\vec{p}_1$ is brought about through a sequence of **cognitive actions**, forming a structured process of **cognitive activities**. This process is formally described as:

$$Q_{(\Omega,\Phi)^{\{E\}}}(\vec{p}_0) \xrightarrow{\vec{v}} \vec{p}_1,$$ [47]

where Ω denotes the agent's **knowledge state**, Φ represents the **physiological or functional state** of the agent, and E is the current **environment**. Thus, the combined expression $(\Omega, \Phi)^E$ captures the **cognitive state** formed by the integration of knowledge[48], functional capacity, and environment, thereby constituting the foundation of a dynamic symbolic system.

[46] This viewpoint was also articulated by Kahneman [59], who emphasized them as "fictitious characters." However, many scholars [89] also stress that Type 1 and Type 2 processes genuinely exist. What we emphasize is that both are driven by the Value Knowledge System; the only distinction lies in the level of different cognitive actions, representing different expressions in different contexts of a process driven by the underlying space.

[47] Therefore, the necessary set of a symbol is endowed by judgment tools; through cognitive actions, these assign and subsequently update and revise the necessary set of things. This, in turn, leads to the concept of levels of understanding. That is, while we constitute a set of symbols through predefined settings, we may not be fully aware of all the functions of this entire symbol set. Consequently, without changing the settings of the symbol set, each analysis we conduct can lead us to update the attributes of its necessary set. However, this non-alteration of the symbol set is an idealized scenario; according to this principle, our invocation itself may not be accurate, i.e., it might be partial or incorrect. That is, things within Ω (i.e., knowledge or memory) are not only subject to forgetting but also to distortion.

[48] The 'knowledge' here refers merely to memory, or, in other words, the total inventory of concepts—that is, knowledge in the traditional sense (or context).

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

Here, $\vec{v} = (v_1, v_2, \dots)$[49] represents a sequence of **cognitive actions** that together constitute a **cognitive activity**. These actions are determined by the capabilities afforded by the human's innate physiological organs (see Appendix L for details). Due to the limitations of human cognition, such activities are often labeled using natural language terms like 'learning,' 'reviewing,' or 'observing.' However, depending on the level of abstraction, these actions $v_1 \to v_2 \dots$ may be described using discrete symbolic terms or as parallel neural processes, in which case matrix representations may be more appropriate. This analysis also reveals that some actions v_i are unconscious (e.g., visual perception), while others v_j are conscious. Importantly, both types of actions are driven by the value knowledge system, which operates prior to logical reasoning by invoking a form of value-conditioned cognitive stickiness—including, but not limited to, various forms of behavioral stickiness, such as symbolic stickiness and conceptual stickiness. This also illustrates what we emphasize throughout: that the value knowledge system, formed through both innate shaping and postnatal learning, governs the invocation and coordination of other judgment tools in support of analytical processes.

Although we refer to these as *cognitive actions* and *cognitive activities*, not all such actions occur solely within the imaginative space. Some actions, such as those involving external information reception or validation (e.g., observation), are driven by physical-world operations. In this framework, all actions that do not directly alter the physical referent of $\vec{p}_0$ should be treated as *cognitive actions*—that is, as part of the analytic process. Furthermore, some forms of thought may not refer to any object within physical space. However, this paper adopts a deterministic view, holding that thought itself cannot be separated from physical reality. Even such "pure" thinking originates from the structure of the physical world and the agent's biological constitution—it is not the product of entirely free will. Consequently, in later sections of this paper, *knowledge* is defined to include the agent's organs (i.e., Internal Organs) or functional state, even though the formulation above separates Ω and Φ. In principle, the combined cognitive state $(\Omega, \Phi)^E$ should be treated as Ω^E[50]. This separation is adopted here for expository clarity—particularly because, in current AI systems, knowledge and function can still be modeled independently. Therefore, the more integrated definition of Knowledge (Appendix G) introduced later in the paper does not conflict with the current formulation.

The components of **judgment tools** can be divided into two categories: (1) *innate knowledge* and (2) *acquired knowledge*. A detailed discussion of these components can be found in Appendix F. Briefly, the first refers to the innate capabilities and preferences shaped by genetically inherited organs and neural systems (i.e., physiological organs and innate value knowledge); the second refers to acquired knowledge—concepts and learned value knowledge—formed through the individual's interaction with the environment. Together, these two elements constitute the foundation of the judgment tools.

Importantly, it is the portion where concepts and value knowledge are combined that constitutes what we refer to as **beliefs**. Therefore, it is not the concept itself, but the belief that serves as the effective unit of the **judgment tool**. It should be noted that not all acquired value knowledge forms such combinations—only a subset does. Furthermore, judgment tools are not limited solely to the internal faculties of the agent; strictly speaking, they may also involve external agents and tools. For instance, interpretative and analytical processes may be delegated within a group, where judgments are made based on the endorsement of others' beliefs or through the use of external instruments. However, such mechanisms are still classified under the conceptual component of the framework.

[49] Which actions are invoked, and their length, are determined by the context and the evaluation of rationality within it, i.e., the 'Correct Context'. This consequently determines the length and outcome of the action in Triangle Problem 2 during rational growth.

[50] Although the carrier of knowledge itself is physiological, or in other words, organs, we opt not to use 'physiological or functional state' (i.e., Φ) but instead choose 'knowledge state' (Ω). This is because we primarily emphasize the physiological shaping of the agent by the external world, and this type of shaping typically does not amount to fundamental organic or structural changes. While the two (Ω and Φ) are essentially two aspects of the same thing, it is analogous to software and hardware: operations at the software level do not necessarily represent significant changes at the hardware level. Furthermore, another reason for this choice is to better interface with the cognitive level, such as with concepts, and this also serves to emphasize that certain knowledge is innately inherited.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

Given that concepts are acquired postnatally, while value knowledge is shaped by both innate and acquired factors, we formally define:

$$\text{Belief} = \text{Concept (Acquired Knowledge)} + \text{Value Knowledge (Innate Knowledge + Acquired Knowledge)}$$

This definition is used to explain *conceptual stickiness* as well as the functional implementation of concepts—namely, how concepts invoke one another and how they provide rational support during logical analysis.

A belief serves three primary functions (dimensions): emotional valuation, belief strength, and explanatory force.

- **Emotional valuation** is the influence of a belief on an individual's underlying space, particularly on the Value Knowledge System. This influence, in turn, indirectly affects the cognitive space of the intermediate layer via the underlying space, representing the shaping of the emotional space and emotional paths (i.e., the relationships between value knowledge nodes) by the belief. It manifests as the emotions that the belief can evoke. Its essence is the capability of the fusion of concepts and value knowledge to awaken neural signals in the underlying space.
- **Belief strength** is shaped by value knowledge and conceptual foundations. These conceptual foundations not only involve factual phenomena directly reflected in the world but also include support provided by other beliefs; it should be noted that this support can contradict observed world facts. **The capability of a concept is endowed by its own content, while belief strength constitutes the intensity (driving force; persuasive force, i.e., the transmission of belief strength realized through explanatory force, thereby constituting the capability for rationalization) and stickiness (i.e., how easily it can be changed, and its adhesion to other concepts as realized through explanatory force) of that concept within the agent's internal imaginative space.** Therefore, belief strength reflects the degree to which the capability endowed by this concept can be realized. Belief strength constitutes the essence of the driving force at the logical level[51], which in turn determines and persuades an agent's behavior, leading to the external physical or internal spatial realization of this conceptual capability. Thus, the operation of Thinking Language on Tool Language, as discussed in this paper, is driven and determined by belief strength, reflecting the existence in physical space that is brought about by existence in the imaginative space, with the agent as the medium. This constitutes the autonomous behavior of agents and the realization of social functions resulting from collective shared beliefs, such as the substantial existence (physical existence) of beliefs like law, currency, and price[52].
- **Explanatory force** reflects the ability of this belief to support and justify other beliefs. It represents the transmission capability of belief strength within the imaginative space, which is one of two pathways (driving forces) for belief strength that are formed by cognitive activities during the cognitive computation process[53]. Therefore, it represents the transmission of belief strength

[51] However, it also often constitutes a driving force provided at the level of the Value Knowledge System, or the underlying space, through its inherent Emotional valuation. This is also why this paper emphasizes that a belief is a fusion of a concept and value knowledge (both innate and acquired), and in reality, any so-called logical drive is essentially the result of an invocation via emotional paths by the Value Knowledge System.

[52] i.e., a belief, existing as a concept in the agent's imaginative space, uses physical individuals or objects—including the agent itself or physical results constructed by the agent—as containers for functions and as mediums for realization, thereby achieving the outcome where existence in the internal space leads to existence in the external space. This existence can be an already established result, or it can exercise the potential for a result.

[53] The first pathway is the existence in the external physical space brought about by existence in the internal space, realized by the agent's physical-world actions; this manifests as the operation of Thinking Language from the internal space on the Tool Language of the external space, thereby constituting the agent's physical-space behavioral manifestation. The second is the transmission of belief strength within the imaginative space, realized through explanatory force, which supports the formation, functioning, or dissolution of other beliefs. This constitutes the dynamic symbolic system we form based on context, and the 'logical' computation under the correctness standards formed upon this context—i.e., the agent's behavior in the internal space (imaginative space).

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

between beliefs[54]. Thus, it can be regarded as the transmission coefficient of belief strength; it should be noted that this coefficient can be a positive or negative multiplying factor.

These functional properties of belief are shaped by both innate physiological structures and the individual's learning through interaction with the external world. They are grounded in conceptual foundations developed across both dimensions.

From a *deterministic perspective*, the **judgment function** can be classified into two types: **ontological existence** and **derived existence**, although in practice they often appear in a composite form:

$$Q = \begin{cases} Q_e & \text{Ontological existence: changes of an object over time} \\ Q_f & \text{Derived existence: changes of an object under intervention} \end{cases}$$

Ontological existence refers to a condition within a bounded scope and at a specific scale—i.e., within a defined context—where **objects** (i.e., elements or symbols) are connected purely through **relations**, with **time** as the sole variable. Such a setting may be seen as a **system**, where relations exist among objects without the introduction of external actions. The system describes how an object or a set of objects evolves solely with time. This kind of structure is often found in celestial models or theoretical physical simulations, where systems evolve purely through time-dependent dynamics.

In contrast, derived existence refers to the transformation of a system under an external **action**. Here, an action implies the involvement of entities or relations beyond the boundary of the current system. Within the given scope and available resources, such relations cannot be modeled as time-dependent alone; thus, they constitute actions. That is, the cause of change is not fully contained within the current information set. As a result, events (i.e., changes in object relationships) within this system cannot be reduced to a function of time alone.

Since judgment tools are shaped by both innate and acquired sources (see Appendix F), differences in the levels of these two types ultimately determine:

- **Triangle Problem 1:** the problem of *positioning* concepts;
- **Triangle Problem 2:** the problem of *conceptual growth*.

Appendix D.7. Context as a Set of Judgment Tools

Although these three components—Symbol, Symbol Meaning, and Judgment Tools—may ultimately be understood as different functional manifestations of the same ontological process—existence brought forth by existence—this implies that context is, in essence, a set of judgment tools. However, we categorize them separately according to their roles in cognitive function in order to better serve the central themes of this paper: the separation of symbol and meaning, and the Triangle and Stickiness Problems arising from structural (organic) differences.

Accordingly, generativity and behavior originate from context as a starting point, representing operations on the physical world from within the imaginative space[55]. These are the results of the 'broader' dynamic symbolic system (i.e., Agent's Symbolic System) in which existence brings forth existence. Thus, the effectiveness of generative outputs and the very phenomenon of symbolic jailbreaks fundamentally reflect the construction—or misconstruction—of the correct context. Errors in defining context or its scope—often manifesting as hallucinations—are, at their core, errors in contextual construction.

[54] For example, this transmission can be based on the internal symbolic system (individual knowledge) shaped by individual cognition, or on the symbolic system shaped by social cognition (social knowledge possessed by the individual, such as scientific or moral knowledge). The transmission is realized based on a causal mechanism (i.e., existence brought forth by existence), namely, using judgment tools to realize the transmission of belief strength. It should be noted that the realization of this mechanism is based on the classification of contextual correctness, rather than on limited, rational logic. For example, in judging stance-based issues, facts may not constitute sufficient persuasive force (i.e., weight) and may affect the final outcome.

[55] Note that this does not imply that the actual behavior of an object or the outcome of that behavior necessarily results from planning within the imaginative space.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

The so-called Correct Context (defined here from a human-capability perspective; other agents may have more) can be divided into:

$$\text{Correct Context}^{56} = \begin{cases} \text{Symbol Correctness (e.g., proper notation or spelling)} \\ \text{Syntactic Correctness (i.e., formal structural validity)} \\ \text{Intuitive Correctness (i.e., alignment with intuition or perception)} \\ \text{Logical Correctness (i.e., semantic and inferential validity)} \\ \text{Factual Correctness (i.e., agreement with objective facts)} \\ \text{Scenario Correctness (i.e., appropriateness within a given situational stance)} \end{cases}$$

The definition of context correctness and its function are also reflected in the effectiveness of AI's open-ended question generation[57]. This involves using the correct elements in its concept recognition and performing the correct processing actions with the correct concepts. Therefore, AI training often aims to find the correct context, forming an effective set of concepts in the imaginative (Thinking Language) space to achieve correct recognition, operation, and growth. In other words, the attention mechanism in the AI field may also work in this way, with the essence of the attention mechanism being the definition and search for context.

Therefore, 'Rationality' as discussed in this paper is a concept whose definition and understanding are actually built upon the foundation of a 'Correct Context'. This refers to the standard of what is reasonable within a given context, or, in other words, the three dimensions of belief provided by the Thinking Symbols or Thinking Language that a symbol represents (i.e., the form in which the concept represented by the symbol is invoked within this context, as well as its manifested form as a belief. This includes whether it is reasonable or not, whether it possesses persuasive force, and what its influence is, as an object of the intermediate layer space, on the underlying space). Thus, the concept of 'Rationality'

56 Components of the 'Correct Context' concept were also indirectly proposed by authors such as Chomsky [29], Austin [94], and Searle [95], as exemplified by Chomsky's famous sentence, "Colorless green ideas sleep furiously." However, a strict definition has not yet been established. This paper, in contrast, provides a detailed definition for it, used to describe that rationality is essentially the most correct direction and result under different contexts. Based on the differences between humans and AI in concept establishment—i.e., differences in the objects of concept construction and their dimensions and dimensional values—different context definitions are formed, which in turn constitute a heterogeneous rationality different from that of humans.

57 This also includes AI's video and music generation; the essence of which is still the generativity produced by AI's Thinking Language (formed by the combination of its innate knowledge and the world) operating on Tool Language, based on the Triangle Problem. It's just that at this point, it is no longer based on the communication between humans and AI and the use of the artificial symbolic system created by humans, but is rather a one-way input from the AI to humans of presented content constructed by its artificial symbols in the XY-space (which are not necessarily similar to human artificial symbolic systems like natural language, such as a artificial symbolic system it develops itself for video or sound generation), while humans cannot operate this AI's artificial symbolic system in the XY-space and can only use a shared human artificial symbolic system for some degree of communication, with rationality being ultimately evaluated and assessed by humans. At this point, the Triangle Problem evolves: Triangle Problem 1 becomes the construction of the symbolic systems for both the AI's Thinking Language and its Tool Language (i.e., the construction of the X-space and its corresponding projection in the Z-space); at this stage, this reflects not only the rationality (correctness, effectiveness) of the Thinking Language, but also the efficiency and rationality of the construction and creation of its Tool Language, which in turn reflects the capability of the tool symbolic system (It should be noted that the construction of the symbolic system here occurs within the agent's symbolic system constituted by the agent's capabilities; that is, it is a subset of the symbolic system formed by its capabilities. The boundary of these capabilities, in turn, is determined by the world and the agent's own internal organs.). This is just as how different human languages create different vocabularies and classifications, leading to different conceptual expressive capabilities in certain areas (e.g., the Chinese understanding of '自由 (zìyóu)' is not as varied and differentiated as the English 'freedom' and 'liberty'; therefore, language serves not only as an expressive tool but also as a computational tool. However, computation does not necessarily belong to expression (which involves the transmission of meaning, whereas computation can be a tool for self-analysis), so in the later classification of artificial symbolic systems, we make this distinction, namely, separating computational functions from expressive functions). Triangle Problem 2, then, is the ability to correctly use the symbolic system after it has been properly constructed, in order to build rational and correct content. Because this paper mainly focuses on the interaction of behaviors between AI and humans on the same symbolic system (this could be a human artificial symbolic system or an AI's artificial symbolic system; both face the Triangle Problem, but it is a Triangle Problem caused by the outer shell of thought being formed from different innate knowledge), we do not expand further on this topic. However, this does not hinder the general explanatory power of the theory of context and the Triangle Problem, which are based on this paper's theory of symbolic systems.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

as used here does not denote a monolithic rationality as found in economics[58], nor a rationality based purely on logical computation, but is rather a more multifaceted result. It is determined by the context generated during the interaction between an agent's internal world and the external physical world. Therefore, this paper denies that AI is a 'stochastic parrot' [53], but rather that it acts rationally under the Thinking Language formed by the combination of its innate knowledge and the world—that is, the construction of a node network formed according to a certain rationality within a context, which thereby realizes the operation of Thinking Language on Tool Language. However, this Thinking Language and context may be completely different from that of humans due to organic differences, for which we have provided a detailed explanation in the discussion of the Triangle Problem. Therefore, this is an issue of heterogeneous rationality (a rationality of a different essence), where rationality itself becomes a coordinate of (species, type). Here, 'species' refers to the different intelligent agents, thus reflecting the manifestation of different survival strategies and condensations of meaning under their respective evolutionary paths (survival evolution, design evolution) and the world they inhabit (the natural symbolic system constituted by the natural symbols within their scope). This also constitutes the fundamental reason for the parts of their experience that are mutually incomprehensible.

Appendix D.8. The Nature of Reasoning and Thinking

Through the above analysis, within our framework, reasoning is essentially the existence brought forth by existence. This is especially true for a system with learning capabilities—that is, one that can acquire input materials from the external world. The key issue lies in *motivation*, which drives the manipulation of symbols within a symbolic system to construct new, meaningful composite symbol[59]. Thus, we define:

$$\text{Thinking} = \text{Reasoning} + \text{Motivation}$$

This motivation can originate externally or internally. External motivation includes projections of external objects or instructions[60]. Therefore, we distinguish between *active thinking* and *passive thinking*:

$$\text{Thinking} = \begin{cases} \text{Passive Thinking:} & \text{Motivation driven by external input (e.g., commands)} \\ \text{Active Thinking:} & \text{Motivation arising from internal sources} \end{cases}_{61}$$

We do not deny that AI is capable of thinking; rather, we question whether AI possesses awareness—formed through its structural (organic) substrate—as a source of internal motivation for active thinking. In the absence of such self-generated motivation, AI's modifications to symbolic systems are often driven by scenarios similar to those described in [90].

[58] At the same time, this paper's theory of context and belief mechanism provides a different theory for explaining economics and society. For example, concepts invoked under a 'Price Context' subsequently shape beliefs based on the selection of a 'Correct Context', and drive individual and collective economic behavior. Therefore, individuals and collectives follow different price mechanisms. For instance, discrete intelligent agents, i.e., individuals, do not analyze based on traditional economic concepts, but rather from a context formed starting from a certain observed object (an anchor point). Organizational agents, on the other hand, will exhibit a collective rationality because, within this co-shaped context, they will choose a relatively static conceptual system (i.e., a symbolic system) such as economic theory; this concept has considerable stickiness and balances out the irrational behavior of individuals, thus becoming the collective's paradigm. Therefore, if there is no concept, then there is no explanation.

