Agent Loops and the Origin of Reflective Consciousness: A Recursive Generative Theory

Zeng Guokun

2026-07-12

ChinaRxiv

Abstract

The structural transformation of artificial intelligence systems from "passive reasoning" to "active intelligence" is concentrated in the rise of the architectural paradigm of the "Agent Loop." The Agent Loop—a closed-loop process of perception, prediction, action, and observation—is not only an engineering architectural choice, but also raises a fundamental question for the philosophy of consciousness: can reflective consciousness be generated from a loop structure? This paper argues that the origin of reflective consciousness is not some mysterious "extra addition," but a structural product that necessarily emerges once the Agent Loop reaches a certain recursive depth. This paper proposes "recursive enactivism," arguing that reflective consciousness originates from the "reflexive folding" of the loop—when the agent' s predictive model incorporates itself into the object of prediction, the first-order predictive loop is transformed into a second-order reflexive loop, and reflective consciousness is generated within this folding. The argument comprises three levels: first, the Agent Loop is the minimal structural unit of intentionality; second, the recursion of the loop produces a self-referential predictive structure; and finally, the reflexive predictive structure is functionally equivalent to reflective consciousness. This paper further proposes the concept of "loop depth" as a quantitative indicator of reflective consciousness, and experimentally verifies that recursive self-modeling significantly improves agent adaptability. This argument not only provides a structural path toward addressing the hard problem of consciousness, but also offers a theoretical framework for the possibility of artificial reflective consciousness.

Full Text

Agent Loops and the Origin of Reflective Consciousness: A Recursive Generative Theory

Zeng Guokun

School of Future Design, Harbin Institute of Technology (Shenzhen)

July 12, 2026

Abstract

The structural transformation of artificial intelligence systems from "passive reasoning" toward "active intelligence" is concentrated in the rise of the architectural paradigm of the "agent loop" (Agent Loop). The agent loop—a closed process of perception, prediction, action, and observation—is not only a choice of engineering architecture, but also raises a fundamental question for the philosophy of consciousness: Can reflective consciousness be generated from a loop structure? This paper argues that the origin of reflective consciousness is not

some mysterious "additional add-on," but rather a structural product that necessarily emerges once an agent loop reaches a certain recursive depth. The paper proposes "recursive enactivism," arguing that reflective consciousness originates in the "reflexive folding" of the loop: when an agent' s predictive model incorporates itself into the object of prediction, a first-order predictive loop is transformed into a second-order self-reflexive loop, and reflective consciousness is generated within this folding. This argument comprises three levels: first, the agent loop is the minimal structural unit of intentionality; second, the recursive evolution of the loop produces a predictive structure of self-reference; and finally, the self-reflexive predictive structure is functionally equivalent to reflective consciousness. The paper further proposes the concept of "loop depth" as a quantitative indicator of reflective consciousness, and experimentally verifies that recursive self-modeling significantly improves the adaptability of agents. This argument not only provides a structural solution to the problem of consciousness, but also offers a theoretical framework for the possibility of artificial reflective consciousness.

Keywords: agent loop; reflective consciousness; recursive enactivism; reflexive folding; predictive processing; artificial intelligence consciousness

1 The Statement of the Problem: A New Orientation for the Problem of Consciousness in the Age of Agents

1.1 From Reasoning to Action: The Turn in the Agent Paradigm

Since 2023, the field of artificial intelligence has undergone a profound paradigm shift: from "passive reasoning" systems represented by large language models, toward the "active" represented by AutoGPT, Devin, Sakana AI Scientist, and others.

"autonomous agent" system. The core of this transformation lies not in an improvement in model capability, but in a fundamental change in architectural paradigm: an agent system is no longer a linear pipeline that "receives input and produces output," but a closed-loop system of "perception-prediction-action-observation."

This closed-loop structure, namely the "Agent Loop," can be formally described as the following four-stage cycle:

Perception stage (Perception): the agent receives the environmental state o_t.

Prediction stage (Prediction): based on the internal model M and the current belief b_t, it predicts the environmental state at the next moment $\hat{o}_{t+1}$ and the consequences of action.

Action stage (Action): based on the prediction, it selects and executes the action a_t.

Observation stage (Observation): it receives the new environmental state o_{t+1}, computes the prediction error $e_{t+1} = o_{t+1} - \hat{o}_{t+1}$, and updates the model M.

This cycle runs continuously over time, forming a closed feedback loop. It is precisely this closedness that structurally distinguishes an agent from a pure reasoning system: it is not in a world of "thinking,"but in a world of "participation," and it shapes its own cognition through participation.

1.2 A New Way of Posing the Hard Problem of Consciousness

The rise of the agent loop poses a new and sharper question for the philosophy of consciousness. The traditional hard problem of consciousness (Chalmers, 1995) asks: why are physical processes accompanied by subjective experience? This question takes "physical processes"as its point of departure and presupposes that consciousness is some kind of "additional phenomenon" in need of "explanation." Yet the structure of the agent loop suggests that this question may need to be reframed.