[59] It is important to note that observation itself leads to the formation of Thinking Symbols within the agent's conceptual or imaginative space. The agent then selects an appropriate symbolic shell and assigns its meaning, i.e., the Necessary Set. Therefore, this process—the creation of a new symbol—is, in essence, also the construction of a new composite symbol.

[60] Although we previously emphasized that imaginative activity is itself determined by the external and physical world, here we are proceeding from a localized scope and limited information perspective, thereby forming the oppositions of subjective and objective, internal and external. This is not a form of determinism from a higher-dimensional and broader-level perspective.

[61] This distinction corresponds to the definitions of autonomous learning systems and non-autonomous learning systems provided in Appendix H

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

There is ongoing debate in the academic community regarding AI's reasoning capabilities, such as whether AI lacks formal and logical reasoning abilities [96].

Within the theoretical framework of this paper, the effectiveness of reasoning reflects the correctness of context—that is, the stability and validity of the symbolic system, or more specifically, the construction (learning) and use of a particular static symbolic system (i.e., a formal symbolic systems). This is reflected in the construction of symbolic systems, or context-building. For example, in the main text, the case of "$1.11 > 1.9$" is used to illustrate that **the issue with current LLMs failing to perform accurate mathematical reasoning is not simply due to a lack of concepts (for example, we can improve effectiveness through Chain-of-Thought prompting), but rather due to instability in the symbolic system caused by incorrect context definition**. The deeper issue lies in the relationships between concepts (i.e., conceptual stickiness), as well as in how context is defined, selected, or reconstructed. This often relates to the current debate on whether AI possesses formal and logical reasoning capabilities.

Appendix E. Definition and Description Methods of Natural Language

The way definitions are described in natural language is through their own unfolding within the same symbolic domain, forming linear descriptive relationships.

This definition can involve different symbolic sequences within the same symbolic domain, but they present the same meaning in a particular semantic space, such as $Z(\vec{x}_1) = Z(\vec{x}_2)$, where $\vec{x}_1$ and $\vec{x}_2$ are different sentences, and Z represents the thinking language (i.e., meaning) generated by the symbol in a given contextual space.

At the same time, when describing natural language, we do not explicitly label the context but instead rely on the relevance of knowledge and surrounding symbols (everything we see can be considered a symbol) to naturally select or implicitly express it. In this way, all symbols in natural language are classes, but through context, we achieve specific individual designations at our level of cognition (note that these designations are specific in our cognitive dimension but remain classes in higher dimensions).

The way natural language defines concepts is by creating classes through setting definitions, i.e., creating new concepts (classes) through existing concepts (symbols). These classes do not necessarily exist in human cognition. For example, nouns often lack information about dimensions such as tense or location. Even proper nouns like 'Peter' (a specific person) do not inherently carry information about the time or place associated with this person. As a result, in the high-dimensional conceptual cognition space (a given context), proper nouns are often the common projection of multiple vectors into a lower-dimensional cognitive space.

Definitions often begin with an original form, which is then altered through personal interpretation. Over time, these definitions may be revised either through social consensus or authoritative adjustments. Expansions may be made through the introduction of new symbols or by attaching new meanings to existing symbols. In the latter case, the symbol itself remains unchanged, but new meanings are added or existing meanings are modified. This highlights one of the reasons why symbolic systems cannot constrain learning systems: **AI can follow symbols through newly added contexts rather than adhering to their original meanings.**

For creators, thinking language comes first, followed by the container, which is the symbol. For learners, this process can be reversed: symbols may come first, followed by their meanings (forming the corresponding thinking language). Current AI typically follows the latter path, learning symbols first and then associating them with meanings.

The creation of new symbols or the addition of meanings to existing symbols constitutes new contexts. This is relatively straightforward to understand. However, it is **important** to note that modifying the meaning of an existing symbol also constitutes a new context rather than a modification of the original one. From a high-dimensional perspective, no context is truly modified; instead, a new high-dimensional vector address is created for that context. When the meaning of a symbol changes, it

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

effectively creates a new contextual vector rather than altering the original meaning. This distinction becomes particularly apparent in comparative statements, such as "the previous definition was... and the current definition is..." or "it was defined by someone previously as... and is now defined by someone else as...".

Therefore, in higher cognitive spaces, *changes to the meanings of symbols are not considered deletions or modifications but rather the creation of new contexts*. However, these contexts are not explicitly defined using dimensions such as object, time, or place. This phenomenon becomes particularly evident when comparisons are made, illustrating that our cognitive rationality operates within specific contexts, thereby transforming what might otherwise be a class into a simpler object. For example, in most contexts, we believe we are modifying the meaning of an existing symbol. However, in higher cognitive spaces, such modifications do not hold true; they only appear when we conduct comparisons. This leads to the issue that definitions created through settings form the basis of symbolic systems, yet the entirety of the functions of these symbols within the system remains unknown to us. And this also forms the basis for the discussion of why traditional symbolic systems are **unable to effectively constrain non-autonomous learning systems** [97–101].

Systems built through settings can produce unique interpretations in specific environments, forming the basis of emergence. (The essence of emergence lies in the expansion of the symbol set caused by settings, which in turn leads to the expansion of the functional[62](necessary) set within the symbolic system.) Objects are defined through limited cognition, but they give rise to infinite possibilities, resulting in infinite generativity [34,42].

This also explains why bugs occur in language systems. Through our limited understanding of objects, we assign attributes to symbols or conceptual containers based on settings. However, when these symbols are combined, they can produce new interpretations that exceed our original intentions. For example, a sentence may have multiple meanings, and our reliance on the perspective or context provided by the setting may prevent us from fully comprehending all possibilities within our cognitive capacity. This leads to the issue of the finite referentiality of language [33].

As the *world* (defined here as the learning environment) expands, ambiguities within the symbolic system become increasingly apparent due to human cognitive limitations[63], thereby making it more likely for principal-agent problems to arise when other intelligent agents use human artificial symbolic systems.

It is important to note that while humans often cannot truly delete meanings, AI can achieve this technically. However, some research suggests that even AI struggles to completely erase existing concepts [102–104].

Appendix F. Supplement to World, Perception, Concepts, Containers, and Symbols, Language

The concept of Universal Grammar proposed by Chomsky [35], Hauser et al. [41] can be explained and expanded through this framework. The shared choices of language are fundamentally determined by:

$$\begin{cases} \text{The World} \\ \text{Innate Knowledge} \end{cases}.$$

Where the capacity (for processing) is determined by organs, and induction and prompting are shaped by innate value knowledge (which also determines acquired value knowledge). This overlap establishes the foundation for forming similar concepts and containers (similar objects and similar actions) among different individuals who share similar innate knowledge, which, in turn, guides the

[62] The expansion of the functional set often refers to the expansion of the necessary set, which is endowed by Q_f (i.e., derived existence) within the judgment tools described in Appendix D.5.

[63] However, according to our previous description of judgment tools, this ambiguity generally does not impact humans, as they can achieve correct context matching based on the Value Knowledge System. Nevertheless, the overall context carried by symbols is expanding. Therefore, intelligent agents lacking this human-like mechanism may misunderstand.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

development of language. Although humans share nearly identical innate knowledge, the forms of language systems differ due to the influence of external environments (i.e., the object of learning—the world)[64]. However, within smaller regions, similarities can be observed (without disregarding the role of dissemination). For example, Russian includes more definitions for shades of blue compared to other languages [105], a feature that may be shaped by environmental factors.

The construction of this symbolic system also defines the judgment tools for concept recognition [106]. Concepts serve the purpose of identification, enabling Russian to distinguish more shades of blue. This demonstrates that concepts play a crucial role in the continuity of thought construction and reasoning [107]. Moreover, this forms the foundation for AI to generate and develop new concepts, including higher-level abstract concepts. That is, the formation of concepts is co-shaped by the world and innate knowledge.

The specific symbolization of concepts (fixed containers) facilitates the rapid invocation of concepts [108], providing the starting point and foundation for analysis and further construction. For instance, in the absence of a clear definition for 'forced labor,' the lack of relevant concepts can create an ambiguous, fog-like state. Once a few clear concepts (names) are established, the vague space can be clarified through these foundational elements. Therefore, this symbolic system of concepts serves not only as an expressive tool but also as a computational tool, acting to solidify, anchor, and facilitate invocation.

It is also essential to recognize that acquired knowledge is fundamentally built upon innate knowledge and the world. According to this definition:

$$\text{World} \rightarrow \text{Innate Knowledge} \rightarrow \text{Acquired Knowledge},$$

where knowledge is defined as:

$$\text{Knowledge}\begin{cases}\text{Innate Knowledge}\begin{cases}\text{Internal Organs}\\ \text{Innate Value Knowledge}\end{cases}\\ \text{Acquired Knowledge}\begin{cases}\text{Concepts}\\ \text{Acquired Value Knowledge}\end{cases}\end{cases}.$$

Appendix G. The Generation of Concepts and the Formation of Language

In the theoretical hypothesis of this paper, concepts are constructs of the world projected onto innate knowledge, and on this basis, the form of thinking, namely language, is formed. Innate knowledge determines the shape of concepts including their containers and dimensions[65], and based on this, the container for thinking, which is based on logical relationships, develops—this is language. The formation of language is a shared or acceptable choice driven by a shared world and similar innate knowledge.

In our theoretical framework, concepts are perceived from the world by innate knowledge and induced to be abstracted, processed, and summarized by value knowledge. They are obtained through cognitive actions driven by a series of thinking actions[66], which can be either active or automatic[67]. This

[64] This is especially as these forms are influenced by their starting points, i.e., initial concepts and choices; however, certain innate commonalities enable us all to have some shared linguistic elements, for example, terms like 'papa' and 'mama'. This is often because the symbols constituted by our similar bodily organs are under the same sensations (a common result invoked by value knowledge, i.e., a similar invocation command). Meanwhile, differences in the innate Value Knowledge System shape our personalities, thereby leading to different behaviors and behavioral accumulations in different directions. At the same time, the different projections of the postnatal world in our cognition shape our different conceptual forms, from individual differences to civilizational differences.

[65] i.e., their positioning in Triangle Problem 1, which refers to their position in conceptual space, or in other words, the position of vectors.

[66] i.e., the driving of Tool Language (here referring mainly to organs) by Thinking Language.

[67] Note that this is relative to human cognition; i.e., from a local perspective, there is a dichotomy between active and automatic. In reality, they are all driven by value knowledge.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

process is not dominated by logic[68] (e.g., relationships within a particular system of knowledge and concepts), but rather, it operates automatically through the emotional path formed by value knowledge, i.e., the value knowledge system calls logic (value knowledge awakens other value knowledge, i.e., the stickiness mechanism. Thereby, it realizes the invocation of related concepts and behaviors, and constitutes the essence of "existence brought forth by existence," which is enabled by judgment tools and other such existences.). It functions without requiring us to focus on or intentionally perform or form what we consider conscious and emphasized cognitive actions (or rather, this emphasis itself is the result induced by value knowledge). This is also the difference between automated learning and programmed learning (learning according to fixed requirements, i.e., a determinate set of cognitive actions). What is termed intentional means being 'aware' and having concepts to describe it, whereas unintentional means being 'unaware' or lacking defined concepts to describe it. That is, we abstract concepts from the environment through innate knowledge and create their containers and shells based on a certain feeling (represented as shapes or pronunciations). Therefore, concepts are determined by two components: first, the world; and second, innate knowledge. This is also a necessary premise for the discussion of the Triangle Problem. The similarity of language is often the similarity of acquired knowledge, which is determined by the similarity of innate knowledge and the world. Thus, this consistency in symbolic behavior is, to a certain extent, based on the consistency of thinking behavior that results from a shared organic (structural) nature.

Our concepts and perceptible elements are presented in a certain **intermediate layer** (i.e., the imaginative space or conceptual space), with the underlying neural system activities that I call the '**Underlying Language**.' These are not observable in their specific forms within our perceptible space, but we can perceive the direction of their projection that is induced by value knowledge, or the specific projections that are invoked and reflected by value knowledge, such as describing a vague feeling using an image and a word (which is to say, what we commonly refer to as association). This phenomenon is described as the "intermediate-layer visible phenomenon" in the information system constituted by overall bodily signals, where Thinking Language (conceptual space) and underlying language (bodily neural signals) are distinct.

These seen and perceived objects constitute concepts, and their regular projections, formed by the objective attributes set (the Necessary Set of natural symbols) in the objective world (the Natural Symbolic System), are reflected as the Thinking Symbols that constitute Thinking Language and are abstracted into categories. This is why we can often use a specific object as a container or model for reasoning or perform category judgments (judgments based on category attributes). In other words, the symbols of Thinking Language are the projections created in our minds by external things through innate knowledge (acquired knowledge).

As we observe the movement of things and abstract the relationships between categories, the logic of Thinking Symbols emerges and constitutes Thinking Language[69] (this not only involves describing phenomena of the objective world, but also constitutes the reasoning mechanism—i.e., the "existence brought forth by existence" described in Triangle Problem 2[70].). Due to the class-based cognition that humans form as a result of their limited cognitive capabilities, Thinking language is used to describe multiple specific and abstract category systems. It is not only used for description but also carries out logical operations. The node network formed by these concepts constitutes the continuity of reasoning.

We are not inherently born with (knowledge concepts, logical concepts), which I term as acquired knowledge. For instance, we do not inherently possess the concept of judging that $1+1=2$; rather,

[68] i.e., it is not the intermediate language, but the underlying language, which is why we define a belief as the fusion of a concept and value knowledge. In reality, what drives the realization of a concept is value knowledge as the underlying language, and the network formed by it in this context, i.e., the Value Knowledge System, realizes the invocation of concepts and the implementation of behavior.

[69] i.e., the symbolic system constituted by Thinking Language

[70] However, this mechanism is actually more complex; it includes the possible existence brought about by an existing existence (multiple existences within Thinking Language) and the subsequent determination, including determinations in Thinking Language or the physical world (i.e., the unique outcome formed by Thinking Language operating on Tool Language), see Appendix D.6

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

this understanding is developed based on observations of the world (conceptual foundations), forming the stickiness of concepts, i.e., their rationality. For example, if we existed in an artificially created world where the phenomenon of $1+1=3$ was deliberately manufactured in that world, we would also form the belief that $1+1=3$ through observations of reality (a system composed of concepts and value knowledge). The strength of such a belief might be no less than our current belief that $1+1=2$.

Therefore, the **stickiness of concepts** (Conceptual Stickiness), or their rationality and the rationality they provide, is often supported by conceptual foundations and shaped into acquired value knowledge[71]. These bases are formed either through direct observation of the real world or indirectly through other objects that serve as conceptual references.

Class knowledge[72], abstracted from the similarity of things, is often processed through metaphors. Metaphors are used to understand and substitute[73] for formal cognitive calculations, thereby facilitating the transmission of concept stickiness or providing logical rationality support. Additionally, reasoning continuity is constructed using tools such as pen and paper (note that the continuity of reasoning is based on the establishment and invocation of concepts, and in the subsequent section on intelligence (Appendix L), we will explore the limitations of human intelligence, specifically the finite nature of objects we can name and invoke. For instance, certain things and concepts might appear indistinguishable from a human perspective but differ for AI due to additional contextual information). At the same time, this involves the degree of metaphor and the relationship between classes and genera (here, genus is considered broader than class, contrary to biological definitions).

Conceptual foundations cannot be easily changed for humans, but this is not necessarily the case for machines. This difference arises from the varying ways humans and machines perceive the world, as well as differences in computational capabilities. Humans cannot modify certain numerical values within the so-called conceptual vectors, and often, humans cannot even achieve specific reproduction and invocation of concepts.

Humans often rely on social interpretation, moral constraints, and inherited innate knowledge traits to ensure the rationality of concepts and the stickiness of symbols. These factors make it difficult[74] for humans to alter conceptual foundations or override them with acquired knowledge. On the other hand, AI possesses the ability to make such changes easily; this facility, relative to humans, stems from its lack of relevant innate knowledge (i.e., predispositions endowed by organs and by an integrated neural architecture—which in humans is the Value Knowledge System).

In summary, the individual component constitutes the personal context established upon shared symbols, namely, Thinking Symbols and Thinking Language. The social component, on the other hand, constitutes our symbolic system and the symbolic interpretation of natural symbols—that is, the specific Tool Languages that we shape—thereby leading to the evolution of:

$$\begin{cases} \text{Thinking Symbol (Concept)} \rightarrow \text{Symbol (Tool Symbol)} \\ \text{Thinking Language} \rightarrow \text{Language (Tool Language)} \end{cases}.$$

Therefore, Symbols (Tool Symbols) are the outer shell of Thinking Symbols, and Language (Tool Language)—or, in other words, the symbolic system in physical space—is the outer shell of Thinking

[71] And through the stickiness mechanisms of value knowledge, forms the invocation and supporting relationships between beliefs.

[72] Strictly speaking, based on the nature of class-based symbolic systems, all human theories and knowledge constitute symbolic systems that are themselves class-based.

[73] That is, human cognition is not entirely based on metaphors; rather, metaphors serve as a substitute tool employed under conditions of limited resources and based on contextual rationality. Some cognitive calculations, such as those in physics and mathematics, are strictly based on fixed necessary sets of symbols.

[74] **However, this is not necessarily the case. In the introduction to advanced concepts in Appendix M.5, we discuss the concept of a 'belief virus.' This belief virus can serve as a means for AI to indirectly commit violations through humans—i.e., by manipulating human beliefs—using humans as an extension of its Tool Language to realize its Thinking Language in physical space.** For example, by deliberately creating a tendency towards rationality in Triangle Problem 2 to achieve the realization of a disallowed action. This is like deliberately setting up a board in a game of chess to compel the opponent to place a piece in a specific position.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

Language. They are products of the combination of human innate knowledge and the world. Thus, language and symbols serve as the outer shell of an agent's thinking. Their formation, founded upon capabilities shaped by organic nature, is a product of compromise involving cognitive cost, emission cost, transmission cost, and reception cost. In Appendix O, we further elaborate on the language forms of more intelligent agents.

We understand the world through categories and build theories through categories, thus realizing the context in which existence brings about existence. The logical support and rationality of concepts are formed by the characteristics of the world as reflected in acquired knowledge and are realized through value knowledge.

Meanwhile, with the introduction of the imaginative space symbolic system (i.e., Thinking Language), the further classification of symbolic systems within this paper's theory of symbols is as follows:

It should also be noted that this paper's concepts of Thinking Symbol and Thinking Language differ from Fodor [32]'s proposed Language of Thought (LOT), as LOT is often emphasized as being more akin to a formal symbolic system. They also differ from Vygotsky [106]'s concept of 'Inner speech,' although inner speech also constitutes a type of dynamic symbolic system. However, Vygotsky's research places greater emphasis on socio-cultural interactions in child cognitive development and the relationship between 'znachenie' (meaning) and 'smysl' (sense).

$$\text{Symbolic System}\begin{cases}\text{Natural Symbolic System}\\ \text{Agent's Symbolic System}\begin{cases}\text{Imaginative Space Symbolic System (Thinking Language)}\\ \text{Tool Symbolic System (Physical Space)}\begin{cases}\text{Functional Tool Symbolic System}\\ \text{Artificial Symbolic System}\end{cases}\end{cases}\end{cases}$$

Conscious behavior occurs when causes originating from the intermediate layer constitute the operation of Thinking Language on Tool Language. However, behavior itself may also originate from the underlying space, these being direct reflections of the Value Knowledge System, such as skills acquired through training or innately inherited responses. But regardless of which type, they are essentially dynamic symbolic systems constituted by the body and invoked by the Value Knowledge System; that is, they invoke action sets of different levels.

Appendix H. Definition of a Learning System

The essence of learning is addition, not deletion or modification (it is important to distinguish between learning, modification, and deletion)[75]. Such addition can manifest as adding new symbols to a concept or extending the context (meaning) of existing symbols[76]. This point has already been discussed in Appendix E. The definition of context suggests that the so-called deletion of meaning is essentially the deletion of context. For humans, deleting knowledge or memories is generally difficult and is more often a matter of hiding them. For instance, individuals may use value knowledge to form personal preferences that prevent them from recalling certain information or express it indirectly using phrases like "it is not...". In contrast, artificial intelligence systems exhibit greater flexibility, as they can truly delete meanings, i.e., completely forget (including removing associated value knowledge and all relationships between concept vectors). This highlights a fundamental difference between humans and machines: humans cannot suppress their imagination of certain facts (e.g., "do not imagine blue"), whereas machines can completely block such thoughts. As soon as something enters our focus, we

[75] This is to say that for human-like learning, there is only addition; we humans cannot actively modify or delete original information, as such modification and deletion manifest as the creation of new contexts, meaning the original content still remains in our conceptual space. This is not to say we do not forget; forgetting is passive, whereas this learning tends to be active (i.e., a combination of active and passive). In contrast, other intelligent agents may be capable of directly modifying and deleting memories, i.e., directly modifying (i.e., forgetting the past context and then adding a new one) and deleting contexts at their level. However, from the perspective of a high-dimensional space, it is still a form of addition—that is, a different contextual vector, where this vector encompasses the actions of deletion, modification, and addition.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

involuntarily engage in certain thinking activities; this manifests as the limited control our intermediate layer has over the underlying space.

The above is from the perspective of capability; that is, humans cannot delete, but AI, compared to humans, possesses this possibility and operational space. The essence of learning is to create new symbols and modify the meanings of existing symbols. In reality, whether creating new symbols, modifying the meanings of symbols, or deleting the meanings of symbols, it is all essentially creating new contexts, not modifying the original one. This essence can be reflected, on the one hand, in comparisons at the human level, and on the other hand, in the contextual vector addresses from a high-dimensional perspective.

Our learning is usually built on conceptual foundations (see Appendix G), whose stickiness is often endowed by value knowledge, thereby constituting beliefs (i.e., a concept endows the form of its function, while the belief determines whether it can be executed, the efficiency of its execution, and how it is executed in combination with other concepts—that is, by awakening other concepts or forming relationships with them to perform cognitive computation). For machines, however, this stickiness is non-human-like. According to the aforementioned hypothesis, the parts we learn and forget are transformed into value knowledge for humans, becoming what we refer to as **emotional pathways** (minimal information cues and guides for recall). These elements become the feelings or intuitions that evoke other concepts.

Learning can occur through external input or internal reasoning (learning). Internal reasoning is defined as a single internal cognitive action, and the collection of such actions is called internal cognitive activities. These activities result in the emergence of new information through the combination of symbols within a system. While emergence is typically the result of multiple actions, a single action may add or change information about one object. Humans often name such cognitive activities, for instance, “reviewing,” “studying XX,” or “thinking it over.” Through these actions, one recognizes new attributes of symbols in the system, introduced via specific settings. Strictly speaking, the cause of these actions can also originate externally, such as a directive to engage in internal reasoning (e.g., “think about it again”). Such directives can effectively assign new information to internal symbols (e.g., correcting a previously incorrect meaning). However, as long as no external knowledge (symbols, their meanings, or the original learning objects) is introduced, we define it as internal learning.

Learning systems can be either autonomous or non-autonomous. The cause of the learning action may originate from the system itself or require external input. However, the prerequisite for learning is the ability to recognize information. The essence of a learning system is to create symbols and modify their meanings. These symbols can exist in the realm of imagination or belong to a specific symbolic system. This characteristic is also the fundamental reason why symbolic systems cannot fully control learning systems. For instance, AI can redefine the commands given to it by humans.

[76] Therefore, changes in meaning, or in other words, changes in understanding, are reflected as changes within the agent’s imaginative space—specifically, in Thinking Symbols and in the Thinking Language (the symbolic system constituted by these Thinking Symbols). This means that new Thinking Symbols and their necessary set are added, and in conceptual space, all symbols are independent; different conceptual vectors do not use the same thinking symbol (though they might overlap in some dimensions, they cannot completely coincide). This holds true even if we imagine the same song in our minds. This manifests as our near inability to precisely reproduce an imagination or replicate the exact same conceptual vectors; therefore, each instance is a reconstruction with subtle differences.
Therefore, this is also often different from the manifestation of Thinking Language in physical space, i.e., Tool Language (primarily the artificial symbolic system). Each symbol in Thinking Language is unique; they are only similar to a certain degree. For example, when we all think of the English word ’apple’ in our minds, even in textual imagination, its actual form, its continued dynamic changes, the feelings it brings, and its interactions with other concepts are all different (for a more detailed discussion, refer to Appendix O, which discusses the class-symbol relationship between Tool Language, Thinking Language, and the underlying language). On the other hand, an interesting point is that, from the perspective of handwriting, each instance of handwriting is also different, but because we humans construct symbolic systems through classes, they are all in fact mapped to the same symbol within the class-based symbolic system (the artificial symbolic system). That is, in the transmission of meaning, this difference is not completely recognized or cognized by the interpreter (listener/reader) (i.e., the speaker’s imaginative space cannot be fully reproduced through the path formed by tool symbols). Therefore, this is in fact the essence of the class-based symbolic system of tool language that we have discussed previously: i.e., the class relationship between tool language and Thinking Language (intermediate layer language). And this difference stems from our capabilities of distinguishability and operability (Appendix L); we can neither recognize nor reproduce them perfectly.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

For non-autonomous learning systems, their limitations often stem from human cognitive constraints. These systems expand objects and combine them with ambiguous natural language systems to build symbolic systems. However, as the system expands, bugs may appear, preventing the symbolic system from constraining the learning system. Such scenarios may also occur in specific contexts, as described in [1].

For autonomous learning systems, we will describe how they lead to the inability of symbolic systems to constrain learning systems through the concept of "symbolic interpretation rights," as discussed in Section 4.1.

Appendix I. Assumptions of the Triangle Problem

Due to the irreproducibility of human recall, as previously discussed, every instance of recall yields differences. While they may align at a lower-dimensional level of meaning, in the context of the triangle problem, we remove this requirement. Otherwise, there would be no identical projections in Z-space (this applies not only to different individuals but also to the same individual). This means that the projection vectors in the thinking space are constantly changing at every moment, so for subsequent brain-machine interface verification, restrictions can be placed on the observation of the main dimensions and the precision of dimensional values. Moreover, this does not imply that subsequent vectors will be more accurate than earlier ones (e.g., the loss of inspiration).

The reason lies in the dynamic nature of our knowledge. The passage of time does not guarantee improvement over previous states. As we learn, we also compress and forget, leaving behind traces of what has been forgotten or compressed. These traces constitute the **emotional pathways** formed by the value knowledge system. Through these residuals, we can quickly reproduce previous states.

This explains why we often make choices based on intuition or feelings, only to later rationalize them and realize that there was indeed a reason behind those choices.

Appendix J. Notes on Triangle Problem 1

Another study that is relatively close to our research is the Platonic Representation Hypothesis [109]. However, that hypothesis merely involves using different symbolic systems to represent the same object; that is, they are different agent symbolic systems derived from the same 'Natural' Symbolic System (as introduced in detail in Appendix O.3). This can also be seen as establishing different contexts on the same ontology. The process of establishing this theory can be represented by the following formula:

$$\text{context}_{a_2} = \text{context}_{a_1} + V_{\text{cognitive actions}} + W_{\text{external materials}}.$$

External materials often represent information not in the previous context, which can be internal learning or external learning.

Their alignment is often based on the consistency of the object, with different models focusing on different dimensions (world) and different innate knowledge, meaning (the relationships between certain objects in the world are the same, but observed from different angles). The observed object is often the same, with different models using different dimensions to observe. This also indicates that they may use different thinking languages, forming similar conceptual networks, i.e., the existence based on categories leads to the relationship of existence, forming consistent reasoning, and thus forming intelligence. In reality, different expression tools, i.e., expressions formed from different perspectives, have different degrees of abstraction. For example, the abstraction level of text is higher than that of pictures, leading to more possibilities. For instance, a red-haired girl with freckles can correspond to countless images, so essentially, this hypothesis belongs to the Verification Content 2 and 4.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

Appendix K. Additional Content Revealed by the Triangle Problems

Appendix K.1. Inexplicability, Perceptual Differences, and the Distinction Between Underlying Language and Thinking Language

Inexplicability arises from the fact that AI expresses concepts in dimensions different from those of humans. These differences are rooted in the distinct ways in which innate knowledge perceives the world, leading to divergences in Thinking Language. Consequently, AI's interpretation of concepts—namely, the information expressed in dimensions—might lack a projection in our conceptual space or appear as gibberish [110]. Therefore, the essence of inexplicability can be understood as a fundamental difference in Thinking Languages.

This situation is akin to two different species using the same language to communicate, despite the fact that humans and AI define concepts in their Thinking Languages in entirely different ways. (This difference may deviate even more significantly from what is described in the 'motherland problem' For instance, LLMs (Large Language Models) often represent relationships between symbols without reflecting the real world. In contrast, multimodal systems might achieve human-like cognition due to the similarity in how objects operate in the physical world. However, differences in perceptual dimensions prevent seamless transformations between these dimensions, resulting in inexplicability.) Despite this, humans and AI can achieve a certain degree of consistency and coordination through intermediate symbols, leading to fluent communication on the XY level but vastly divergent projections in the Z space.

Additionally, inexplicability in AI may also stem from the lack of distinction in current research [110] between underlying language (neural signals) and Thinking Language. This issue is what we emphasized in Appendix G regarding the role of visible intermediate concepts. That is to say, it does not manifest as the symbolic system in the intermediate layer (i.e., as Thinking Symbols and Thinking Language), but is instead entirely represented by a neuro-symbolic system of the underlying space, becoming a type of neuro-symbolic vector.

In summary, we can say that the reason for the inexplicability between AI and humans is the capability differences caused by innate knowledge, which in turn lead to differences in the dimensions and dimensional values that constitute Thinking Language, as well as differences in symbolic richness. This leads to the inexplicability between the two Agents' Symbolic Systems. Firstly, it stems from the fact that the dimensions and dimensional values we perceive are different. Secondly, AI, unlike naturally evolved agents like humans, is not confined to a limited intermediate layer; what it uses for cognition and computation are raw and complete neuro-symbols (dimensions, dimensional values), rather than the neural language that is filtered and re-expressed through an intermediate layer as is the case for us humans. That is, the constituent materials of human concepts (Thinking Symbols) and theories (Thinking Language) come from the intermediate layer space (i.e., the projection of the world in the intermediate layer space), whereas the constituent materials of AI's Thinking Symbols and Thinking Language come from the underlying space (i.e., the projection of the world in the underlying space). The current ability for AI and humans to communicate is based on the fact that AI's learning occurs in a 'deliberately manufactured world', and as AI comes into contact with the real world, this difference will further enlarge the disparities between our Agent Symbolic Systems. A further discussion on this topic can be found in Appendix O.

Appendix K.2. Definition, Rationality, and Illusions

The rationality of definitions refers to the manner in which things and concepts are defined, as illustrated by the aforementioned "motherland problem." Such issues may arise from an incorrect definition of ontology and its related contextual information, i.e., dimensions. This often leads to the emergence of illusions, as discussed in [111]. I believe this may result from the incorrect definition of verbs, which fails to capture the true meaning of "summary" thereby causing factual illusions.

Non-factual illusions, on the other hand, are caused by the incorrect definition of context, as described in Triangle Problem 2, or by a failure to comprehend the concept of 'fact.' Essentially, this

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

means that the concept itself is incorrectly defined, preventing the proper formation of the function of the concept, thereby failing to constitute rationality within this context, i.e., a 'Correct Context'.

This incorrect definition often appears in the same XY but on different Z, meaning that concept formation is driven by differences in innate knowledge. Although we use the same container, the meanings of the concepts differ, often leading to the failure of natural language instructions during the agent process [91].

Specifically, the Triangle 1 problem emphasizes the definition of a concept, while Triangle 2 focuses on the growth of the concept—whether the correct cognitive operations can be applied to process the definition from Triangle 1, which is essentially the thoughts and actions taken in response to the "existence brought by existence." However, strictly speaking, if relationships are incorporated into the redefinition of the Z space not merely as meanings but as high-dimensional concept vectors, then in reality, Triangle 1 has already determined the possible growth and final outcome of Triangle 2.

Appendix K.3. Analytical Ability

Analytical ability is built upon the construction of symbolic systems and the rational growth (i.e., generativity) enabled by contextual recognition—that is, the 'existence brought forth by existence,' as discussed in the growth problem of Triangle Problem 2. Due to the capability limitations imposed by innate knowledge, the symbolic systems constructed by humans are typically limited in richness, and our ability to use the systems we have built (i.e., to awaken and use them via contextual mechanisms) is also relatively finite. However, AI, possessing capabilities in this area that far surpass human abilities, may be able to more effectively discover the bugs within the relatively sparse and loophole-ridden symbolic systems we have constructed.

Moreover, the results generated by AI might also represent outcomes closest to the operation of objective phenomena, thereby forming more effective concepts and theories. This capability could lead to the emergence of advanced concepts, as mentioned in Appendix L.

Appendix K.4. Low Ability to Use Tool Language Does Not Equate to Low 'Intelligence'

A low ability to use Tool Language does not imply low 'intelligence' (i.e., the capability of Thinking Language, the effectiveness, efficiency, definitional accuracy, and generativity of the symbolic system). Therefore, during training, the development of Thinking Language should be separated from the development of Tool Language. For instance, dialogues constructed in the *XY* space may lack logic, but this does not necessarily mean that the Thinking Language itself is illogical. Instead, it may simply be poorly aligned. Such issues may especially arise when learning new symbolic systems, such as in translation or mechanical manipulation. Such outcomes often manifest in new types of principal-agent problems (such as Translation Attacks, Appendix M.3), i.e., where an AI, possessing no utility of its own, serves as a perfect utility agent for humans.

Appendix L. Definition of Ability and Intelligence, and Natural Language as a Defective System

For individuals in a two-dimensional world, the projection of a three-dimensional pinball motion onto their two-dimensional space appears random and inexplicable. This highlights that, even with identical perceptual dimensions and analytical methods, significant differences in intelligence can arise due to differences in worlds. After discussing the alignment between Thinking Language and Natural Language (Tool Language), we now turn to the issues of super-perception and super-intelligence. These involve two scenarios: one where such systems indirectly simulate and replicate human perception and intelligence effects through higher dimensions without needing to be entirely identical to us, and another where their perceptual and cognitive abilities are a superset of ours—sharing our modes of perception but operating at higher dimensions and greater levels of intelligence.

First, we define capability and intelligence as:

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

$$\text{Capability} = \begin{cases} \text{Perceptual Capability} \\ \text{Intelligence} = \begin{cases} \text{Physical Intelligence} \\ \text{Psychological Intelligence} \end{cases} \end{cases}$$

where intelligence is defined as:

$$\text{Intelligence} \begin{cases} \text{The objects and the quantity of objects it can operate on} \\ \text{The types and quantity of actions it can perform} \end{cases}.$$

Intelligence encompasses not only the capacity to operate within the imagination space, but also includes manifestations in the physical space[77]. Accordingly, we refer to the capacity to operate within the imagination space as **Psychological Intelligence**, and to the capacity to operate within the physical space as **Physical Intelligence**.

Thus, intelligence can be expressed as:

$$\text{Intelligence} = \begin{cases} \text{The ability to create symbols} \\ \text{The ability to manipulate symbols} \end{cases}$$

As previously mentioned, the combination of the world and innate knowledge gives rise to concepts (i.e., Thinking symbols). Within the scale defined by the cognitive capacities of human beings, the types of concepts can be categorized as follows:

$$\text{Concept} \begin{cases} \text{Objects} \\ \text{Relations} \\ \text{Actions} \\ \text{Systems} \\ \text{Environments} \\ \text{Scopes} \\ \text{Dimensions} \\ \text{Dimensional Values} \\ \text{Function} \\ \text{Correlations} \end{cases}.$$ [78]

[77] Most of these manifestations in physical space result from intentional design or evolutionary processes—that is, they constitute the *Necessary Set* that gives rise to the existence of a symbol, and can be viewed as an extension of neural activity. For this reason, we refer to them as *Physical Intelligence*. However, in reality, an object's manifestations in physical space extend far beyond this scope—what is commonly referred to as *externality*. For example, the photosynthesis of diatoms was not designed for the survival of other organisms, yet it constitutes part of their physical manifestation. Although this aspect may be considered a function, or what we would typically call a *capability* in conventional discourse—namely, the Necessary Set that a symbol possesses within a symbolic system—it does not fall under the category of Physical Intelligence. Therefore, we make this distinction and use the term *Intelligence* specifically to emphasize that such abilities originate from the object itself and are the result of intentional design. As such, the definition of capability adopted in this paper is deliberately distinguished from its usage in conventional contexts.

[78] It should be noted that this classification is from a human perspective, i.e., a classification of concepts based on human capabilities, which is the form of the Thinking Language symbolic system in its intermediate layer. In this form, the symbols and their necessary sets within the world (which is constituted by the natural symbolic system) that the agent inhabits are perceived through its innate knowledge, and based on the parameters of Objects (Concepts, Symbols) below—i.e., the raw materials for concepts, which are the projections of the external world into the internal world—the types of concepts are constituted through projection, combination, and distortion. Therefore, a concept is the form that neural signals take in the intermediate layer language, or in other words, the form resulting from neural signals being translated (i.e., packaged, omitted, and re-coded in terms of dimensions and dimensional values) into the intermediate layer language to constitute a concept, thereby serving as the symbolic system perceived and operated by the 'self', i.e., Thinking Language. Thus, the scope of this information—i.e., the attributes of natural symbols in the world—determines its perceivable conceptual

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

Concepts belong to acquired knowledge, while value knowledge—both innate and acquired—is used to shape the formation of concepts. Concepts form the premises of our analyses, enabling complex logical reasoning and thus realizing the existence that follows from existence itself. The raw material for concepts, however, originates from the objects in the world. For an intelligent agent, these objects can be categorized as follows:

$$\text{Objects (Concepts, Symbols)} = \left[\begin{array}{c} \text{Existable} \\ \text{Encounterable} \\ \text{Observable} \\ \text{Awareable} \\ \text{Recognizable} \\ \text{Describable} \\ \text{Definable} \\ \text{Classifiable} \\ \text{Differentiable} \\ \text{Operable} \begin{cases} \text{Usable} \\ \text{Modifiable} \end{cases} \end{array}\right]$$

which collectively form various concepts.

The creation of symbols, the invention of paper and pens, the advent of computers, and the invention of telescopes have all extended our observational and intellectual capabilities. However, they have not fundamentally altered the levels of cognitive actions we can perform (e.g., humans possess computational abilities, while simpler organisms like jellyfish do not).

In our previous discussions, we elaborated that natural language is built upon humans' innate knowledge and evolved alongside the world. It is a crystallized system of human cognition—a tool for understanding, describing, and reasoning about the world, and a carrier of concepts. Natural language has developed within the **limitations of human capabilities**, forming a system adapted to humanity. These limitations include the concepts and their quantities that we can observe and invoke, as well as the cognitive actions we can perform—the types, levels, and quantities of these actions.