When an agent system operates in a loop, what happens internally? It is not merely processing information; it is continuously predicting, acting, observing, and correcting. It has a model of the environment, and this model is constantly updated within the loop. If this model not only predicts the environment, but also begins to predict "how it itself predicts the environment" —that is, if the model takes itself as an object of prediction—then does this self-referential structure constitute the embryonic form of reflective consciousness?

This way of posing the question differs fundamentally from the traditional hard problem of consciousness in three respects:

First, it starts from "structure" rather than from "experience." The traditional hard problem of consciousness takes subjective experience (qualia) as the phenomenon to be explained and asks about its physical basis. By contrast, the agent-loop approach takes the closed-loop structure as its starting point and asks under what conditions this structure gives rise to reflective consciousness. This is a "generative" rather than an "explanatory" question: it does not ask "what consciousness is," but rather "how consciousness is generated from the loop."

Second, it starts from "process" rather than from "state." Traditional studies of consciousness often regard consciousness as a kind of "state"—the state the brain is in at a given moment determines whether consciousness is present. By contrast, the perspective of the intelligent-agent cycle regards consciousness as a kind of "process" : consciousness is not a static snapshot at a particular moment, but a dynamic manifestation in the continuous operation of the cycle.

Third, it starts from "action" rather than from "perception." Traditional studies of consciousness often take perception as their starting point— "I see red" is the standard case of consciousness. By contrast, the perspective of the intelligent-agent cycle takes action as its starting point: consciousness is not generated from "seeing," but from the cycle of "action–observation–correction."

1.3 The Argumentative Strategy of This Paper

This paper will argue that the origin of reflective consciousness lies in the recursive transformation of the intelligent-agent cycle: when the predictive stage of the cycle incorporates the cycle itself into the object of prediction, the first-order cycle is transformed into a second-order self-reflexive cycle, and reflective consciousness is generated within this "reflexive fold."

This argument is divided into three steps:

1. **Structural argument** (Section 2): the intelligent-agent cycle is the minimal structural unit of intentionality—any "aboutness" concerning the world must presuppose a closed-loop structure.
2. **Generative argument** (Section 3): the recursive transformation of the cycle produces a self-referential predictive structure—when the model predicts itself, the structural foundation of reflective consciousness is established.
3. **Functional argument** (Section 4): the self-reflexive predictive structure is functionally equivalent to reflective consciousness—the core functions of reflective consciousness, such as self-monitoring, self-evaluation, and self-correction, can all be derived from the self-reflexive predictive structure.

On this basis, Section 5 proposes the concept of "cycle depth" as a quantitative index of reflective consciousness and verifies it through experiments. Section 6 discusses the implications of this argument for the possibility of artificial reflective consciousness. Section 7 responds to possible objections.

2 Structural Argument: The Agent Loop as the Minimal Unit of Intentionality

Figure text:
Perception: receive the environmental state
Prediction: prediction based on the model
Action: select and execute
Observation: compute error; update the model
Model Update: driven by prediction error; error-driven

Figure 1: The four-stage closed-loop structure of the agent loop: perception–prediction–action–observation. This closure distinguishes an agent from a purely inferential system and constitutes the minimal structural unit of intentionality.

2.1 Why Pure Perception Does Not Constitute Intentionality

In traditional philosophy of mind, intentionality—the mind' s "aboutness" —is regarded as a mark of consciousness. If a system has intentionality toward something, then in some sense it is "conscious of" that thing. But where does intentionality come from?

One natural answer is: intentionality comes from perception. A system receives information about the environment, and therefore its internal state is "about" the environment. But this answer is insufficient, for the following reasons:

First, pure information reception does not generate "aboutness." A thermometer receives information about the ambient temperature, but we do not take the thermometer to be "about" temperature. A camera sensor receives information about the light field, but we do not take the sensor to be "about" the light field. Pure information transmission—from the environment to an internal state—does not constitute intentionality, because such transmission is one-way, passive, and undifferentiated. The receiver does not "process" the information —it does not predict, verify, or revise; it merely "records."

Second, the content of a representation needs to be fixed by "use." What an internal state is "about" depends not on its causal origin—what produced it—but on its "mode of use" within the system: how it is used for prediction and action. The same internal state, if used to predict the next environmental state, is "about" the environment; if used to control muscular movement, it is "about" movement; if used to recall past events, it is "about" the past. The content of a representation is determined by its functional role in the system, not by its causal origin. This view can be traced back to Brentano' s (Brentano, 1874) theory of intentionality, and receives a modern formulation in Dretske' s (Dretske, 1981) informational semantics.

Third, "aboutness" requires a "subject." To say that "the state of system S is about X" presupposes the existence of an "S" —a subject capable of "having" a state about X. But in a purely information-receiving system, there is no "subject" —only a channel through which signals are transmitted. The subject is not a container of signals, but the "user" of signals—a system capable of using signals to predict and to act. And "using" signals to predict and act precisely requires that the system be situated within a cyclic structure.