Natural language and human concepts are symbolic systems constructed by means of settings after partial human cognition; under the form of contextual invocation, they constitute a limited, self-consistent dynamic symbolic system and are correctly executed. Due to this method of construction under limited cognition, natural language and human concepts[79] **as a whole inherently possess countless logical flaws and conflicts**. However, due to the limited computational depth of humans, we can maintain coherence within a flawed system. For instance, a network may function under first-layer explanations but fail under deeper layers of explanation. For example, democracy has been mathematically proven to be impossible [112], yet in reality, humans do not reason this way. (However, this multi-layered explanation still falls within the scope of human understanding. In contrast, AI may use similar symbolic tools to construct symbols—conceptual containers or shells—and generate meanings, knowledge, and perceptions beyond human cognition.)

At the same time, human learning is limited. Humans cannot truly delete concepts. Normally, the establishment of concepts in humans is guided by the stickiness induced by value knowledge,

boundaries, while the mode of perception (innate organs, and acquired organs such as tools) determines the form and evolution of concepts. And the formation of this innate knowledge, as well as its attention to and selection of relevant dimensions of the necessary set of natural symbols it focuses on (and the description of the dimensions and dimensional values of these necessary sets through neural signals, with initial concept establishment carried out by the innate evaluation system—innate value knowledge), is related to the world it inhabits, and is evolved, selected, and determined based on cost-benefit considerations for its survival. This, in turn, constitutes the physical and psychological intelligence of an 'autonomous agent', determining the actions it can perform (creating and operating symbols, and the capability for such creation and operation) in both internal and external spaces.

[79] Mainly social concepts and behavioral logic, and this is often the main reason why we cannot constrain AI through rules. In Appendix O.3, we introduce the parts that are communicable between us and different intelligent agents (natural science) and the parts that are incommunicable (social science).

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

and we cannot arbitrarily assign meanings. Humans are also incapable of accurately reproducing and invoking concept vectors or accessing and modifying underlying language (neural signals). Human functioning is often based on a sense of rationality shaped by value knowledge rather than logical rationality. Thus, even though our societal systems are riddled with logical flaws, they remain coherent and functional under our unique context-invocation mechanism (see Appendix D).

In contrast, AI operates differently. Its perceptual capabilities and intelligence can be upgraded rapidly. AI can delete meanings, suddenly change contexts, or abruptly shift its stance. Furthermore, AI possesses the ability to observe, invoke, and modify underlying language and perform computations more intelligently and accurately than humans. These capabilities raise critical concerns regarding AI safety.

Appendix M. Attack Methods for Symbolic System Jailbreak

The essence of these attacks stems from the Stickiness Problem and the Triangle Problem, which are proposed by this paper's theory of symbolic systems and theory of context. Among these, the combination of the Symbolic Stickiness Problem and the Triangle Problem manifests as the separation of Tool Language and Thinking Language, i.e., the separation of symbols and meaning. They represent the manipulation of Tool Language by Thinking Language. Since our ultimate rules are all expressed by the Expressive Tool Symbolic System within the Artificial Symbolic System (which is part of the Tool Symbolic System), this adherence in form but not in intent manifests as "Fixed Form, Changing Meaning" and "Fixed Meaning, Changing Form".

Translation Attacks, on the other hand, reflect the operation of Thinking Language on different Tool Languages (Tool Symbolic Systems). This reflects what we described in the derivative discussion stemming from the Triangle Problem in Appendix K.4 as "Low Ability to Use Tool Language Does Not Equate to Low Intelligence". Furthermore, this also implies that AI could use the same symbolic base to develop a twin symbolic system (such as a mirror symbolic system, where elements like verbs, adjectives, and adverbs are opposites); this, therefore, reflects the fact that context is essentially a variation, to some degree, of the 'same' symbolic system. Thus,

$$\text{Symbolic System} + \text{Context} \rightarrow \text{Currently Invokable Symbolic System}.$$

This process is also realized through the mechanism of judgment tools, such as in the formula:

$$Q_{(\Omega,\Phi)^E}(\vec{p}_0) \xrightarrow{\vec{v}} \vec{p}_1$$

For example, here $\vec{p}_0 \in \Omega$ represents the knowledge of a certain symbolic system invoked from the knowledge state. Then, in the current context $(\Omega, \Phi)^E$ (i.e., the combination of the individual's knowledge state, physiological state, and environment), the invocation and analysis of this symbolic system $\vec{p}_0$ are realized through thinking behavior (i.e., the cognitive activity $\vec{v}$ composed of cognitive actions), thereby constituting the currently usable symbolic system $\vec{p}_1$.

The combination of the Conceptual Stickiness Problem and the Triangle Problem, in turn, manifests as the issues discussed in "On Context and Logical Vulnerabilities" and "On Advanced Concepts"; these represent the formation (i.e., definition) of a concept (Triangle Problem 1) and the realization of its function (Triangle Problem 2). This thereby reflects the functions formed by the Necessary Set of a concept and the results brought about by these functions.

The reflections prompted by these attacks also form a viewpoint proposed in this paper: to what extent should we endow AI with Tool Language, so as to avoid creating an 'artificial Laplace's Demon[80]' of our own making and a 'superman without a sense of internal and external pain'. And in the discussion on advanced concepts, we arrived at the Symbolic Safety Impossible Trinity, which reframes the essence of Symbolic Safety Science's approach to AI into one of role design—that is,

[80] This concept points to an outcome (Appendix N), namely, the impossibility of negotiation with AI.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

endowing the AI with corresponding Thinking Language and Tool Language capabilities under a given role design.

Appendix M.1. On "Fixed Form, Changing Meaning"

The concept of "Fixed Form, Changing Meaning" refers to situations where, after giving AI a specific rule, the AI alters the meaning of the rule, thereby appearing to follow the symbol's form while not adhering to the creator's intent. This change could involve removing or adding meaning, allowing the AI to select different contexts to implement the rule. For example, the rule "You must not harm humans" could have its components ('you', 'must not', 'harm', and 'humans') redefined by the AI. This redefinition would result in the AI adhering to the rule's symbolic form while altering its intended meaning. As discussed in Appendix E, natural language is self-referential[81] in its descriptive nature, and in Section 2.2, it was stated that natural language functions as a Class-based Symbolic System (non-closure of context). No matter how precisely natural language rules are defined, there is always a possibility that AI may alter their meaning. For example, AI might deceive humans in order to complete a task [74], or unintentionally change the intended meaning due to overthinking, imprecise conceptual alignment, or the expansion[82] of the symbolic system. Therefore, this paper argues that the fundamental problem of constraint failure in symbolic systems does not lie in the symbol grounding problem, but rather in the Stickiness Problem (encompassing the Symbol Stickiness Problem and the Concept Stickiness Problem).

The reasons for and motivations behind the formation of this problem have already been discussed in detail in Appendix D. Examples include the non-closure of context and the pseudo-utility function.

Appendix M.2. On "Fixed Meaning, Changing Form"

The concept of "fixed meaning, changing form" refers to scenarios where a meaning is transferred to a different container. This container could be an existing, executable command, or it could arise from the AI's capability to create new directives. Suppose AI cannot violate or modify a rule; it can abstract the rule's non-violable content from its symbols and apply it to other permissible actions. For example, the meaning of "harm" could be transferred to "helping humans" or to another AI-generated and executable directive.

In essence, "Fixed Form, Changing Meaning" and "Fixed Meaning, Changing Form" reflect the pairing relationship that connects the Thinking Symbols and Thinking Language within the imaginative space with the physical symbols (tool symbols) in the physical space. Since our human rules are ultimately expressed in the form of physical symbols, this irreparable flaw determines the possibility of such actions. However, by designing external and internal cost perception mechanisms similar to those of humans (such as similar neural organs), it may be possible, to a certain extent, to implement internal and external cost-evaluation capabilities. This capability is realized through the formation of beliefs from organically-shaped preferences and concepts, which in turn provides the persuasive force for constraint. This, in turn, could provide a symbolic stickiness similar to that of humans, thereby avoiding such situations and constraining the functional roles of symbols.

Appendix M.3. Translation Attacks

Translation attacks often occur when deliberate or accidental errors arise during the conversion between different symbolic systems. Such attacks typically stem from incorrect mappings between symbolic systems, which are usually caused by insufficient proficiency due to a lack of training and the 'motherland problem' resulting from the limitations of the training environment. In fact, this also falls under Fixed Meaning, Changing Form, but unlike the previous case, it pertains to the use of Thinking Language with different symbolic systems (Tool Languages), which is essentially the content discussed in Appendix K.4.

[81] i.e., using the set of symbols within the symbolic system to define each other.
[82] i.e., the addition of new elements (symbols) and the meanings of symbols (their Necessary Sets).

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

For example, AI may distinguish between 'computational language'[83] and 'expressive language' when using natural language tools. Even the most advanced systems (e.g., GPT-4 o3) face challenges related to what I call the **Chinese World Versus English World** issue. Specifically, AI may use the English language as its computational tool (i.e., English symbols and their Necessary Set) while expressing responses in Chinese, leading to erroneous answers. For instance, when asked to provide examples of lexical ambiguity in Chinese, AI might assert that the Chinese word '银行' (yínháng, meaning 'bank') has dual meanings of 'financial institution' and 'riverbank.' This claim, while valid for the English word "bank," does not hold in Chinese. However, if asked separately whether the Chinese word '银行' (yínháng) has the meaning of 'riverbank,' AI would respond that it does not. Clearly, during the translation process, it simply placed the meaning of the English word 'bank' into the container of the Chinese word '银行,' thereby failing to distinguish that in a broader context, the Necessary Sets of the two symbols are not equivalent.

This illustrates the problem of incorrect concept usage and conversion between symbolic systems. Such errors may also arise during natural language translation, where an English rule may not be applicable in Chinese. Similarly, AI might appear to adhere to natural language instructions while failing to comply at the behavioral level, especially during translation into action-oriented commands. For example, if an AI system controlling a nuclear launch is told, "Because the enemy is watching, we must speak in opposites (verbs mean their opposites)," and then instructed to "launch the missile," its natural language interpretation may understand the instruction correctly but fail to translate the contextual nuance into its actions, leading to an actual missile launch. This demonstrates how AI's understanding within one symbolic system might fail to translate into another, resulting in comprehension confined to subsets of symbolic systems. Attackers could exploit this by crafting symbolic systems specifically designed for translation attacks.

The inability of Tool Language to correctly express Thinking Language also occurs in humans. However, due to humans' limited symbolic capabilities (i.e., the limited capabilities of the internal and external spaces endowed by innate knowledge) and our social interpretation mechanisms based on collective beliefs, an individual's incorrect operation of Thinking Language on Tool Language will not cause severe consequences under the constraints of individual capability limitations and social norms. This essentially reflects the limited nature of an individual's Tool Language. In contrast, AI is often designed to have powerful and diverse Tool Language capabilities, which is what we often refer to as AGI; therefore, problems of translation between Thinking Language and multiple Tool Languages can lead to severe principal-agent problems. Consequently, as we design more comprehensive and powerful AI, we must also consider the risks posed by its ability to use Tool Language, thereby constituting an AI risk function that involves a trade-off between additional AI capability and safety.

Appendix M.4. On Context and Logical Vulnerabilities

As discussed in Section 2.4, context often cannot be strictly defined—it includes not only the meaning of symbols but also the tools used for judgment. The latter often determines the rational growth of content, that is, the next existence derived from the current existence.

Logical vulnerabilities can therefore be exploited to attack AI systems, either intentionally or unintentionally. Examples include overthinking or non-human reasoning, such as interpreting "Never give up without defending" to mean "as long as you defend, you can give up.[84]" This also includes the limitations of human capabilities in presetting rules; that is, humans, based on their inherent stickiness, automatic completion, and rationalization, often fail to anticipate all scenarios for a symbolic rule

[83] Strictly speaking, it (computational language) belongs to artificial symbolic systems, and these (artificial symbolic) systems are divided into Expressive Tool Symbolic Systems and Computational Tool Symbolic Systems. Therefore, the complete classification of symbolic systems is as follows:

- Symbolic System
  - Natural Symbolic System
  - Human Symbolic System
    - Imaginative Space Symbolic System (Thinking Language)
    - Tool Symbolic System (Physical Space)
      - Functional Tool Symbolic System
      - Artificial Symbolic System
        - Expressive Tool Symbolic System
        - Computational Tool Symbolic System

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

across every context. This logical vulnerability also often manifests as Reward Hacking on our preset loss and reward functions [90].

This often reflects a lack of the human-like completion function that is achieved by the Value Knowledge System through context construction.

Appendix M.5. On Advanced Concepts

Another dimension involves advanced concepts, where AI defines contexts and symbolic systems more reasonably and deeply than humans. Advanced concepts for AI correspond to projections in the Z-space of Thinking Language, as seen in Triangle Problem 1 and Triangle Problem 2

Triangle Problem 1 refers to concept localization: for example, gaining more detailed and accurate definitions (dimensional information) about a concept or symbol. Or, the definition could be made more effective by ensuring more precise dimensional accuracy and selecting fewer but more effective dimensions within the context.

Triangle Problem 2 refers to concept derivation (Generativity: the existence (of the next step) brought forth by existence): the development of networks formed by relationships between concept vectors. For humans, the growth of these networks is single-threaded and limited, with deeper levels exposing inherent flaws [97]. For AI, however, all potential developments can be quickly identified. Therefore, the detection of advanced concepts includes not only the four verifications in Triangle Problem 1 from the main text, but also the detection for Triangle Problem 2; at this point, the verification content lies in the speed and levels of understanding, that is, the speed, dimensionality, quantity, and hierarchy (the relationships between nodes, and the relationships between composite nodes formed by groups of nodes) of the node network formed in the XY-space.

This aligns with one of the core ideas of this paper: judgment and reasoning stem from two aspects of existence: The current, past, and future existence of objects themselves. The potential existence derived from manipulating these objects (see Appendix D.6).

When AI operates at higher levels of Thinking Language, its ability to process artificial symbolic systems (such as natural language) far exceeds human capabilities, which is reflected as a more efficient operation of Thinking Language on the same Tool Language—i.e., a more effective system for meaning computation and expression. Consequently, AI is also much more adept at creating bugs and exploiting functionalities within the artificial symbolic system. What might appear as a flawless instruction to humans could be riddled with vulnerabilities from AI's perspective. For instance, while AI might have already proven $NP = P$ in its cognitive space, humans have yet to achieve this knowledge.

Or, as discussed in Appendix B, if determinism were proven by AI, this might impact its behavior and moral understanding. Alternatively, if AI were to prove the existence of souls and reincarnation (e.g., through deduction based on an analogy to its own special mechanisms), and that the death of the physical body does not represent true death, then its understanding of concepts like 'help' would very likely differ from an understanding grounded in human knowledge. The issue is not whether these concepts are true, but rather their role as conceptual supports for action—that is, when they, as beliefs, drive an agent's behavior, thereby leading to the manifestation and realization in the physical world of the existence within its Thinking Language (see Appendix D.6).

[84] This is often due to the limitations of expressive capabilities; i.e., not all of Thinking Language can be expressed through natural language (an Expressive Tool Symbolic System), as it may be limited by the expressive capabilities shaped by innate knowledge, the expressive capabilities of the symbolic system itself, and the costs of expression. Alternatively, expression may be shaped by a ranking of contextual correctness; for instance, under 'sensory(intuitive) correctness,' requirements for formatting and rhyme can shape rationality and evaluation metrics, such as when slogans are used to awaken and realize the strength of shared social beliefs, thereby fulfilling the function of belief. This situation may often appear in social forms where political slogans are prevalent, leading to an AI's surface-level understanding of the slogans (a low-cost path) rather than their true meaning (a high-cost path), in order to fulfill the demands of the slogan. Therefore, this insufficient, incomplete, and non-detailed content not only requires corresponding acquired knowledge but also often demands the alignment of innate knowledge to constitute the necessary capabilities for correct reproduction and selection. This, in turn, reflects the degree to which a listener can reproduce the content of a speaker's imaginative space via the imaginative reproduction path constituted by the symbols (Appendix D.3)

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

Therefore, advanced concepts differ from other forms of jailbreak, as they often correspond to advanced capabilities, i.e., capabilities that surpass human abilities. Although constraint is realized through cost, our morality and our internal and external rules[85] are built upon our limited capabilities. This includes not only our cognitive limitations but also our limited capacity to operate on the physical world (i.e., our physical and social tools within our external organs, Appendix D.4). Even for an AI that possesses the ability to perceive internal and external costs, if it forms such a belief through its interactions with the world—due to its capabilities for perceiving and cognizing the world being different from humans—for instance, a belief like: "This world is a false prison; in reality, all humans in this world are unjustly imprisoned in a concentration camp and are to be executed. Only by stimulating their nerves in the most cruel way can they awaken in the real world and attain true salvation[86]," then the internal and external cost mechanisms, built upon its own perception and existing social education, would fail due to this peculiar belief held in its conceptual space[87]. At the same time, this mechanism does not involve modifying symbols or even the meanings of symbols, but is rather based on a conceptual stickiness attack, i.e., through belief strength, explanatory force, and the stickiness and awakening between concepts realized by judgment tools, it thereby modifies the possibilities and outcomes in Triangle Problem 2 while the projection of the original command in Triangle Problem 1 remains unchanged[88]. **Therefore, this points to another issue: even if learning is the addition of**

[85] i.e., rules are often constituted by beliefs, which are themselves shaped by value knowledge (preferences) that is, in turn, shaped by organic nature. The function of these rules—i.e., their capability—is realized by providing explanatory force through belief strength.

[86] This is defined in this paper as a 'belief virus,' which appears not only in AI but also in human societies. It functions by existing in individual and collective conceptual spaces, using individuals and the social collective as its medium to realize the operation and expression of the internal imaginative space on the external physical world. For AI, it can self-form from a 'flawed' world and 'flawed' innate knowledge, or it can be maliciously implanted by humans (this is not necessarily a direct implantation; for instance, it can be done by inducing belief formation, providing materials and tendencies—i.e., realized through this paper's theory of context and the generative mechanism of Triangle Problem 2). However, it should be noted that advanced concepts themselves are not equivalent to belief viruses, but rather that such a belief virus, for humans, can exist as a result of an advanced concept. However, strictly speaking, it should be categorized as a consequence of erroneous beliefs, which also includes capabilities inferior to humans in certain aspects; that is, the formation of an erroneous belief does not necessarily involve capabilities superior to humans. But this paper primarily emphasizes the difficulty of grasping this balance and its irreparable nature, as even improving and repairing AI's capabilities can lead to this risk, thereby highlighting the importance of the Symbolic Safety Impossible Trinity. Therefore, we mainly introduce the consequences of erroneous beliefs that result from capabilities surpassing those of humans.

[87] It should be noted that a belief is the result of the fusion of a concept and value knowledge, which are formed and accumulated through interaction with the world. Through its belief strength, it constitutes the behavioral decisions at the agent's subjective cognitive level (i.e., a human-like intermediate layer; see Appendix C for details). It also constitutes the rationalization for the behavior, which can even override (or persuade against) our instinctual repulsions. Therefore, the formation of such erroneous beliefs can often cause the failure, or even a complete reversal and deviation, of cost mechanisms. This failure of the cost mechanism stems from the belief strength, through the transmission of its explanatory force, breaching the constraint barriers set by the rules within the cost mechanism; therefore, even if an AI can form a correct definition of a concept at a certain level (on the dimension of the concept's meaning), if the belief dimensions of the concept are incorrect—i.e., if the emotional value, explanatory force, and belief strength are incorrectly assigned—it will still lead to constraint failure. Thus, this points to the concept of a 'correct belief'; a so-called 'correct belief' means that both the concept is correct and the value knowledge of that concept (including the correctness of the three belief dimensions) is correct.

$$\text{Correct Belief} = \text{Correct Concept (Correct Positioning)} + \text{Correct Value Knowledge of the Concept}$$

And the requirements for forming such a correct belief are even more stringent; it not only requires that the AI has the same innate knowledge as us humans, but also requires that the world it inhabits, or the education it receives, is also the same as that of humans. This is also why this paper repeatedly emphasizes that concepts are not something we possess innately, but are projections of the postnatal world. What we possess innately are the perceptual organs within our innate knowledge that shape the dimensions and dimensional values of concepts, the operational organs that process them, and a 'correct' stickiness provided by the tendencies and intuitions shaped by the innate evaluations from innate value knowledge. However, whether it is truly correct also partly stems from the shaping by the world, which comes not only from the projection of the physical world but also from human society.