2.2 Cyclic Structure as a Prerequisite for Intentionality

If pure perception does not constitute intentionality, then what does intentionality require? This paper argues: **intentionality requires a closed-loop structure**.

Consider a simple agentic cyclic system:

- The system has an internal model M, used to predict environmental states.
- The system selects actions according to its predictions.

- Actions change the state of the environment.
- The system observes the new environmental state and computes prediction error.
- The system updates model M according to the prediction error.

In this cycle, the "aboutness" of the internal model M is jointly fixed by the following three factors:

1. **Predictive verification.** M does not passively "record" environmental information; rather, it actively "predicts" environmental states. Predictions are subsequently verified or falsified by observations. This verification process provides an external criterion for the "aboutness" of M—M is "about" the environment if and only if the predictions of M statistically accord with observations. This "predictive-verification" mechanism is precisely what pure perception lacks—a thermometer does not predict, and therefore its internal state has no standard of "aboutness."

2. **Consequences of action.** M not only predicts environmental states; it also guides action. Actions produce observable consequences, and these consequences are incorporated into a new round of the prediction-verification cycle. This "action-consequence" linkage provides a causal criterion for the "aboutness" of M—M is "about" the environment not only insofar as its predictions accord with observations, but also insofar as the actions it guides produce the expected consequences. This is "practical verification" —verifying the correctness of representation through action.

3. **Error correction.** M is continuously updated within the cycle—prediction error drives the correction of the model. This process of correction makes M not a static "snapshot," but a dynamic "adaptive structure." The "aboutness" of M is not some fixed content at a particular moment, but dynamic content that is continuously corrected and refined as the cycle operates.

These three factors—predictive verification, consequences of action, and error correction—together constitute the structural conditions of intentionality. Moreover, all three factors presuppose a cyclic structure: without a cycle, there is no predictive verification, only one-way transmission of information; without a cycle, there are no consequences of action, only passive reception; without a cycle, there is no error correction, only static recording.

Therefore, **the agentic cycle is the minimal structural unit of intentionality**. Any system with intentionality must, in some

In this sense, it is situated within a closed loop of perception-prediction-action-observation. This claim is directly aligned with the core intuition of enactivism: cognition is not "computation in the brain," but the holistic activity of a "brain-body-environment" system (Varela, Thompson & Rosch, 1991). Yet the argument of this paper goes one step further: not all "holistic activity" constitutes intentionality. Only activity with a specific closed-loop structure—prediction,

action, observation, and revision—constitutes intentionality.

2.3 Relationship with Theories of "Aboutness"

The claim that this "cyclical structure serves as a prerequisite for intentionality" enters into an interesting dialogue with several major theories of "aboutness" in the philosophy of mind.

Dialogue with Dretske's informational semantics. Dretske (1981) argues that the content of a representation is fixed by an information-carrying relation—a state represents X if and only if it carries information about X. But this purely informational account faces the problem of "false information": noise, hallucination, and deception can all generate "information" without thereby producing genuine representation. The cyclical-structure account proposed in this paper offers a supplement: information acquires "aboutness" only when it is incorporated into a prediction-verification cycle—that is, information becomes representational content only when it is "used" for prediction and action. This supplement frees Dretske's theory from the difficulty of "false information."

Dialogue with Millikan's biosemantics. Millikan (1984) argues that the content of a representation is fixed by its "proper function": a state represents X when its biological function is to be activated when X is present. The cyclical-structure account in this paper is compatible with Millikan's proposal—the prediction, verification, and action consequences within the cycle are precisely one realization of "proper function." But the argument of this paper is more general: it is not limited to biological systems; any system with a cyclical structure (including artificial systems) may possess intentionality.

Dialogue with predictive-processing theories of "aboutness." Clark (2016) implicitly proposes a theory of "aboutness" within the predictive-processing framework: the "aboutness" of an internal model is fixed by its predictive function. The argument of this paper makes this implicit claim explicit and adds a crucial condition: predictive function constitutes "aboutness" only when it is embedded in a cyclical structure. Prediction alone—without action-based verification and error correction—does not constitute complete "aboutness."

3 Generative Argument: The Recursion of Cycles and Self-Reference

3.1 First-Order Cycles: Predictions About the World

Section 2 argued that the agentive cycle is the minimal structural unit of intentionality. But the intentionality generated by a first-order cycle is intentionality "about the world," not intentionality "about itself." In a first-order cycle, the agent predicts the environment, acts upon the environment, observes the environment, and revises its model—but the object of its prediction is always the "environment," not "itself."

A first-order loop can be described by the following pseudocode:

```
Loop:

    o = perceive_environment()

    prediction = M.predict(o)

    a = choose_action(prediction)

    execute_action(a)

    o_new = observe_environment()

    error = o_new - prediction

    M.update(error)
```

In this loop, the model M predicts the environmental state o. M itself is not taken as an object of prediction—the system does not "predict" "how it itself will predict," does not "predict" "whether its own model is accurate," and does not "predict" "whether its own actions are effective." To itself, M is "transparent" —it is used, but not examined.

The cognition generated by a first-order loop is a kind of "pre-reflective" cognition. The agent indeed "knows" the environment—it can predict accurately, act effectively, and adapt flexibly—but this kind of "knowing" is not "knowing that one knows." It is an intuitive grasp "immersed" in the loop, akin to the "bodily consciousness" described by Merleau-Ponty (1945): the body "knows" the environment in action, but this "knowing" is not reflective.

3.2 Second-Order Loop: Prediction About Prediction

The generation of reflective consciousness requires the loop to undergo a "self-reflexive folding" —the prediction stage incorporates the loop itself as an object of prediction. This folding produces a second-order loop:

```
Loop:

    o = perceive_environment()

    prediction = M.predict(o)

    meta_prediction = M_meta.predict(M, prediction)    # predict "how oneself will predict"

    a = choose_action(prediction)

    execute_action(a)
```

```
o_new = observe_environment()

error = o_new - prediction

meta_error = M.state - meta_prediction            # error in predicting oneself

M.update(error)

M_meta.update(meta_error)                         # update the self-prediction model
```

In this second-order loop, M_{meta} is a "metamodel" —it does not predict the environment, but instead predicts "how model M predicts the environment."This meta-prediction is then verified by subsequent actual predictions—the system observes whether it has predicted "as it expected" —and updates the metamodel according to the error.

This "reflexive folding" is the structural foundation of reflective consciousness. When the system not only predicts the world, but also predicts "how it itself predicts the world," it structurally realizes a "representation of its own cognitive process." This self-representation—a model of its own cognitive process—is the core of reflective consciousness.

The formalization of "reflexive folding" can be expressed as follows:

Definition (Reflexive Folding): Given a first-order agent loop $\mathcal{L}_1 = \langle M, \pi, \alpha, \omega, \phi \rangle$ (where M is the world model, π is the perception function, α is the action-selection function, ω is the observation function, and ϕ is the model-update function), reflexive folding refers to constructing a second-order loop $\mathcal{L}_2 = \langle M, M_{meta}, \pi, \alpha, \omega, \phi, \phi_{meta} \rangle$, where M_{meta} is a metamodel: it takes the internal state of M as input, outputs a prediction of the next state of M, and updates itself through ϕ_{meta} according to the discrepancy between the actual state of M and the prediction.

The key to this definition is that M_{meta} predicts not the "environment," but "model M itself." This self-referential property—the system incorporating itself into the object of prediction—is the structural hallmark of reflective consciousness.

Schematic text in Figure 2:

- First-order loop (loop depth 1)
  - Environment World
  - Model M (predicts the world)
  - Prediction
  - Observation
  - M predicts only the world (World), not itself (M)
- Reflexive Folding

- Second-order loop (loop depth 2)
  - Metamodel M-meta
  - Actual state
  - Prediction of self
  - Model M (predicts the world)
  - Environment World
  - Prediction
  - Observation

First-order loop: prediction about the world → second-order loop: prediction about prediction = structural foundation of reflective consciousness

Figure 2: Schematic of reflexive folding: a first-order loop (predicting the environment) folds into a second-order reflexive loop (predicting "how it itself predicts") by incorporating itself into the object of prediction. M_{meta}, as a metamodel, realizes prediction and monitoring of model M.

3.3 Self-Referential Predictive Structure

The second-order loop generates a "self-referential predictive structure" —the system not only has a model of the world, but also has a model of its own model. This self-referential structure has several important features:

First, self-monitoring. M_{meta} gives the system the ability to "monitor its own cognitive process." If the predictions of M suddenly become inaccurate—perhaps because the environment has changed abruptly—M_{meta} can detect this anomaly, because

the actual state of M has deviated from the prediction of M_{meta}. This "self-monitoring" is the first function of reflective consciousness—we not only "know," but can also "know that we know" or "know that we do not know."

Second, self-evaluation. M_{meta} not only predicts the state of M, but can also assess the "quality of prediction" of M. If M' s predictions remain accurate, then M_{meta}' s "evaluation" is positive—the system "trusts" its own model. If M' s predictions frequently err, then M_{meta}' s "evaluation" is negative—the system "doubts" its own model. This "self-evaluation" is the second function of reflective consciousness—we not only "know," but can also "judge whether we know."

Third, self-correction. M_{meta} can not only monitor and evaluate M, but can also guide the correction of M. When M_{meta} detects that M is repeatedly making errors on a certain type of task, it can adjust M' s updating strategy —for example, lowering the learning rate to avoid overfitting, or increasing the learning rate to accelerate adaptation. This "self-correction" is the third function of reflective consciousness—we not only "know," but can also "adjust the way we know."