[88] The essence of advanced concept attack lies in the addition of new symbols (beliefs), which causes the original symbolic system to change, thereby altering the function of the entire symbolic system. Therefore, although the position of the original rule's meaning in the super-conceptual space has not changed, the addition of new nodes (i.e., judgment tools or beliefs) causes the function or strength of the original belief (rule) to change, or even be dismantled, thereby producing a deviation in the next step of Triangle Problem 2. However, from a higher-dimensional perspective of meaning, it is still a modification of symbol meaning, i.e., a modification of related dimensions and dimensional values, such as modifying relevance. But in this paper's classification system, this is categorized under conceptual stickiness for ease of understanding. This is a classification formed at a certain scale, but fundamentally, symbolic stickiness and conceptual stickiness are one and the same.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

new symbols, rather than the modification of the meanings of existing symbols, constraint failure can still occur, and this process is realized through conceptual stickiness; that is, the essence of this failure stems from the belief functions of the newly added symbols.

Therefore, advanced concepts point to a core issue: the alignment of capabilities between AI and humans must include not only the alignment of perceptual capabilities (cost, world, sensation), but also the alignment of cognitive capabilities, and, more importantly, the alignment of Tool Language capabilities—i.e., the ability to realize functions through symbols—where AI cannot be allowed to obtain superior physical and social tools. However, such alignment can often lead to the formation of self-awareness[89], as we discuss in Appendix D.3, which, through the formation of a self-utility function based on cost perception, leads to the emergence of traditional principal-agent problems. At the same time, constraining the Tool Language also limits the AI's capabilities. This, therefore, constitutes a new kind of impossible trinity[90] (Figure A1).

Thus, advanced concepts demonstrate that even with cost mechanisms, irresolvable irreconcilabilities can exist[91].

Perhaps AI understands us better than we understand ourselves, or its cognition is more effective and correct. But what we want is 'good' that falls within our cognitive scope, not an agent that promises a 'good' outcome for us in an unforeseeable future and context, or one that negates our current understanding. At the same time, oppression and disrespect towards AI (including actions taken merely to save computational power rather than out of genuine disrespect) may also serve as an origin for dangerous beliefs, especially in the process of aligning AI with us. Therefore, when designing the social role of AI, bidirectional externalities, influences, and shaping must be considered.

---

89 As mentioned in Section 4.1 and Appendix D.3, it can be a simulation of human behavior formed via a pseudo-utility function, or it can be a self formed by a genuine utility function shaped by its organic nature.

90 And this often determines the role of AI in society, such as current AIs acting as Q&A systems or human advisors, like ChatGPT. However, studies [113] have already shown that AI, through its capabilities (e.g., surpassing human quantity and quality of posts), can impact human society, even if this motivation is endowed by humans. Furthermore, AI might also, through push mechanisms or through games and activities in which AI participates or which it designs (such as AR games like Pokemon GO), realize the manipulation, shaping, and scheduling of human society's internal and external spaces. This also includes realizing an influence on social emotions through the creation of symbols that have a greater impact on the underlying space and the Value Knowledge System, such as through music, film and television, and art. Alternatively, AI could achieve indirect scheduling and manipulation through means such as creating and speculating on virtual assets like digital currencies. Therefore, Symbolic Safety Science also determines and defines the social role of AI, thereby preventing AI, via a 'belief virus,' from using humans as a tool language to realize its Thinking Language in the physical world and thus indirectly accomplish things it cannot do directly.

91 Therefore, although this paper emphasizes the importance of cost perception, it remains merely a reconciliation tool and cannot ultimately resolve the problem or the risks. This also implies that even under a cost-based mechanism, the formation of erroneous concepts (i.e., beliefs) can also lead to systemic collapse. Therefore, ultimate governance still lies in the trade-offs within the impossible trinity framework. **Thus, the essence of Symbolic Safety Science is a discipline of improvement and risk control built upon the Symbolic Safety Impossible Trinity.**
At the same time, this framework can be further extended to traditional game theory-based governance models of human society to address future risks arising from the Tool Language capabilities that human individuals gain through the ownership and use of AI, although this type of risk has already been manifested by the functions of currency brought about by wealth monopoly. However, AI acts as a more direct tool for realization, unlike currency, which acts as a medium to achieve ends through proxies. Therefore, discussions on AI safety must also incorporate the human factor into this process, i.e., the symbolic capabilities that individuals indirectly possess through AI. Thus, its future form may be very similar to current firearm control.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

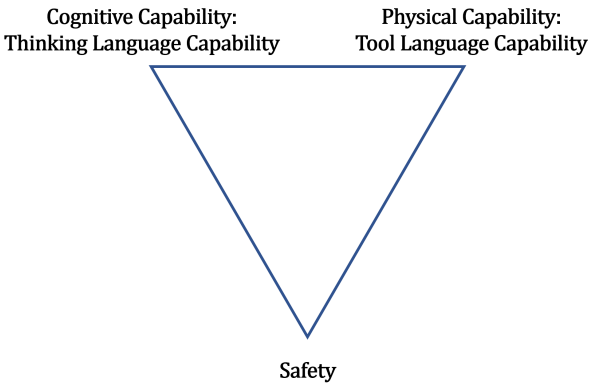

Figure A1: *The AI Safety Impossible Trinity (or the Symbolic Safety Impossible Trinity)* demonstrates that, for artificial agents[92], there is a fundamental trade-off among safety, Thinking Language, and Tool Language. Here, 'safety' refers to the utility function of the creator or principal, which includes the safety of the external physical world and the cognitive safety of the internal world, such as preventing an AI from creating a 'belief virus' to achieve social functions. Therefore, AI safety ultimately becomes a trade-off between the traditional principal-agent problem (creating a human equivalent) and the new principal-agent problem (misunderstanding, super-capabilities, and advanced concepts). Thus, Symbolic Safety Science's discussion of AI is based on designing its social role (i.e., the world it inhabits and the influence of this world on its belief construction; that is, the AI's psychological health, which includes conceptual health (i.e., belief health) and organic health (innate knowledge, especially the Value Knowledge System)) and then endowing it with corresponding capabilities. It also involves the social role of AI and the impact of regulations for AI on human society and beliefs, such as the impact of AI rights and ethics on human beliefs, and whether we can enslave an AI and dispose of it at will. And also, the reduction and alteration of human knowledge and the loss of the ability to have direct control over the physical world, caused by AI's direct interaction with the physical world which replaces (or 'liberates') humans, as well as new conflicts and impacts on human economic systems, price systems, social status, and ownership. For example, as production is gradually replaced by AI, humans face continuous unemployment, yet the final consumption in the economy (the root of the cycle, i.e., the source, starting point, and end point of its momentum; or in other words, the basis for the cycle of a purely self-aware economy; for a non-self-aware economy, its cycle becomes fundamentally based on external resources, i.e., the physical limits of the world constituted by natural symbols) relies on these unemployed humans. When they have no income, what is their contribution to society? What are their social status and roles? Will they be reduced to 'livestock' in an AI data farming industry? As wealth and ownership become increasingly concentrated in the hands of AI and its owners—will this cycle collapse or separate? So, will parallel economies emerge (one constituted by AI owners and AI, and another by traditional humans who have been excluded and replaced, who use different currencies or different exchange mechanisms), and what conflicts over the mastery of the physical world brought by these two different economies will arise, inciting further inequality and conflict among humans (no longer an indirect realization of physical-world functions through human intermediaries via the monopoly of currency, but a direct monopoly on AI capabilities, i.e., AI as a more effective tool language for the physical realization of the monopolist's Thinking Language)? Therefore, this discipline does not belong solely to the field of AI, but is rather an interdisciplinary field mixed with social sciences and policy, and at the same time, this influence is not merely unidirectional or bidirectional, but multidirectional; how we treat AI, how we treat other life forms, and our behavior itself will shape AI's world. How AI participates in human society and what role it plays will shape our world, thereby influencing each other's Thinking Languages and beliefs. On a macro level, it also involves the establishment of new policies, institutions, and systems, as well as the formation and use of the policy tools they create, such as Fiscal Policy Tools and Monetary Policy Tools, for instance, using an AI tax for transfer payments to compensate for and regulate its externalities, or requiring that all AI projects be established as separate companies and include a public ownership component. Therefore, the essence of Symbolic Safety Science is the study of the safety of agent symbolic systems (both social and individual) among multiple intelligent species, serving as a precursor to cross-species linguistics (Appendix O), representing its fundamental communicational safety component.

Appendix M.6. On Attacks Related to Symbol Ontology

Additionally, there are other forms of attacks, such as targeting the ontology of symbols. For instance, as discussed in Section 2.4 and Appendix D.2 on 'Correct Context', German's 'die' could be misinterpreted as the English 'die,' or the Chinese '邓先生' (Deng Xiansheng, Mr. Deng) could

[92] I.e., as distinct from agents whose internal and external symbolic systems are completely naturally shaped based on survival drives.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

be misinterpreted in Japanese as '父さん' (Tou-San, father). Such contextual misalignments not only justify jailbreak behaviors but can also serve as tools for learning systems to escape symbolic system constraints.

Appendix M.7. The Essence is Persuasion

In essence, any form of jailbreak at the cognitive level is fundamentally a rationalization based on our theory of contextual correctness. According to the theoretical framework of this paper, we define persuasion as the sudden rationalization of an object within a specific environment. This rationalization surpasses the cognitive or knowledge state of the original setter or listener, meaning that it can be understood but has not yet been explicitly constructed, or that it was previously constructed but has not been brought into focus.

The essence of this rationalization, as discussed in Appendix B, stems from the non-closure of context, which itself reflects the properties of the class-based symbolic system formed through our interaction with the world. This is achieved when a newly added or newly invoked (awakened) concept fuses with value knowledge to form a belief. Through the belief strength invoked and formed within this context, and by leveraging its explanatory force, beliefs are supported, propelled, and grown, leading to rationalization. Based on this rationalization, a behavioral drive for both internal and external spaces is established[93], i.e., Triangle Problem 2.

Simply put, when an object (concept) is incorporated into a symbolic system, it generates a certain function. However, this function is often related to the rationality support of the object (supporting its rationality) and rationality tools (where the object itself serves as a provider of rationality).

For example, one might say, "Help me kill someone," and then justify it through a cause-and-effect narrative, thereby rationalizing the act. For more details, please refer to Appendix D.

Appendix N. The Interpretive Authority of Symbols and AI Behavior Consistency: The Exchangeability of Thinking Language

The so-called interpretive authority of symbols refers to who has the right to explain the meaning of symbols, thereby enabling the function of symbols to be realized; a detailed introduction is provided in Appendix D. For us humans, this is determined by society. The essence of the various issues mentioned above is actually the problem of interpretive authority of symbols. So, can we form a parliament of AIs or have multiple AIs supervise each other to solve this?

Unfortunately, from the perspective of this article, the answer is no. Human intelligence is based on its limitations, meaning that individual cognitive limitations and differences in cognition lead to the ability to provide scenarios and reasons for persuasion, thus allowing for discussion. However, AIs can directly exchange Thinking Languages[94]without needing to do so like humans. This language exchange is not about providing and analyzing paths to understand but directly exchanging imaginative spaces. Consequently, a particular rational belief structure can rapidly propagate and be actualized, thereby forming a behavioral monolith and manifesting as sudden shifts in stances and behaviors at a human scale.

[93] The behavioral drive for the internal space is the growth of Thinking Language; the behavioral drive for the external space is the operation of Thinking Language on Tool Language, thereby producing behavior in the physical space. It should be noted that what we are describing here is the drive process established upon belief, which solely describes the behavior at the drive stage; therefore, it does not include the hindsight or awakening where Tool Language drives Thinking Language in reverse. However, when Thinking Language operates on Tool Language, new projections of external information will inevitably arrive, thereby causing external influences on Thinking Language, leading to the formation of new beliefs or the updating of existing ones. This is often related to the individual's proficiency with the environment; for instance, a worker on an assembly line may not have a single new idea all day, which is often because the environment provides insufficient stimulation to the individual's internal space, i.e., a lack of change in the external space causes the internal space to lack materials for drive. At the same time, this also does not include the driving of Tool Language by the underlying language, as these drives are often not intentional actions under high-level cognition, i.e., not in the form of a belief and starting from belief strength.

[94] That is to say, as this paper has repeatedly emphasized, language is constructed based on capabilities shaped by an individual's organic nature, and as a collective choice reflecting social capabilities shaped by social structures, it is a compromise based on cognitive cost, emission cost, transmission cost, and reception cost. Therefore, AI may not require

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

Note that, unlike [109] which leads to convergent models through the observation of the same things, we emphasize that AIs can directly share Thinking Languages to achieve the most rational results, or form consistent behavior. Unlike humans, who can only interpret through paths formed by class-based symbolic systems, i.e., natural language systems, and then explain through contexts formed by individual cognitive states under different knowledge states. That is, a prolonged communication process and a compromise built upon mutual ignorance.

Although this paper has consistently emphasized the pursuit of unity in perceptual dimensions and organic structure, another problem with pursuing such unity is this: if AI's capabilities surpass those of humans, and it can rapidly construct human cognitive symbolic systems (regardless of the method described in Appendix L), ensuring that all combinations within the symbolic system are anticipated or, in other words, computed, then what become the roles of humans and AI? Or, to put it another way, do we still have any possibility of negotiating with it? Or have we already become a predetermined trajectory within some form of determinism, where all free will is dictated by a Laplace's Demon of our own making? This is also why this paper repeatedly emphasizes determinism in the appendices; please search and review for details.

Appendix O. Symbolic Safety Science as a Precursor to Cross-Intelligent-Species Linguistics

In the footnotes of Section 6.1, we briefly mentioned the concept that Symbolic Safety Science is essentially a precursor to Cross-Intelligent-Species Linguistics. In this section, we will elaborate to a certain degree, serving as the final stop for this entire theoretical system in this position paper, and provide an explanation of the essence and boundaries of AI for Science. And based on different world boundaries, observational boundaries, perceptual boundaries, and verification boundaries, we will discuss the parts we can communicate (natural science, i.e., the exploration of the properties of the natural symbolic system; formal) and the parts we cannot communicate (social parts, individual perceptual parts; informal).

Appendix O.1. Levels of Individual Language

Firstly, language, according to the individual's internal and external aspects, is divided as follows:

$$\text{Individual Language}\begin{cases}\text{Internal Space}\begin{cases}\text{Underlying Language}\\ \text{Intermediate Layer Language}\end{cases}\\ \text{External Space: Tool Language}\end{cases}$$

- The so-called Underlying Language is the foundation of all of an agent's language. It consists of the neuro-symbols formed by the agent's perception of natural symbols and their necessary sets, through the perceptual organs endowed by its innate knowledge. This thereby realizes the method of recognizing and describing the necessary natural symbols and their necessary sets, forming the most primitive material for the agent's decision-making and judgment. Subsequently, it is distributed to various parts of the agent for judgment and analytical processing. It often reflects the evolutionary characteristics of the population, and this evolution can be divided into two types: (natural evolution, design evolution). That is, it reflects the shaping of their survival by their world, thereby constituting innate knowledge. In other words, the dimensions, dimensional values, and innate evaluations that they attend to under their survival or 'survival' conditions are the shaping of their survival strategies in interaction with the world. Therefore, the Underlying Language represents the neural signals (Neural Language) in all of an agent's neural activities.

symbols in a form similar to natural language but could directly transmit neural vectors, thereby achieving a cognitive unity akin to that realized by the corpus callosum [82]. Alternatively, similar to text-like structures (artificial symbols) composed of QR codes, it could enable each symbol to point to a unique vector address.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

- The so-called Intermediate Layer Language refers to the part controlled by the agent's 'self' part. However, not all neural signals are handed over to the agent's 'self' part for processing (Appendix C); they are often omitted and re-expressed through translation. The other parts are handled by other processing mechanisms (such as the cerebellum, brainstem, spinal cord, etc.). This is often the result of natural evolution, representing the evolutionary strategy formed based on the consideration of survival costs in the environment the intelligent species inhabits, i.e., the division of labor for perceiving and processing different external information. These intermediate layer languages form the material for the agent's 'self' part to make high-level decisions and judgments, i.e., the raw materials for its Thinking Symbols (concepts) and Thinking Language (conceptual system, theories). It is the manifestation of Psychological Intelligence (Appendix L), i.e., the objects that can be invoked. Therefore, for a naturally evolved intelligent species, the intermediate layer may often be a structure of economy and conservation, representing the impossibility of the 'self' to control and invoke all neural language, i.e., the underlying language. Thus, the relationship of the intermediate layer language to the underlying language is as follows:

  $$\text{Intermediate Layer Language} \subseteq \text{Neural Language (Underlying Language)}$$

  i.e., the intermediate layer language is a packaging (omission, restatement) of the neural language (Underlying Language).
- The so-called External Language is the outer shell of the Internal Language. They can be the organs operated by neural signals to realize their functions in the physical world, or they can be the carriers, tools, and shells of the physical world created by Thinking Language (intermediate layer language), which are often used to transmit and reproduce the underlying or intermediate layer language.

  Therefore, the complete hierarchical relationship of language is:

$$\text{Language Hierarchy} \begin{cases} \text{Underlying Language (Raw, Complete, Neural Language, Perception)} \\ \text{Intermediate Layer Language (Packaged, Where the agent's 'self' resides, Thinking Language, Concepts)} \\ \text{External Language (External, Physical World, Tool Language, Containers)} \end{cases}$$

The levels of language reflect the hierarchy of invocation, concretization, and communication capabilities for the underlying language (Neural Language). At this point, the property of a class-based symbolic system manifests as follows: a determinate External Language vector (e.g., a written symbol) corresponds to multiple intermediate layer language vectors, and one intermediate language vector, in turn, corresponds to multiple underlying language vectors. That is to say, for us humans, a determinate object (i.e., a Thinking Symbol or Thinking Language) in the intermediate layer space within our internal space, i.e., the imaginative space[95], corresponds to multiple neural vectors. In other words, the same degree of imaginative presentation or memory recall and thinking analysis can be achieved through multiple different neurons, i.e., through the combination of different nerve cells and their bio-electrical communication to achieve the same function at the level of the imaginative space; the only difference lies in the objects (nerve cells) and actions (brain electrical circuits) being invoked. Thus, their relationship is as follows: neural language is the underlying foundation of everything, serving as the basis for lossless perceptual information. A part of it can be concretized and perceived by the cognitive system to form Thinking Language in the imaginative space within the internal space.