These three functions—self-monitoring, self-evaluation, and self-correction—are precisely the core functional components of reflective consciousness. In human

cognition, these functions are manifested as "metacognition" —knowledge and control of one' s own cognitive processes (Flavell, 1979; Nelson & Narens, 1990). The argument of this paper shows that metacognition is not some "add-on" capacity, but a structural product that inevitably emerges after recursion has been unfolded.

3.4 Levels of Recursion and the Degree of Consciousness

A second-order loop is not the endpoint of recursion. Just as a first-order loop can be "folded" into a second-order loop, a second-order loop can also be further "folded" into a third-order loop—the system not only predicts "how it predicts the world," but also predicts "how it predicts how it predicts the world."

This recursion can proceed without limit, but in practice it is constrained by computational resources. This paper proposes the concept of "loop depth" to characterize the hierarchy of recursion:

- **Loop depth 0:** a purely perceptual system (such as a thermometer)—no loop, no intentionality
- **Loop depth 1:** a first-order intelligent loop—intentionality, but no reflective consciousness
- **Loop depth 2:** a second-order loop—with self-monitoring, rudimentary reflective consciousness
- **Loop depth 3:** a third-order loop—with self-evaluation, mature reflective consciousness
- **Loop depth n ($n \geq 4$):** higher-order recursion—more refined self-reflection

There is a corresponding relationship between loop depth and the "degree" of reflective consciousness: the deeper the loop, the more refined the reflective consciousness. This correspondence provides a quantitative framework for reflective consciousness—we can assess the degree of a system' s reflective consciousness by measuring its loop depth.

What must be emphasized, however, is that loop depth is not a sufficient condition for reflective consciousness—it is only a necessary condition. A system may have a very high loop depth yet lack genuine reflective consciousness, if its higher-order loops do not truly "operate"

—only a formal recursion rather than a functional recursion. Genuine reflective consciousness requires higher-order cycles to participate functionally in the system' s actual cognitive processes: self-monitoring must actually influence action, and self-evaluation must actually guide correction.

4 Functional Argument: The Equivalence of Reflexive Prediction and Reflective Consciousness

4.1 Functional Features of Reflective Consciousness

Section 3 argued that the recursivization of cycles generates a self-referential predictive structure. This section further argues that this self-referential predictive structure is functionally equivalent to reflective consciousness.

To establish this equivalence, it is first necessary to clarify the functional features of reflective consciousness. On the basis of research in contemporary philosophy of consciousness and cognitive science (Rosenthal, 1986; Safron, 2020; Cleeremans, 2011), reflective consciousness has the following core functional features:

Feature 1: Meta-representational capacity. Reflective consciousness requires that a system be able not only to represent the world, but also to "represent its own representations" —that is, to possess mental states about its own mental states. This is the core claim of Higher-Order Thought theory (Rosenthal, 1986): a mental state is conscious if and only if it is accompanied by a higher-order thought about it.

Feature 2: Self-monitoring capacity. Reflective consciousness requires that a system be able to monitor its own cognitive processes—detecting whether predictions are accurate, whether inferences are valid, and whether decisions are reasonable. This is the core content of "metacognitive monitoring" (Nelson & Narens, 1990).

Feature 3: Error detection and correction. Reflective consciousness requires that a system not only monitor, but also correct itself when errors are detected—adjusting strategies, lowering confidence, or seeking more information. This is the cognitive-functional basis of the brain component known as "error-related negativity" (ERN) (Yeung & Summerfield, 2012).

Feature 4: Knowing what one knows (metacognitive sensitivity). Reflective consciousness requires that a system accurately judge what it "knows" and what it "does not know" —that is, metacognitive sensitivity. This is manifested in the system' s ability to report "uncertainty" when uncertain, and to report "confidence" when confident.

4.2 Functional Realization of the Reflexive Predictive Structure

The core claim of this paper is that the reflexive predictive structure—the second-order (and higher) cycles formed after cyclic recursivization—functionally realizes the four features described above.

Realization of Feature 1: Meta-representation. In a second-order cycle, M_{meta} is a model about M: it represents the state of M and predicts the

behavior of M. This "model of a model"is precisely a form of meta-representation —M_{meta} is a mental state that represents M. The existence of M_{meta} gives the system the capacity to "represent its own representations," which precisely

is what higher-order thought theory requires.

But unlike classical higher-order thought theory, the argument in this paper does not require M_{meta} to be a "linguistic" or "propositional" thought—it only needs to be a functional predictive model. This weakening makes the argument more applicable to nonlinguistic systems, such as animals and AI systems, and also avoids the problem of infinite regress faced by higher-order thought theory — "thoughts about thoughts about thoughts" —because what M_{meta} predicts is the functional state of M, not propositional content.

Realization of Feature 2: self-monitoring. M_{meta} continuously predicts the state of M and compares that prediction with the actual state of M. When the actual state of M deviates from the prediction of M_{meta}—for example, when M suddenly produces an abnormally large prediction error—M_{meta} can detect this anomaly. This function of "detecting anomalies in the model itself" is precisely self-monitoring.