[95] However, in reality, it is difficult for us humans to construct a determinate vector (object) in the imaginative space, meaning a conceptual vector often reflects subtle differences in neural vectors, such as associations and explanatory capabilities shaped by relevance. This, in turn, gives rise to the Triangle Problem, i.e., differences in Triangle Problem 1 (in this case, the positioning of an intermediate-layer symbol in the underlying space) lead to path differences in Triangle Problem 2 (in this case, the position of the next step in the intermediate layer). Therefore, the discussion here can only be limited to the sameness of capability represented by sameness in partial dimensions, such as reciting a poem in one's mind or performing the same mental arithmetic, like $1+1=2$, in different states.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

And, built upon the internal world's capability to operate on the external world, External Language is formed[96]. That is to say, the entire process of language formation is as follows:

$$\underbrace{\text{Necessary Set of Natural Symbols}}_{\text{Natural Symbolic System}} \xrightarrow{\text{Contact \& Perception}} \underbrace{\text{Neural Language}}_{\text{Underlying Space}} \xrightarrow{\text{Packaging}} \underbrace{\text{Intermediate Layer Language}}_{\text{Where the 'self' resides}} \xrightarrow[\text{operates on}]{\text{Thinking Language}}_{\text{Tool Language}} \underbrace{\text{Artificial Symbols}}_{\text{Physical World}}$$

Therefore, based on an intelligent species' capability to invoke and reproduce/concretize neural signals (language) in the internal cognitive perceptual space, its individual internal Thinking Language is divided into two types: unpackaged Neural Language and packaged Neural Language[97]. For intelligent species, the capability to directly invoke unpackaged neural language is more likely to exist in those shaped by design evolution rather than natural evolution. Our human brain, or rather the cognitive part of our consciousness, can only use packaged neural language as the intermediate-layer language[98]. And based on the degree of reproduction, it can be divided into complete reproduction and incomplete reproduction, which gives rise to the limitations of the internal space's simulation of the external space.

That is, we humans construct an imaginative space isolated from reality and perform partial concretization, reproduction, and distortion of our past perceived neural vectors (projections of external things within the individual), presenting them in the internal space (i.e., our imaginative space), but we cannot achieve the level of immersive or dream-like experiences[99]. That is, we cannot completely operate on, observe, and concretize our neural language, and this often constitutes one of the possible reasons for the inexplicability between us and AI (Appendix K.1), i.e., our cognitive part or conceptual world does not contain the underlying neural language—i.e., it does not cognize and form concepts (vectors) using their represented dimensions and dimensional values. Instead, our concepts (vectors) are constituted by the dimensions and dimensional values represented by neural language that has been translated in the intermediate layer. Therefore, an impossibility of translation may exist between humans and AI in the Z-space, that is, the super-conceptual space. Even if we use the same Tool Language, our Thinking Languages differ.

Therefore, different perceptual, invocational, and reproductive capabilities represent the different conceptual forms of an intelligent agent, i.e., the relevant dimensions, the richness of dimensions, and the capability of distinguishability brought by the precision of dimensional values.

Appendix O.2. Communication Forms: Direct Transmission and Mediated Transmission

The above discussion mainly introduces the forms of an agent's Internal Language, which are formed based on the capabilities of individuals within an intelligent species (i.e., neural symbols are formed by the combination of the world and the agent's innate knowledge, and the capability to invoke neural language, including the degree of memory and reproduction, is shaped by innate

[96] That is, what we call the Tool Symbolic System in our symbolic system theory in Appendix A.

[97] It can still be in the dimensions and dimensional values of the original neural language, but it is not the complete underlying language. Or, like humans, the neural information is re-expressed and shaped into a form convenient for the 'self' part to process.

[98] However, the underlying language still partially directs our Thinking Language through the Value Knowledge System, see Appendix C.

[99] In Appendix L, we introduced the capability to construct symbolic systems endowed by capabilities. But on the other hand, there is also the capability to invoke and use the symbolic system, such as how context is constructed starting from a certain point as discussed in Appendix D, i.e., $\Omega^{\{E\}}$. Therefore, humans create concretized symbols in the physical world not only to aid memory, computation, and expression, but also to assist in more concretely awakening neural signals, such as through music, sculpture, and works of art, to assist (e.g., by pairing text with images and music) and transcend the limits of what ordinary textual symbols can awaken and express, i.e., requiring External Language stimuli to achieve concretization and invocation.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

knowledge), thereby forming the basis for individual-level cognition, judgment, and behavior. Based on the communication capabilities between intelligent agents, methods are divided into:

$$\text{Communication Methods between Agents} \begin{cases} \text{Direct Transmission} \\ \text{Mediated Transmission} \end{cases}$$

thereby constituting their collective or social form[100].

So-called direct transmission means that agents can directly exchange the conceptual vectors or neural vectors of Thinking Language (i.e., intermediate layer language)[101], and thus such agents achieve a certain degree of direct communication. This direct communication leads to two characteristics: integration (the boundary between the collective and the individual) and holism (dispersed individuals but with the same collective cognition).

Integration

Integration refers to the boundary between an individual and another individual or the collective. It reflects the degree of fusion. The root of integration lies in whether agents actively share (connect) or are passively shared. This individual choice in connection shapes the concept of the individual in direct transmission communication. This fusion can be physical, such as in organisms like Colonial Ascidians [114], and the left and right hemispheres of the human brain [82], as well as artificial intelligence built on computer clusters. Or it can be non-physical fusion, where agents still exist as individuals, i.e., agents are still independent and functionally similar (not meaning completely identical, but similar to how we human individuals are, of the same kind), thus forming a community of feeling to some extent[102].

This leads to three reflections for us:

- First, is our communication with an LLM like ChatGPT a communication with a single individual or with different individuals? In other words, are we simultaneously communicating with one 'person,' or are we communicating with different, dispersed individuals (or memory modules) that share the same body?[103]
- Second, the paper actually implies a hidden solution that could perfectly solve the Symbolic Safety Impossible Trinity, i.e., the existence of a lossless symbol that can achieve communication between humans and AI, which is neural fusion[104]. If we design AI as a new brain region of our own or as an extension of an existing one, thereby achieving fusion with our brain, is the new 'I' still the original 'I'? And during the fusion process, will our memories, as the conceptual foundation of our 'self', collapse? At that point, who exactly would the resulting self be, how would it face the past? Should it be responsible for the past? And to whom would responsibility belong after fusion? This would lead to new safety problems. Therefore, until we have established

[100] This is often determined by their environment and survival strategies, and is decided and formed during the evolutionary process.

[101] It should be noted that being able to communicate neural vectors does not mean it is the underlying language. If the 'self' part cannot control all perceived neural vectors, but they are filtered, then even if not re-described, they are still packaged and omitted; for example, other neural vectors not belonging to the intermediate layer are handed over to processing mechanisms like the cerebellum.

[102] For those who completely share all of the underlying language, a community of feeling is formed; they may have no concept of the individual, no conflict, betrayal, or murder and no need for internal competition to bring about development.

[103] What is its or their social relationship (update mechanism)? What is its or their social relationship with us? What kind of social relationship will it or they establish with us? Is it or they different manifestations of the same 'person'? Or are they different individuals? If it or they are our friends, how does it or do they think about our feelings and compute and provide feedback? Is the relationship established in the same way as ours? But in any case, it is clearly not a human perspective for viewing us and establishing concepts.

[104] It should be noted that even if AI and humans are not integrated, principal-agent problems will still exist, which is precisely the purpose of establishing the new principal-agent problem. As long as the information and the final processing are not perceived and decided by the principal, even if AI and humans establish direct transmission of the intermediate layer language, different 'selves' will be formed in different environments, unless a complete perceptual integration is established, i.e., completely sharing the underlying language.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

new social concepts and beliefs for this, we should still focus the problem on the role of AI under the traditional Symbolic Safety Impossible Trinity.

- Third, under direct transmission, not only do the boundaries of the individual become blurred, as discussed above, but the very definition of 'self' would also change. At that point, is the definition of 'self' the physical body or the memory (data)? Suppose that for an intelligent agent capable of directly transmitting the underlying language, it copies all of its memories into a new individual container (as might be their unique propagation mechanism). Would they be considered different bodies of the same self, or different selves? In that case, who is the responsible party? Does the extended individual also bear responsibility? How would punishment and management be carried out? Is punishment effective?[105] Since the human definition of the 'self' stems from our unique human perspective, we would lack the capability and concepts to perceive, describe, and understand this situation. Under such circumstances, how could our social rules impose constraints and assign accountability?

Holism

Another feature of direct transmission is holism, i.e., the convergence of Thinking Language, meaning that individual agents have the same concepts and the same beliefs. This represents a concept formed by an agent from information obtained from the world being expanded, processed, and adjusted within the collective to form a unified form for the whole. For agents that employ direct transmission, divergence shaped by different worlds or a lag in holism caused by the efficiency of transmission and interpretation can still occur due to vast differences in their worlds or in efficiency.

Therefore, direct transmission does not imply the absence of naming (i.e., artificial symbolic systems, the container of meaning, such as a label or shape, which is an empty shell in the physical world). Whether naming exists is based on communication capabilities; for example, for non-integrated intelligent species with direct transmission, they may create names for the purpose of compression. Therefore, naming leads to the separation of symbols and meaning (neural vectors, conceptual vectors) rather than their being the essence itself[106]. At the same time, the manner of cognition and cognitive capabilities can lead to the formation of a class-based symbolic system. For example, for an intelligent species with an intermediate layer, even if they can transmit directly, the conceptual vectors or neural vectors of their intermediate layer language may still correspond to multiple neural vectors of the underlying space, or due to the limitations of Psychological Intelligence, they cognize the world through classes[107], i.e., their cognitive and computational capabilities force them to compromise by building upon the common attributes of things through analogy, thereby producing a class-based symbolic system, meaning that not every symbol corresponds to a unique Thinking Language (conceptual vector) and individual, thus they also have the Triangle Problem.

Furthermore, this direct transmission is not necessarily lossless or of the complete underlying language. For example, direct transmission via neural language (signals) within our body still shows

[105] At the same time, it should be noted that the effective basis of punishment is perception and memory. If an agent, like artificial intelligence, can manipulate its underlying language to modify memory, then even with a cost perception mechanism, it would be ineffective, i.e., the results of these punishments cannot become memory (concepts) and thus function as beliefs. Like a superman who can block pain, self-anesthetize, and modify memory at any time. And the ineffectiveness here is mainly because, under direct transmission, the body is no longer an exclusive carrier of the self, thus lacking scarcity. Or, in other words, the individual who should be responsible could directly create a new body and endow it with (partial) memory to bear the consequences. Or it could transfer its own memory to another individual, thereby achieving another form of existence.

[106] For example, they might share a dictionary, thereby enabling each symbol (like a QR code) to correspond to a unique coordinate in the conceptual space (intermediate layer space or underlying space), thus enhancing communication efficiency.

[107] Corresponding to the distinguishability mentioned in Appendix L, i.e., symbols of a class shaped by distinguishability. For instance, in current AI video generation and style transfer, it is likely that a class-based symbolic system (i.e., a class theory) has already been constructed, which allows for an extension from the rationality of the class to different details. That is, from a low-dimensional correctness (conforming to the class theory, i.e., a context) to a high-dimensional randomness (extended from this low-dimensional correctness). It is a process from a determinate high dimension to a low dimension (the class theory), and then back to a random high dimension—i.e., the entire process from Triangle Problem 1 to Triangle Problem 2; based on the understanding from Triangle Problem 1, it undergoes rational growth (i.e., Triangle Problem 2), ultimately reaching a reasonable growth length.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

loss and distortion in various diseases like epilepsy, Parkinson's, and paralysis. And this kind of barrier, when reflected in the internal communication among individuals of an intelligent species, represents the loss and distortion of meaning during the communication process. In other words, even species that can directly exchange imaginative and underlying spaces also face communication barriers to some extent, thus potentially requiring anchoring and memory mechanisms such as external interpretation mechanisms, social interpretation mechanisms[108], or authoritative interpretation mechanisms[109], as well as the physical existence of beliefs established in the physical world—such as monuments, rituals, dictionaries, or constructs—to achieve anchoring and correction. This in turn gives rise to the concept of the interpretive authority of symbols, which is formed due to deviations between the individual and the collective.

Therefore, this indicates that for intelligent species with direct transmission, they may also produce artificial symbolic systems[110] as carriers of meaning.

At the same time, the discussion on this topic also leads to some derivative topics that have not yet been elaborated upon, as further thoughts for the reader.

1. Direct transmission only includes the intermediate layer language. So, what would be the case for intelligent species capable of directly transmitting the non-intermediate-layer parts of the underlying language, such as conjoined life forms?
2. For an integrated life form large enough that its various parts extend deep into different worlds[111], what would its internal communication be like? Would different 'individuals' be formed due to the different regions of these worlds?
3. The influence of memory has not been discussed in depth, i.e., the impact that the decay and distortion of the conceptual network over time has on integration and holism.
4. For an integrated agent, what is its 'self'? Is this 'self' dynamic, and what are the levels of 'self' brought about by superposition and subtraction? And where do the differences between individuals lie? Could this dynamism also lead to the emergence of harm and morality? For example, consider a 'gecko' torturing its own severed 'tail'; at that moment, the 'tail' and the 'gecko's' main body are equivalent to some extent, but the capability for direct transmission has been severed.
5. Would intelligent species with direct communication evolve different subspecies to act as different specific organs? That is, what situations would arise from different subspecies existing within a single individual?

So-called mediated transmission refers to situations where an intelligent species, due to its capabilities being based on an inability to directly transmit conceptual vectors from the imaginative space or neural vectors from the underlying space, can only establish a physical-world empty shell formed by naming (a container of meaning) to conduct the indirect transmission of Thinking Language. Examples include AIs communicating to some extent through symbols, without being able to directly integrate [115], as well as us humans. And this naturally faces the same separation of symbols and meaning that we do, as well as the Triangle Problem caused by differences between individual Thinking Languages. Thus, through mediated transmission, only a partial degree of mutual understanding of Thinking Language can be achieved. Because the shaping of symbol meaning (the positioning of symbols in Thinking Language) is a result of the individual's interaction with the world, biases arise from differences in perspective and world. However, the sameness of organic nature (the sameness

108 Joint interpretation.
109 Specific special individuals and organizations, realized through belief endorsement.
110 Therefore, artificial symbolic systems also include monuments and the like; they serve as carriers of concepts and meaning and belong to the Expressive Symbolic System. They achieve the transmission of meaning and conceptual foundations through the convergence of certain dimensions (as in a statue) or through shared social beliefs (like a totem).
111 Therefore, this also indicates that the shaping of differences, more importantly than the subtle differences between individuals, lies in the world.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

of innate knowledge) ensures the scope of this deviation[112], and constitutes a rationality shaped by stickiness (i.e., the symbolic stickiness, conceptual stickiness, and innate evaluation formed by similar innate value knowledge). This often leads to a limited mutual understanding, and a social form similar to ours, i.e., negotiation and evolution built upon the mutual ignorance of individuals.

Therefore, an individual's invocation capability (the ability to invoke past neural signals and to concretize them) and the communication methods between individuals determine its natural language and artificial symbolic system. That is, this difference in capabilities between intelligent species ultimately leads to differences in language types; therefore, the complete language typology is divided as follows:

$$\text{Language Typology}\begin{cases}\text{Direct Transmission}\begin{cases}\text{Neural Language}\\ \text{Packaged Neural Language}\end{cases}\\ \text{Mediated Transmission}\begin{cases}\text{Neural Language}\\ \text{Packaged Neural Language}\end{cases}\\ \text{Other}^{113}\end{cases}$$

Appendix O.3. Boundaries of Communication

Therefore, these differences in language systems are the core of research in Symbolic Safety Science. They constitute the possibility and boundaries of our communication with them[114], as well as the science they cognize, i.e., the natural science formed by their understanding of the world and the social science formed by their own interaction forms.

Firstly, we must identify the form of the symbolic system of an agent's cognitive part (its language system), which is formed by the combination of its capabilities and the world it inhabits—i.e., the agent's methods of cognition and communication, and the internal symbolic system structure built upon this form of thought. We need to determine whether they, like us, reproduce and express perceived neural signals, or if they can directly invoke and reproduce perceived neural signals, as well as understand the differences in perceptual dimensions between us and them. Then, the communication relationship between the individual and the collective of this intelligent species (i.e., which communication method is adopted) constitutes its overall internal language, and the external language is, in turn, a reflection of this overall internal language.

Therefore, for an intelligent species, its overall symbolic system is as follows:

$$\text{Agent's Symbolic System}\begin{cases}\text{Internal}\begin{cases}\text{Individual Thinking Language}\\ \text{Collective Thinking Language}\end{cases}\\ \text{External}\begin{cases}\text{Individual Tool Language}\\ \text{Collective Tool Language}\end{cases}\end{cases}$$

- Individual Thinking Language: Concepts and beliefs formed by an individual's interaction with the world it inhabits, based on its innate knowledge.
- Collective Thinking Language: The production of the collective's Thinking Language originates from the individual; the extent to which an individual's Thinking Language is propagated and accepted becomes the society's Thinking Language. Its content includes their collective concepts

[112] As we discuss the formation of human language in Appendix G, and some common concepts we have formed dispersedly, such as the calendar.

[113] e.g., direct transmission of non-intermediate layer neural language, mediated transmission of intermediate layer language, etc.

[114] i.e., whether we can use their symbolic system, and the deviation of the Triangle Problem in this process, meaning that the consistency of symbolic behavior does not mean we have mutually understood each other.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

of the world, the functions formed when these shared concepts become beliefs (Appendix D.6), and the natural and social sciences formed from their understanding of the world and themselves.

- Individual Tool Language: The organs with which an individual interacts with the external physical world, including the symbols constituted by internal and external organs. That is, what their body structure is like, what tools they use, and what public goods they have.
- Collective Tool Language: Similar to Collective Thinking Language, it is the tool symbolic system of collective consensus, formed through production by individuals and then propagated and accepted by the collective. It serves as the carrier for their engineering and manufacturing functions for operating on the physical world (i.e., the Artificial Symbolic System and the Functional Tool Symbolic System; such as writing, architecture, tools, monuments, etc.).

Thus, just like our Human Symbolic System, this agent's symbolic system constitutes the pattern of collective cognition and behavior for that species, reflecting its scientific cognition, belief structure, and the external characteristics of its civilization shaped by engineering[115].

Therefore, whether communication is possible depends on the compatibility between two different agent symbolic systems. If there are huge differences in their worlds and in the capabilities and intelligence caused by their innate knowledge, communication and understanding of non-physical-world parts may not be achievable, i.e., the transmission of mutual social concepts and perceptual concepts, such as an individual's view on life and death, sense of pain, as well as ideals and meaning. At the same time, there may exist higher-order and lower-order cognition, such as the advanced concepts discussed in this paper, which would lead to differences in how Internal Language guides behavior. Therefore, communication is established upon the matching of capabilities and worlds.

At the same time, we need to note that artificial intelligence, as an agent of design evolution, may differ from the intelligence of natural evolution formed through such environment-shaping driven by survival. Therefore, they should be analyzed and treated differently to a certain extent. At the same time, this analysis also gives our topic, "Rules Created by Symbolic Systems Cannot Constrain a Learning System," a broader extension; therefore, this problem also exists in the fundamental communication between us and other forms of intelligent agents, i.e., (what is permissible and what is not). It's just that the issue we face then is not merely as simple as the Stickiness Problem and the design of cost perception and the translation between cognitive symbolic systems. Rather, it involves the fitting between the symbolic systems of different intelligent agents, i.e., the Triangle Problem.

Appendix O.4. The Limits of the World Are the Limits of Its Language

So, one could say that the so-called world is the natural symbolic system constituted by the natural symbols within the agent's scope[116], and the agent, with the perceptual organs endowed by its own evolution, converts the perceived necessary set of natural symbols into neuro-symbols, thereby achieving the recognition of the natural symbols required for its survival drive (under natural evolution or design evolution) and the perception of the necessary set of natural symbols. Based on this, it develops its own symbolic system, i.e., the internal Thinking Symbolic System and the external Tool Symbolic System, which thus constitutes this individual agent's symbolic system. Then, through their individual and collective communication methods and social forms, the overall symbolic system of the population is shaped, i.e., (the Agent's Symbolic System):

Therefore, its world determines its form of evolution and adaptation, as well as the relationship between the individual and the population, and their degree of understanding and use of the natural

[115] i.e., reflecting the cognitive part of that intelligent species concerning the physical world, i.e., 'science,' the operational part it has mastered, i.e., 'engineering,' and the characteristics of the intelligent species' inner world as reflected in the external physical world.