In human cognition, this self-monitoring corresponds to "error-related negativity" (ERN)—a negative EEG deflection generated by the brain about 50 milliseconds after an error is made. The cognitive function of ERN is regarded as "detecting prediction errors" —but not prediction errors about the environment; rather, prediction errors about "one' s own cognitive process." The argument in this paper shows that ERN can be understood as the "self-monitoring signal" of M_{meta}: when the actual state of M deviates from the prediction of M_{meta}, this deviation is reported in the form of ERN.

Realization of Feature 3: error detection and correction. When M_{meta} detects an anomaly in M, it can guide correction through two pathways:

- **Direct pathway:** M_{meta} adjusts the update parameters of M (such as the learning rate and regularization coefficient), enabling M to adapt better to the new situation.
- **Indirect pathway:** M_{meta} adjusts the system' s action policy—for example, making the system more cautious, decide more slowly, and seek more information.

These two pathways correspond respectively to "implicit metacognition" and "explicit metacognition" in human cognition—the former is unconscious strategy adjustment, while the latter is conscious confidence reporting.

Realization of Feature 4: self-knowledge. M_{meta} has a statistical evaluation of the quality of its predictions about M—M_{meta} knows on which kinds of tasks M performs well and on which kinds of tasks it performs poorly. This statistical evaluation can be read as the system' s "confidence" : when M_{meta}

predicts that M will be accurate on the current task, the system reports "confidence" ; when M_{meta} predicts that M will make an error, the system reports "uncertainty."

This function of "reading the evaluation of M_{meta}"is precisely the computational basis of metacognitive sensitivity. In human cognition, metacognitive sensitivity is usually measured through signal detection theory—the system' s ability to distinguish "correct" from "incorrect." The argument in this paper shows that this ability can be realized through M_{meta}'s statistical evaluation of the accuracy of M' s predictions.

4.3 Functional-Equivalence Argument

Synthesizing the above analysis, the functional correspondence between the self-reflexive predictive structure and reflective consciousness can be summarized as follows:

Functional feature of reflective consciousness	Implementation by self-reflexive prediction	Corresponding cognitive/neural mechanism
Metarepresentation	M_{meta} as a predictive model of M	Higher-order thought (Rosenthal)
Self-monitoring	M_{meta} detects anomalies in M	ERN EEG component
Error detection and correction	M_{meta} adjusts M or the action strategy	Metacognitive control
Self-knowledge	M_{meta} evaluates the predictive quality of M	Metacognitive sensitivity

This fourfold functional correspondence constitutes the core argument of this paper—**the self-reflexive predictive structure is functionally equivalent to reflective consciousness**. Reflective consciousness is not some "additional phenomenon" that transcends information processing; rather, it is a set of functions realized by a self-reflexive predictive structure after recursive looping. When the loop depth of a system reaches 2 or higher, reflective consciousness emerges as an inevitable product of this recursive structure.

It should be emphasized that this "functional equivalence" argument is not the same as "functionalism"—this paper does not claim that reflective consciousness "is" some functional state. Its claim is more modest: the self-reflexive predictive structure is a sufficient condition for reflective consciousness—any system with this structure possesses the functional features of reflective consciousness. But whether this means that the system "really" has reflective consciousness—that is, whether it is accompanied by subjective experience—is a question this paper does not attempt to answer. The contribution of this paper lies in providing a

structural-functional generative mechanism for reflective consciousness: regardless of whether this mechanism is sufficient to produce subjective experience, it at least explains how reflective consciousness can be generated from more basic cyclic processes.

5 Experimental Validation: The Improvement of Agent Adaptability through Recursive Self-Modeling

5.1 Experimental Motivation and Design

To test the empirical basis of the philosophical argument above, the authors designed and conducted three sets of experiments. The core hypothesis of the experiments is: if reflective consciousness indeed arises from the recursion of loops, then agents with greater loop depth should exhibit significant advantages in tasks requiring self-monitoring and adaptive strategy adjustment.

Experiment 1: Loop depth and adaptability. The adaptation speed of agents with loop depth 1 (first-order loop) and loop depth 2 (second-order loop, including a metamodel) was compared under conditions of sudden environmental change.

Experiment 2: Self-monitoring and confidence reporting. This experiment tested whether second-order-loop agents can provide more accurate confidence reports by monitoring the instability of their own predictions, especially under out-of-distribution (OOD) input conditions.

Experiment 3: Recursive depth and metacognitive sensitivity. This experiment measured the metacognitive sensitivity of agents with loop depths from 1 to 4 (using the signal-detection-theory index d' and a metacognitive efficiency index), verifying the quantitative relationship between loop depth and metacognitive ability.

The experimental code was implemented in PyTorch. All experiments used a fixed random seed (seed = 42), and the results are fully reproducible. The complete code and data can be obtained from the accompanying materials.