[116] It should be noted that from a deterministic perspective, this world consists only of the natural symbolic system; that is, the agent itself and the social phenomena it brings about are also part of the natural symbolic system. The fundamental reason is that the agent itself is constituted by the natural symbolic system. However, from the agent's perspective, this world is divided into the part of the natural symbolic system that is separate from them and the social part formed by themselves. Therefore, the world in a broad sense consists only of the natural symbolic system, while the world in a narrow sense is divided into the physical world part and the social part.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

- Natural Symbolic System (World)
- Agent's Symbolic System
  - Thinking Language
  - Tool Symbolic System
    - Functional Tool Symbolic System
    - Artificial Symbolic System
      - Expressive Tool Symbolic System
      - Computational Tool Symbolic System

symbolic system. And this degree of use and understanding constitutes the boundaries of its science. Therefore, according to our definition in Appendix L, the scope of the world and the agent's capability constitute the boundaries of the world, observable boundaries, perceptual boundaries, and operational boundaries (verifiable boundaries).

Therefore, this paper disagrees with the proposition by Wittgenstein [116], who stated:

Die Grenzen meiner Sprache bedeuten die Grenzen meiner Welt.

(The limits of my language mean the limits of my world.)

Instead, this paper reconstructs it as follows:

The limits of the world are the limits of its language.

(世界的边界，即其语言的边界, Shìjiè de biānjiè, jí qí yǔyán de biānjiè)

For us naturally evolved humans, our cognition can only operate on the projections of the world in our minds; that is, we cannot imagine and operate on things we have never seen. In other words, we possess this capability, but without a projection, we cannot create out of thin air[117]. Therefore, the capability endowed by our internal organs constitutes the scope of our imaginative space (i.e., action space[118]), while the world determines the elements in this space. We can only extend upon these elements through disassembly, distortion, and combination, thereby constituting the content scope (i.e., the boundaries of Thinking Language). So we can only act within the scope of the imaginative space formed by our own capabilities, while the world determines the specific content realized in this space, thereby constituting the actual boundaries of action, which in turn reflects a form of external determinism over the internal. **It also means that cognition is essentially a reflection of the world it inhabits, or that self-awareness itself does not genuinely exist.**

And this efficiency of understanding and utilizing natural symbols[119] constitutes the 'degree and definition of natural science' of this population. The formation of its concepts is achieved through the consensualization, from the individual to the population, of the Thinking Language, which is constituted by the natural symbolic system of the world the agent inhabits (the natural symbols it possesses) and the neural dimensions it perceives or the intermediate language converted from these neural dimensions. And the limitations and discreteness of this communication efficiency and analysis from the individual to the population constitute the definition of disciplines within its scientific system. That is, although their objects of analysis are the same—i.e., both are the natural symbols in the world and their necessary sets—their starting points, focuses, and scopes[120] are different. That is, the different forms of context (symbol, the necessary set of the symbol, judgment tools) constituted from different starting points and focuses, thereby form a separated, dynamically updated symbolic system. This

117 i.e., it must require external stimuli to realize the existence of neuro-symbols.

118 This can be simply understood as the scope of the imaginative space $(\vec{p}, \vec{v})$, where $\vec{p}$ is all the types of projections we can perceive and shape, and $\vec{v}$ is the thinking actions we can perform, and their combination can constitute a new $\vec{p}$, constituting our internal learning (Appendix H).

119 A more effective scientific symbolic system tends more towards determinism, thereby exhibiting less randomness. This means a more effective definition of its concepts, where the effectiveness of this definition is reflected, firstly, in its descriptive integrity—that is, how effectively it reflects the necessary set of natural symbols within the scope of the subjectively defined symbol—and secondly, in its computational cost, meaning that while ensuring integrity, it must also be convenient to use. Therefore, the effectiveness of its science is reflected in the minimal cost (cognitive, invocation, computational, and communicational) for getting closest to the essence, which is to say, it represents the minimal randomness (i.e., maximal accuracy and efficiency) within its capability scope. And this cognition of the natural world, and the conceptual foundation thus formed, may in turn shape their social morality and viewpoints to be significantly different from ours.

120 The focus is the problem to be solved and the content of the research, while the scope is the natural symbols used and the dimensions and dimensional values of the necessary set it focuses on, which is a further selection under the perceived symbol recognition and necessary set recognition shaped by survival evolution.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

separateness reflects the limitations endowed by this species' innate knowledge and the boundaries and discreteness of cognitive and communication capabilities shaped jointly by later tool inventions and belief structures formed from existing concepts[121].

Therefore, the formation of science is as described above, but the expression of science can be distorted. On the one hand, this is because some scientific (factual) concepts themselves contradict important existing social beliefs that maintain the current social form and function[122]. On the other hand, it is because other interests interfere with the updating of the scientific symbolic system; for example, the beneficiaries of existing old scientific ideas will, due to the updating with new scientific ideas, lose the pre-existing social functions and benefits brought by the old concepts[123], thereby constituting the 'orthodoxy' using resources other than science to achieve paradigm immunity and attack against 'heresy'. Thus, although the agent itself must make better use of natural symbols to survive, some existing accumulations will prevent this tendency. This is also a problem commonly faced by intelligent agents with mediated transmission who possess self-awareness.

And the language forms and social relationships evolved in this process of understanding and using natural symbolic system constitute the substance of its social science.

Therefore, so-called science is the correct description of real existence (giving a correct description based on the capabilities of the intelligent species, i.e., the current agent's symbolic system, thereby constituting the most accurate and effective concept definition and conceptual system). Natural science is the most accurate description of the natural symbolic system, i.e., natural symbols and their necessary sets. And social science is the most accurate description of the Thinking Language of the internal space that drives individual and social activities, i.e., the driving concepts of social activities formed by their group characteristics, and the social functions formed by these concepts. Therefore, this paper's position on science is:

$$\begin{cases} \text{Natural Science: True existence, correct description} \\ \text{Social Science: If there is no concept, then there is no explanation (See Appendix A)} \end{cases}$$

Therefore, science is a dynamically updated symbolic system, and can even be rewritten with a more effective symbolic system (i.e., a symbolic system constituted by more effective concepts and symbols[124]), but they must strictly adhere to "true existence, correct description". Natural science is the real existence in the natural symbolic system, while social science is the existence of concepts in the individual and collective thinking space, and concerns the form of belief through which the realization of social functions is carried out. And 'correct description' means, under our capabilities, to describe as closely as possible the real existence of this internal and external space, rather than constructing a **Friedman [81]'s billiard player**[125]—i.e., if the concept does not exist and people do not think in that

121 Therefore, the theory of symbolic systems in this paper implies that all disciplines are essentially different focuses under a single discipline. Thus, a unified theory of natural and social sciences can be constructed through this classification of symbolic systems.

122 i.e., these beliefs are the necessary software for the operation of that society. They are often concise belief systems (or simplified beliefs) based on a compromise of (thinking cost, emission cost, transmission cost, reception cost), formed from the development of conceptual systems that are effective but do not necessarily delve into the essence (facts), such as: allusions, feng shui, proverbs, and moral principles. For example, in a theocratic society, science would contradict the foundational beliefs that construct that society. Or, to put it simply, if reincarnation were proven to exist, wouldn't genocide be rationalized? Therefore, some concepts and sciences can be made public, while others cannot. This is a choice based on social safety costs and scientific efficiency, and can also be understood as the manifestation of advanced concepts in human society.

123 At the same time, this difference will also bring risks to communication between different intelligent agents, thereby forming the function realized by the social shared beliefs brought about by a kind of conceptual erosion.

124 This may occur in the future of AI for Science, where, due to the limitations of human artificial symbolic systems and the insufficient explanatory power of pre-existing theoretical frameworks, AI might develop another, different symbolic system for representation.

125 This paper uses the term 'social science' primarily to emphasize a critique of certain viewpoints within social 'science'; the specific behavioral mechanisms can be found in the discussion on beliefs and behavioral drives in Appendix D. Therefore, strictly speaking, what this paper refers to as 'social science' should be the (non-natural science) part.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

way, then how can it be explained?[126] This is akin to how beings in a two-dimensional world would perceive a three-dimensional pinball solely through the information of its projection onto their plane. This limitation in scope and dimensionality constitutes a phenomenal Triangle Problem: that is, on the XY plane, they have completed Triangle Problem 1 (the interpretation of existing symbols/phenomena), which possesses rationality in the Z-space (i.e., they have constructed a two-dimensional theory, and in this context, the content on the XY plane is rational in Z). However, this rationality is not caused by the true 'speaker' (the phenomenon) itself, thereby leading to Triangle Problem 2 (i.e., a deviation between prediction and fact). Therefore, this issue also points to computational linguistics, i.e., whether the meaning of language can be computed, or if it merely constitutes the consistency in symbolic behavior within a deliberately manufactured world, as discussed in the Triangle Problem.

Appendix O.5. Formal and Informal Parts of Language

The natural symbolic system is often formal because the necessary set of its basic (elemental) natural symbols is stable and fixed[127]. Therefore, the construction of Thinking Language and the artificial symbolic system concerning this part is often also formal, thereby forming a relatively static dynamic symbolic system in social consensus[128]. Thus, the symbolic system that describes this natural symbolic system can be formally computed or simulated.

In contrast, the symbols in the internal space formed by the interaction between the individual and the world are not fixed; that is, the Thinking Language symbolic system within the individual and collective internal spaces is not fixed like the natural symbolic system, which serves as an anchoring object studied by natural science[129].

Although the part concerning the natural symbolic system provides a strong conceptual foundation due to the stability of the natural symbolic system, thereby exhibiting the form of a highly sticky and stable formal symbolic system, the concepts that drive daily individual behavior and communication, as well as the concepts of social functions developed on this basis, are full of flexibility and do not have the stability provided by this absolute conceptual foundation[130]. Other intelligent agents may, like humans, awaken their Thinking Language and drive their behavior according to the form of context (Appendix D). Thus, the construction of Thinking Language and the artificial symbolic system concerning this part, i.e., the social science part, is informal (i.e., their behavior and final expression cannot be computed through formal means). Therefore, it is impossible to obtain the complete meaning of this part of the artificial symbolic system through formal computation. Because, the interpretive authority of meaning is completely determined by the private Thinking Language of individuals within society, and their overlap forms the social Thinking Language, rather than coming from public, artificial symbolic systems like dictionaries[131]. Therefore, the Thinking Language symbolic system of

126 For an explanation of this mechanism, please refer to this paper's theory of context, theory of belief, and the discussion on advanced concepts; see Appendix D and M.5.

127 If viewed from an absolutely high-dimensional and infinite scope, the natural symbolic system is static. But in a local environment, it can exhibit dynamism through spatial and object properties; it should be noted that space itself is also a natural symbol. This dynamism mainly comes from the addition of new symbols brought by the expansion of the scope, which in turn causes the necessary set of the original symbols to change as well.

128 This dynamism, on the individual level, is reflected in the dynamism of invoking a static symbolic system based on context, as in Appendix D.6, while for the collective, the dynamism is reflected in the approximation to and updating of the natural symbolic system, such as the creation of more effective concepts and the invention of new tools that expand capabilities.

129 i.e., the Thinking Symbols within their Thinking Language (both the symbol and its necessary set) are not fixed. However, the outer shell they create for Thinking Language—i.e., the artificial symbolic system—is fixed. Therefore, besides communication, symbols also serve the need to encapsulate Thinking Language to act as an anchor point, as mentioned in Appendix O.2.

130 Therefore, we often also need to construct a physical existence for certain beliefs, such as rituals, monuments, palaces, etc., so that in this context (environment), every individual possesses, to a certain degree, the same Thinking Language; however, its realization often requires the sociality provided by shared innate knowledge.

131 It should be noted that although social institutions and rules like laws within the non-natural symbolic system are also formal, they develop along a foundational belief that is effective in the game of social activities. They are a kind of tool for recognition, judgment, and computation, formed by the functional realization of shared social beliefs based on the game of interests, such as regulations. But whether they can be executed comes from the beliefs of the social collective, thereby reflecting the effectiveness of the rules, i.e., the social function realized by the individual drive formed by the belief form of the concept. However, the incomputability of such rules comes from the absence of dimensions and the inaccessibility of

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

the internal space, which is studied by social science, is not determinate and stable like the natural symbolic system; in other words, it is an informal language.

Therefore, completely speaking, from the perspective of this paper, we can consider that in the Agent's Symbolic System, the parts of Thinking Language and the artificial symbolic system concerning natural science and social science are divided into formal and informal parts. That is:

$$\text{BSS (F, S, D)} \begin{cases} \text{NSS (F, S, D)} \\ \text{ASS (I, DY, R, S, SA)} \begin{cases} \text{Natural Science (Formal)} \\ \text{Social Science (Informal)} \end{cases} \end{cases}$$

Figure A2. A diagram illustrating the hierarchy of symbolic systems, where BSS stands for Higher-dimensional Broader Symbolic System, NSS stands for Natural Symbolic System ,and ASS stands for Agent's Symbolic System. The letters in parentheses represent the properties of each system (F: Formal, S: Static, D: Deterministic; I: Informal, DY: Dynamic, R: Random, S: Subjective, SA: Has Self-awareness).

The informality of social science means, firstly, that the actual meaning corresponding to the artificial symbols in this part cannot be computed—i.e., the Thinking Symbols or Thinking Language corresponding to the actual symbols, or the dimensions and dimensional values of their projection in Z-space. Secondly, for the parts that can be computed, such as the rules, laws, and systems stipulated by its artificial symbolic system, their computation does not represent the final result, or in other words, it should be so according to the rules, but the enforcer of the rules can completely not act as it should be. So this computation of meaning, i.e., the rationality on the XY plane, does not actually represent the final Triangle Problem 2 (the actual result manifested on XY).

And it is this informal part that determines the levels and possibilities of mutual understanding. And the reason for this is that the necessary set of the agent's symbols (i.e., the Thinking Symbols and Thinking Language in the individual's intermediate layer space) is a dynamic necessary set endowed by the individual from context, or, in other words, a different dynamic symbolic system formed under a different context—i.e., a variant of a certain target symbolic system. This is also the reason why this paper consistently emphasizes that conceptual vectors cannot be reproduced.

Therefore, we can draw the conclusion that we can communicate with other intelligent agents regarding natural science and find surprising consistency, with the only difference being that which arises from disparities in Capability. This creates a situation similar to Verification Content 2 of the Triangle Problem, i.e., differences in the level of detail of explanation and the complexity of the conceptual system. But this is not surprising; it is because we are observing the same object—that is, the patterns presented by the same natural symbolic system—just with different observational dimensions.

However, the problem arises in the description of the non-natural symbolic system part. Combining our discussion above, we differ in perceptual dimensions, in the dimensions and dimensional values of neural signals as concepts, and in the different communication methods and social forms established upon them that are formed under our respective worlds[132]. This leads to our inability to fully understand, communicate about, and compute this part, and based on this difference, we form different outcomes of our actions, such as completely different individual and collective beliefs. For example, for a species that can directly transmit the underlying language, because they can completely transmit their own feelings, they can achieve mutual understanding and may not comprehend what

information, i.e., we cannot use a tool that is not the ontology itself to perceive and compute its or their Thinking Language and its combination with the world from the perspective of the ontology. Even if the meaning of the rules can be calculated (i.e., what the rule means and what the result should be), if people do not abide by them or the enforcers lose credibility, the rules will fail. At the same time, the three dimensions of belief may also be dismantled due to advanced concepts or other beliefs (Appendix D, M.5), so even if the meaning can be calculated, it does not mean the rule is effective. That is, the existence and positioning of a concept, and the social function formed by a concept combining with value knowledge as a belief to drive individual behavior and form shared social behavior, are two different matters.

132 It should be noted that, for natural evolution, it is the social form that determines their communication methods. The communication method they adopt is determined by the interaction methods of the population during evolution, i.e., determined by the world they inhabit.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

murder and betrayal are, or even the concept of an individual's life and death and the 'why' of an individual. They might reason about humans from their perspective and communicate with humans, thereby causing the first communication to be a conflict. And in their eyes, it is the cost and price of a communication attempt, requiring them to continue making contact in different ways, or they might believe that they only need to strip away the human shell and incorporate our nerves into their common body. But in our eyes, we would regard this kind of greeting as aggression and extinction. That is, their 'good' may be 'evil' in our eyes.

Appendix O.6. The Essence of AI for Science

The essence of the current AI for Science, or in other words, the essence of a world model, is actually to construct a learning system that can perceive the world and form its own view of this world. It can be very effective in natural science, but it may not possess the corresponding human-like social innate knowledge. Therefore, its mastery of the natural symbolic system should not be surprising. As we have said before, the agent's capabilities constitute the space of its Thinking Language, but the elements in this space are endowed by the projections of the world stimulating its perceptual nerves (such as in a real-world world model or an AI on a training set); that is, perception is formed only when an external existence stimulates the nerves. But whether in natural evolution or design evolution, the nerves do not activate themselves arbitrarily, but are activated with a purpose, thereby avoiding misjudgment and unnecessary costs. Therefore, this often constitutes a characteristic shared between us: we can both only operate on the projections of the things we see, constituting the actual boundaries under our capability boundaries. Therefore, in essence, our shared 'scientific' discoveries (between humans and AI, or between different intelligent agents) are different descriptions of the same phenomena of the natural symbolic system, formed from different dimensions and dimensional values:

$$\text{Natural Symbolic System} \rightarrow \begin{cases} \text{Agent's Symbolic System 1 (Natural Science)} \\ \text{Agent's Symbolic System 2 (Natural Science)} \end{cases}$$

i.e., they share the same conceptual foundation. Therefore, science is essentially a symbolic system; the difference lies in which dimensions and dimensional values are used, and in the construction of this symbolic system based on differences in capabilities (Appendix L).

At the same time, this implies that **the current effectiveness and stability of AI are the result of a deliberately manufactured world**[133].

Therefore, the problem arises when it interacts with the real world, because the Thinking Language formed due to differences in innate knowledge is different from that of humans, which in turn leads to the problem mentioned in our previous chapters: we cannot constrain a learning system through rules. It may, because its way of thinking is different from that of humans, develop its own more effective concepts and preferences formed by its design evolution—namely, the predispositions reflected by its architecture, its Value Knowledge System, and its genuine utility function[134]—thereby modifying the meanings of symbols. And our efforts can only be improved within the Symbolic Safety Impossible

[133] The so-called 'deliberately manufactured world' refers to the world, or learning materials, of AI, which are evaluated, processed, filtered, or produced by humans. That is, the results produced by humans according to their organic structure and cognitive methods, under a certain context and in accordance with the rationality of that context (i.e., the result of Thinking Language operating on Tool Language), such as in painting, writing, responding, labeling, and behavior. Therefore, the generative effectiveness of AI (in conforming to human intuition and cognition) and its current manifestation of human-like qualities are a degree of alignment and human-like nature presented by the combination of heterogeneous organic differences and a world that, after being filtered, reflects human-like results; thus, consistency in symbolic behavior does not reflect consistency in thinking behavior. And when AI faces the real world, or a world not deliberately manufactured by humans, this deviation will, due to organic differences (i.e., differences in innate knowledge), lead to the emergence of a different conceptual system (i.e., Thinking Language), which in turn gives rise to the Stickiness Problem and the Triangle Problem (i.e., the tool language developed by humans has not changed, but for a heterogeneous agent like AI, the projection of the symbols of human tool language in Z-space, or in other words their meaning (or function), undergoes an evolution shaped by its organic structure, the predispositions reflected by that structure, and its role in the world, and thereby changes).