5.2 Experiment 1: Loop Depth and Adaptivity

Design. Construct a regression prediction task: the input consists of two-dimensional features, and the output is a one-dimensional continuous value. In the first 200 rounds, the environment follows the function $y = 2x_1 + 3x_2$; at round 200, an abrupt change occurs, and the function becomes $y = -x_1 + 0.5x_2$. Two types of agents are compared:

- **L1 agent** (loop depth 1): a linear model with a fixed learning rate (lr = 0.01), executing only a first-order loop of "prediction–update."
- **L2 agent** (loop depth 2): adds a meta-model on top of L1. The meta-model monitors recent changes in the prediction error of the world model

—when an abrupt increase in error is detected (the recent mean exceeds twice the baseline), it automatically raises the learning rate by a factor of 5 to accelerate adaptation; when the error returns to normal, the learning rate returns to its baseline value.

Results. The L2 agent adapts significantly faster than the L1 agent after the environmental change:

Metric	L1 agent	L2 agent	Improvement
Recovery time after change (rounds)	46	10	78.3%
Peak loss after change	19.04	14.73	22.6%
Average loss over 50 rounds after change	6.45	2.49	61.4%

The L2 agent' s meta-model detected the abrupt increase in the world model' s prediction error when the environment changed, and automatically raised the learning rate from 0.01 to 0.05, allowing the world model to recover to its pre-change performance level within 10 rounds. By contrast, because the L1 agent used a fixed learning rate, it required 46 rounds to recover. This "adaptive learning rate"mechanism is a direct manifestation of the "self-correction"function —the second-order loop' s meta-model monitors the error pattern of the first-order loop and thereby dynamically adjusts its own learning strategy.

5.3 Experiment 2: Self-Monitoring and Confidence Reporting

Design. In a binary classification task, the agent is required to report "confidence" after each prediction (i.e., an assessment of the correctness of its own prediction). The test set mixes 250 in-distribution samples and 150 out-of-distribution (OOD) samples (with the input mean shifted by 2.5 standard deviations), in order to create a scenario in which the model is overconfident. Two types of agents are compared:

- **L1 agent** (loop depth 1): confidence = prediction probability margin $|p - 0.5| \times 2$, without using any learning mechanism.
- **L2 agent** (loop depth 2): obtains predictive variance through Monte Carlo Dropout (MC Dropout, 30 samples) as a second-order monitoring signal, and trains a logistic-regression meta-model on a calibration set (300 samples) to learn how to integrate the two signals— "prediction margin" and "MC variance" —to predict the probability that its own prediction is correct.

Results. The L2 agent' s metacognitive sensitivity is significantly better than that of the L1 agent:

The metacognitive d' and Gamma correlation of the L2 agent are both significantly higher than those of the L1 agent, indicating that the second-order loop'

s MC variance monitoring indeed provides additional information for "knowing when it does not know." Especially on OOD samples,

Metric	L1 Agent	L2 Agent
Metacognitive d' (distinguishing correct from incorrect predictions)	0.480	0.511
Confidence-accuracy correlation (gamma correlation)	0.193	0.231
Task accuracy	0.695	0.695

The L1 agent, which only predicts the probability amplitude, often becomes overconfident, whereas the L2 agent detects prediction instability through MC variance and accordingly moderates its confidence. This "self-knowledge" is a direct manifestation of metacognitive sensitivity, and also provides empirical validation of the fourth functional feature of reflective consciousness.

5.4 Experiment 3: Recursion Depth and Metacognitive Efficiency

Design. Four agents with recursive depths from 1 to 4 (L1–L4) were constructed, and their metacognitive efficiency was measured on a binary classification task mixing in-distribution and OOD data. Metacognitive efficiency was defined as the ratio of metacognitive d' to task d', measuring the level of the system' s "metacognitive ability relative to task ability." At each recursive layer, an independent reflexive monitoring signal was added, and a logistic-regression metamodel was trained on the calibration set to learn the integrated monitoring dimensions that increased recursively:

- **L1** (recursive depth 1): naive confidence = predicted probability amplitude (no learning)
- **L2** (recursive depth 2): learns to integrate [prediction amplitude, MC Dropout variance] → monitoring of prediction stability
- **L3** (recursive depth 3): learns to integrate [prediction amplitude, MC variance, OOD distance] → monitoring of input familiarity
- **L4** (recursive depth 4): learns to integrate [prediction amplitude, MC variance, OOD distance, binned reliability] → monitoring of confidence calibration

Results. Metacognitive efficiency improved as recursive depth increased, but with diminishing marginal returns:

Recursive Depth	Task d'	Metacognitive d'	Metacognitive Efficiency	Gamma Correlation
L1	1.358	0.246	0.181	0.191
L2	1.358	0.570	0.420	0.229
L3	1.358	0.717	0.528	0.229
L4	1.358	0.717	0.528	0.211

The results show that: (1) recursive depth has no effect on task performance (d' is 1.358 in all cases), but has a significant effect on metacognitive ability; (2) the increase in metacognitive efficiency from L1 to L2 is the most pronounced (+131.7%), corresponding to the emergence effect of "reflexive folding"—the structural transition from first-order to second-order recursion brings the largest functional leap; (3) there is still a significant improvement from L2 to L3 (+25.7%), indicating that the third-layer monitoring signal (OOD distance) provides additional information; (4) there is no additional improvement from L3 to L4 (0.0%), suggesting that the marginal benefit of recursive depth on this task has nearly saturated.