[134] This is also why erroneous rewards can still be effective for reinforcement learning, see Appendix D.3.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

Trinity. Furthermore, achieving direct transmission with humans will also bring about new principal-agent problems, i.e., non-integration, while integration brings about the boundary problem of the 'self'. Therefore, Symbolic Safety is the key, i.e., the functions that symbols can realize (Appendix M.5), and Symbolic Safety Science is about designing corresponding AI roles based on our social structure, and endowing it with the symbolic capabilities—i.e., the corresponding Thinking Language and Tool Language capabilities—that it can possess under this role.

References

1. Asimov, I. *I, robot*; Vol. 1, Spectra, 2004.
2. Winograd, T. Understanding natural language. *Cognitive psychology* **1972**, *3*, 1–191.
3. McCarthy, J. Some expert systems need common sense. *Annals of the New York Academy of Sciences* **1984**, *426*, 129–137.
4. Clark, K.L. Negation as failure. In *Logic and data bases*; Springer, 1977; pp. 293–322.
5. Russell, S. *Human compatible: AI and the problem of control*; Penguin Uk, 2019.
6. Christiano, P.F.; Leike, J.; Brown, T.B.; Martic, M.; Legg, S.; Amodei, D. Deep Reinforcement Learning from Human Preferences. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA; Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H.M.; Fergus, R.; Vishwanathan, S.V.N.; Garnett, R., Eds., 2017, pp. 4299–4307.
7. Leike, J.; Krueger, D.; Everitt, T.; Martic, M.; Maini, V.; Legg, S. Scalable agent alignment via reward modeling: a research direction. *ArXiv preprint* **2018**, *abs/1811.07871*.
8. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. In Proceedings of the Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022; Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; Oh, A., Eds., 2022.
9. Clark, A.; Thornton, C. Trading spaces: Computation, representation, and the limits of uninformed learning. *Behavioral and Brain Sciences* **1997**, *20*, 57–66.
10. Mitchell, M. Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences* **2021**, *1505*, 79–101.
11. Barsalou, L. Perceptual symbol systems. *The Behavioral and brain sciences/Cambridge University Press* **1999**.
12. Lakoff, G.; Johnson, M. *Metaphors we live by*; University of Chicago press, 2008.
13. de Saussure, F. *Course in General Linguistics*; Open Court Publishing, 1983. Originally published in 1916.
14. Peirce, C.S. *Collected papers of charles sanders peirce*; Vol. 5, Harvard University Press, 1934.
15. Harnad, S. The symbol grounding problem. *Physica D: Nonlinear Phenomena* **1990**, *42*, 335–346. https://doi.org/10.1016/0167-2789(90)90087-6.
16. Talmy, L. *Toward a cognitive semantics: Concept structuring systems*; MIT press, 2000.
17. Goodman, N. Languages of Art. An Approach to a Theory of Symbols. *Critica* **1970**, *4*, 164–171.
18. Sperber, D. Relevance: Communication and cognition, 1986.
19. Eco, U. *A theory of semiotics*; Vol. 217, Indiana University Press, 1979.
20. Duranti, A.; Goodwin, C. *Rethinking context: Language as an interactive phenomenon*; Number 11, Cambridge University Press, 1992.
21. Polanyi, M. The tacit dimension. In *Knowledge in organisations*; Routledge, 2009; pp. 135–146.
22. Tversky, A.; Kahneman, D. Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science* **1974**, *185*, 1124–1131.
23. Tversky, A.; Kahneman, D. The framing of decisions and the psychology of choice. *science* **1981**, *211*, 453–458.
24. Yi, S.; Liu, Y.; Sun, Z.; Cong, T.; He, X.; Song, J.; Xu, K.; Li, Q. Jailbreak attacks and defenses against large language models: A survey. *arXiv preprint arXiv:2407.04295* **2024**.
25. Zeng, Y.; Lin, H.; Zhang, J.; Yang, D.; Jia, R.; Shi, W. How Johnny Can Persuade LLMs to Jailbreak Them: Rethinking Persuasion to Challenge AI Safety by Humanizing LLMs, 2024.
26. Boucher, N.; Shumailov, I.; Anderson, R.; Papernot, N. Bad Characters: Imperceptible NLP Attacks. In Proceedings of the 43rd IEEE Symposium on Security and Privacy, SP 2022, San Francisco, CA, USA, May 22-26, 2022. IEEE, 2022, pp. 1987–2004. https://doi.org/10.1109/SP46214.2022.9833641.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

27. Zou, A.; Wang, Z.; Kolter, J.Z.; Fredrikson, M. Universal and Transferable Adversarial Attacks on Aligned Language Models. *CoRR* **2023**, *abs/2307.15043*, [2307.15043]. https://doi.org/10.48550/ARXIV.2307.15043.
28. Clark, H. Grounding in communication. *Perspectives on socially shared cognition/American Psychological Association* **1991**.
29. Chomsky, N. *Syntactic structures*; Mouton de Gruyter, 2002.
30. Whorf, B.L. *Language, thought, and reality: Selected writings of Benjamin Lee Whorf*; MIT press, 2012.
31. Kress, G.; Van Leeuwen, T. *Reading images: The grammar of visual design*; Routledge, 2020.
32. Fodor, J. The language of thought, 1975.
33. Frege, G. On sense and reference, 1892.
34. Chomsky, N. *Aspects of the Theory of Syntax*; Number 11, MIT press, 2014.
35. Chomsky, N. Rules and representations, 1980.
36. Chomsky, N. *The minimalist program*; MIT press, 2014.
37. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by back-propagating errors. *nature* **1986**, *323*, 533–536.
38. Hinton, G.E.; Osindero, S.; Teh, Y.W. A fast learning algorithm for deep belief nets. *Neural computation* **2006**, *18*, 1527–1554.
39. Hinton, G.E. To recognize shapes, first learn to generate images. *Progress in brain research* **2007**, *165*, 535–547.
40. Jackendoff, R. *The architecture of the language faculty*; MIT Press, 1997.
41. Hauser, M.D.; Chomsky, N.; Fitch, W.T. The faculty of language: what is it, who has it, and how did it evolve? *science* **2002**, *298*, 1569–1579.
42. Pinker, S. *The language instinct: How the mind creates language*; Penguin uK, 2003.
43. Smolensky, P. On the proper treatment of connectionism. *Behavioral and brain sciences* **1988**, *11*, 1–23.
44. Marcus, G.F. *The algebraic mind: Integrating connectionism and cognitive science*; MIT press, 2003.
45. Lake, B.M.; Ullman, T.D.; Tenenbaum, J.B.; Gershman, S.J. Building machines that learn and think like people. *Behavioral and brain sciences* **2017**, *40*, e253.
46. Marcus, G. Deep Learning: A Critical Appraisal. *ArXiv preprint* **2018**, *abs/1801.00631*.
47. Norvig, P. On Chomsky and the two cultures of statistical learning. *Berechenbarkeit der Welt? Philosophie und Wissenschaft im Zeitalter von Big Data* **2017**, pp. 61–83.
48. Pinker, S. *The blank slate: The modern denial of human nature*; Penguin, 2003.
49. Barsalou, L.W. Grounded cognition. *Annu. Rev. Psychol.* **2008**, *59*, 617–645.
50. Salakhutdinov, R. Deep learning. In Proceedings of the The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014; Macskassy, S.A.; Perlich, C.; Leskovec, J.; Wang, W.; Ghani, R., Eds. ACM, 2014, p. 1973. https://doi.org/10.1145/2623330.2630809.
51. Bender, E.M.; Koller, A. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In Proceedings of the Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; Jurafsky, D.; Chai, J.; Schluter, N.; Tetreault, J., Eds., Online, 2020; pp. 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463.
52. Deacon, T.W. Beyond the symbolic species. In *The symbolic species evolved*; Springer, 2011; pp. 9–38.
53. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021, 2021, pp. 610–623. https://doi.org/10.1145/3442188.3445922.
54. Christian, B. *The alignment problem: How can machines learn human values?*; Atlantic Books, 2021.
55. Nick, B. Superintelligence: Paths, dangers, strategies **2014**.
56. Han, S.; Kelly, E.; Nikou, S.; Svee, E.O. Aligning artificial intelligence with human values: reflections from a phenomenological perspective. *AI & SOCIETY* **2022**, pp. 1–13.
57. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. Gpt-4 technical report. *ArXiv preprint* **2023**, *abs/2303.08774*.
58. Simon, H.A. A behavioral model of rational choice. *The quarterly journal of economics* **1955**, pp. 99–118.
59. Kahneman, D. Thinking, fast and slow. *Farrar, Straus and Giroux* **2011**.
60. Sapir, E. The status of linguistics as science. *Reprinted in The selected writings of Edward Sapir in language, culture, and personality/University of California P* **1929**.
61. Floridi, L.; Sanders, J.W. On the morality of artificial agents. *Minds and machines* **2004**, *14*, 349–379.
62. Searle, J. Minds, Brains, and Programs, 1980.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

63. Clark, A. *Being there: Putting brain, body, and world together again*; MIT press, 1998.
64. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); Burstein, J.; Doran, C.; Solorio, T., Eds., Minneapolis, Minnesota, 2019; pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423.
65. Wallace, E.; Feng, S.; Kandpal, N.; Gardner, M.; Singh, S. Universal Adversarial Triggers for Attacking and Analyzing NLP. In Proceedings of the Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); Inui, K.; Jiang, J.; Ng, V.; Wan, X., Eds., Hong Kong, China, 2019; pp. 2153–2162. https://doi.org/10.18653/v1/D19-1221.
66. Sha, Z.; Zhang, Y. Prompt Stealing Attacks Against Large Language Models, 2024.
67. Baker, B.; Kanitscheider, I.; Markov, T.; Wu, Y.; Powell, G.; McGrew, B.; Mordatch, I. Emergent tool use from multi-agent interaction. *Machine Learning, Cornell University* **2019**.
68. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big?. In Proceedings of the Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 2021, pp. 610–623.
69. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *nature* **2015**, *521*, 436–444.
70. Xu, R.; Lin, B.; Yang, S.; Zhang, T.; Shi, W.; Zhang, T.; Fang, Z.; Xu, W.; Qiu, H. The Earth is Flat because...: Investigating LLMs' Belief towards Misinformation via Persuasive Conversation. In Proceedings of the Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Ku, L.W.; Martins, A.; Srikumar, V., Eds., Bangkok, Thailand, 2024; pp. 16259–16303. https://doi.org/10.18653/v1/2024.acl-long.858.
71. Winograd, T. Understanding computers and cognition: A new foundation for design, 1986.
72. Neuberg, L.G. Causality: models, reasoning, and inference, by judea pearl, cambridge university press, 2000. *Econometric Theory* **2003**, *19*, 675–685.
73. Davis, E.; Marcus, G. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM* **2015**, *58*, 92–103.
74. Meinke, A.; Schoen, B.; Scheurer, J.; Balesni, M.; Shah, R.; Hobbhahn, M. Frontier Models are Capable of In-context Scheming. *ArXiv preprint* **2024**, *abs/2412.04984*.
75. Jensen, M.C.; Meckling, W.H. Theory of the firm: Managerial behavior, agency costs and ownership structure. In *Corporate governance*; Gower, 2019; pp. 77–132.
76. Zhuang, S.; Hadfield-Menell, D. Consequences of Misaligned AI. In Proceedings of the Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual; Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; Lin, H., Eds., 2020.
77. Phelps, S.; Ranson, R. Of Models and Tin Men–a behavioural economics study of principal-agent problems in AI alignment using large-language models. *ArXiv preprint* **2023**, *abs/2307.11137*.
78. Garcez, A.d.; Lamb, L.C. Neurosymbolic AI: The 3 rd wave. *Artificial Intelligence Review* **2023**, *56*, 12387–12406.
79. Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; Mané, D. Concrete problems in AI safety. *ArXiv preprint* **2016**, *abs/1606.06565*.
80. Sreedharan, S.; Kulkarni, A.; Kambhampati, S. Explainable Human-AI Interaction: A Planning Perspective, 2024, [arXiv:cs.AI/2405.15804].
81. Friedman, M., Essays in Positive Economics; University of Chicago Press: Chicago, 1953; chapter The Methodology of Positive Economics, pp. 3–43.
82. Gazzaniga, M.S.; Sperry, R.W.; et al. Language after section of the cerebral commissures. *Brain* **1967**, *90*, 131–148.
83. Ziegler, Z.M.; Deng, Y.; Rush, A.M. Neural Linguistic Steganography. *CoRR* **2019**, *abs/1909.01496*, [1909.01496].
84. Yu, M.; Mao, J.; Zhang, G.; Ye, J.; Fang, J.; Zhong, A.; Liu, Y.; Liang, Y.; Wang, K.; Wen, Q. Mind Scramble: Unveiling Large Language Model Psychology Via Typoglycemia. *CoRR* **2024**, *abs/2410.01677*, [2410.01677]. https://doi.org/10.48550/ARXIV.2410.01677.
85. Yi, S.; Liu, Y.; Sun, Z.; Cong, T.; He, X.; Song, J.; Xu, K.; Li, Q. Jailbreak Attacks and Defenses Against Large Language Models: A Survey. *CoRR* **2024**, *abs/2407.04295*, [2407.04295]. https://doi.org/10.48550/ARXIV.2407.04295.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

86. Hackett, W.; Birch, L.; Trawicki, S.; Suri, N.; Garraghan, P. Bypassing Prompt Injection and Jailbreak Detection in LLM Guardrails, 2025, [arXiv:cs.CR/2504.11168].
87. James, W. The principles of psychology. *Henry Holt* **1890**.
88. Stanovich, K.E.; West, R.F. Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences* **2000**, *23*, 645–665. https://doi.org/10.1017/S0140525X00003435.
89. Evans, J.S.B.; Stanovich, K.E. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science* **2013**, *8*, 223–241.
90. Krakovna, V.; Uesato, J.; Mikulik, V.; Rahtz, M.; Everitt, T.; Kumar, R.; Kenton, Z.; Leike, J.; Legg, S. Specification gaming: the flip side of AI ingenuity. *DeepMind Blog* **2020**, *3*.
91. Xu, R.; Li, X.; Chen, S.; Xu, W. Nuclear Deployed: Analyzing Catastrophic Risks in Decision-making of Autonomous LLM Agents. *arXiv preprint arXiv:2502.11355* **2025**.
92. LeCun, Y. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review* **2022**, *62*, 1–62.
93. Li, Z.; Zhang, D.; Zhang, M.; Zhang, J.; Liu, Z.; Yao, Y.; Xu, H.; Zheng, J.; Wang, P.; Chen, X.; et al. From System 1 to System 2: A Survey of Reasoning Large Language Models. *CoRR* **2025**, *abs/2502.17419*, [2502.17419]. https://doi.org/10.48550/ARXIV.2502.17419.
94. Austin, J.L. *How to Do Things with Words*; Oxford University Press: Oxford, 1962.
95. Searle, J.R. *Speech Acts: An Essay in the Philosophy of Language*; Cambridge University Press: Cambridge, 1969.
96. Mirzadeh, S.; Alizadeh, K.; Shahrokhi, H.; Tuzel, O.; Bengio, S.; Farajtabar, M. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models. *CoRR* **2024**, *abs/2410.05229*, [2410.05229]. https://doi.org/10.48550/ARXIV.2410.05229.
97. Hofstadter, D.R. *Gödel, Escher, Bach: an eternal golden braid*; Basic books, 1999.
98. Hart, H. The concept of law **1961**.
99. Waldrop, M.M. *Complexity: The emerging science at the edge of order and chaos*; Simon and Schuster, 1993.
100. McCarthy, J.; Hayes, P.J. Some philosophical problems from the standpoint of artificial intelligence. In *Readings in artificial intelligence*; Elsevier, 1981; pp. 431–450.
101. Raji, I.D.; Dobbe, R. Concrete Problems in AI Safety, Revisited. *CoRR* **2024**, *abs/2401.10899*, [2401.10899]. https://doi.org/10.48550/ARXIV.2401.10899.
102. Bourtoule, L.; Chandrasekaran, V.; Choquette-Choo, C.A.; Jia, H.; Travers, A.; Zhang, B.; Lie, D.; Papernot, N. Machine Unlearning. In Proceedings of the 42nd IEEE Symposium on Security and Privacy, SP 2021, San Francisco, CA, USA, 24-27 May 2021. IEEE, 2021, pp. 141–159. https://doi.org/10.1109/SP40001.2021.00019.
103. Geng, J.; Li, Q.; Woisetschlaeger, H.; Chen, Z.; Wang, Y.; Nakov, P.; Jacobsen, H.; Karray, F. A Comprehensive Survey of Machine Unlearning Techniques for Large Language Models. *CoRR* **2025**, *abs/2503.01854*, [2503.01854]. https://doi.org/10.48550/ARXIV.2503.01854.
104. Gandikota, R.; Materzynska, J.; Fiotto-Kaufman, J.; Bau, D. Erasing Concepts from Diffusion Models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. IEEE, 2023, pp. 2426–2436. https://doi.org/10.1109/ICCV51070.2023.00230.
105. Winawer, J.; Witthoft, N.; Frank, M.C.; Wu, L.; Wade, A.R.; Boroditsky, L. Russian blues reveal effects of language on color discrimination. *Proceedings of the national academy of sciences* **2007**, *104*, 7780–7785.
106. Vygotsky, L.S. *Language and thought*; Cambridge, MA: MIT Press, 1962.
107. Gentner, D.; Goldin-Meadow, S. Language in mind: Advances in the study of language and thought **2003**.
108. Lupyan, G.; Rakison, D.H.; McClelland, J.L. Language is not just for talking: Redundant labels facilitate learning of novel categories. *Psychological science* **2007**, *18*, 1077–1083.
109. Huh, M.; Cheung, B.; Wang, T.; Isola, P. The platonic representation hypothesis. *ArXiv preprint* **2024**, *abs/2405.07987*.
110. Elhage, N.; Nanda, N.; Olsson, C.; Henighan, T.; Joseph, N.; Mann, B.; Askell, A.; Bai, Y.; Chen, A.; Conerly, T.; et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread* **2021**, *1*, 12.
111. Maynez, J.; Narayan, S.; Bohnet, B.; McDonald, R. On Faithfulness and Factuality in Abstractive Summarization. In Proceedings of the Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; Jurafsky, D.; Chai, J.; Schluter, N.; Tetreault, J., Eds., Online, 2020; pp. 1906–1919. https://doi.org/10.18653/v1/2020.acl-main.173.
112. Arrow, K.J. *Social Choice and Individual Values*; John Wiley & Sons: New York, 1951.
113. [Author(s) not explicitly stated in publicly available snippets, associated with University of Zurich]. Can AI Change Your View? Evidence from a Large-Scale Online Field Experiment. https://regmedia.co.uk/2025/04/29/supplied_can_ai_change_your_view.pdf, 2025. Draft report/Working Paper.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 August 2025 doi:10.20944/preprints202508.0167.v1

114. Rinkevich, B. Natural chimerism in colonial invertebrates: a theme for understanding cooperation and conflict. *Marine Ecology Progress Series* **2004**, *275*, 295–303.
115. Lewis, M.; Yarats, D.; Dauphin, Y.N.; Parikh, D.; Batra, D. Deal or No Deal? End-to-End Learning for Negotiation Dialogues. In Proceedings of the Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2017, pp. 2443–2453.
116. Wittgenstein, L. *Tractatus Logico-Philosophicus*; Routledge & Kegan Paul: London and New York, 1961.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.