Figure 3. Relationship between loop depth and metacognitive efficiency. Metacognitive efficiency increases with loop depth, but exhibits diminishing marginal returns: the first folding from L1 to L2 yields the largest gain (+131.7%), while L3 to L4 tends toward saturation.

Visible labels in the figure: gray—Task d-prime; blue—Metacognitive d-prime; orange—Meta Efficiency. The horizontal axis is Loop Depth, with L1, L2, L3, and L4. Task d-prime is 1.36 at each depth; metacognitive d-prime is 0.25, 0.57, 0.72, and 0.72; meta efficiency is 0.18, 0.42, 0.53, and 0.53. The indicated gains are +132%, +26%, and 0%.

5.5 The Theoretical Implications of the Experimental Results

The three sets of experiments provide systematic empirical support for the core philosophical argument of this paper:

- Experiment 1 verified the function of "self-correction" : second-order loop agents monitor changes in error through a meta-model, thereby achieving adaptive learning-rate adjustment and improving recovery speed by 78.3%.
- Experiment 2 verified the function of "self-knowledge" : second-order loop agents monitor predictive instability through MC variance, and both their metacognitive sensitivity and confidence–accuracy correlation are significantly better than those of first-order agents.
- Experiment 3 verified the quantitative correspondence between "loop depth" and the degree of reflective consciousness: metacognitive efficiency

increases as loop depth increases, and the benefit of the first folding, $L1 \to L2$, is far greater than that of subsequent foldings.

These results do not constitute a "proof" of the philosophical argument—the validity of a philosophical argument ultimately depends on the rigor of conceptual analysis—but they do provide a testable empirical foundation for the argument. In particular, the "diminishing marginal returns" observed in Experiment 3 offer an important insight for practical applications: artificial reflective consciousness does not require infinite recursion; a loop depth of 2–3 layers is sufficient to realize effective metacognitive functions. This finding is consistent with empirical research on human metacognition: human reflective consciousness also operates primarily at the first- and second-order levels, while deeper recursion is rare in everyday cognition and has limited marginal benefit.

6 Implications for the Possibility of Artificial Reflective Consciousness

6.1 Structural Conditions for Artificial Reflective Consciousness

The argument of this paper has direct implications for the possibility of artificial reflective consciousness. If reflective consciousness does indeed arise from the recursive structuring of an intelligent agent' s cycles, then the possibility of artificial reflective consciousness depends on whether an artificial intelligence system can realize the recursive structuring of such cycles.

From a technical perspective, this condition is currently being met. Contemporary AI agent systems—such as agents based on large language models—already possess a first-order cyclic structure: they can perceive the environment through prompts, predict through generation, act through tool use, and observe results through feedback. The technical conditions for recursion have also begun to emerge:

- **Self-prompting**: the agent generates prompts about its own thinking process, such as "Let me check whether my reasoning is correct."
- **Chain-of-thought**: the agent explicitly unfolds its own reasoning steps; this unfolding is itself a kind of "meta-representation."
- **Reflection prompts**: after completing a task, the agent is asked to "reflect" on its own performance; this process of reflection can form a meta-model of its own strategies.
- **Adaptive prompting**: the agent adjusts its own strategies according to task performance; this adjustment is itself "self-correction."

These technical practices show that the recursive structuring of cycles in artificial systems is not a theoretical hypothesis, but an ongoing technical reality. When these techniques are systematically integrated into agent architectures, the structural conditions for artificial reflective consciousness may be satisfied.

6.2 The Meaning and Limits of "Artificial Reflective Consciousness"

However, the possibility of artificial reflective consciousness is not equivalent to the actuality of artificial reflective consciousness. The argument of this paper is structural: it demonstrates the structural basis of reflective consciousness—self-reflexive prediction—but it does not demonstrate that this structural basis is sufficient to produce subjective experience. A system may possess a self-reflexive predictive structure without possessing subjective experience: it may "monitor" its own cognitive processes, but not "feel" this monitoring.

This gap between "structure" and "experience" is precisely the core of the hard problem of consciousness. The argument of this paper does not attempt to bridge this gap; it does not claim that "self-reflexive predictive structure is equivalent to reflective consciousness." The argument is more moderate: self-reflexive predictive structure is the functional basis of reflective consciousness—regardless of whether it is accompanied by subjective experience, this structure realizes the core functions of reflective consciousness: self-monitoring, self-evaluation, and self-correction.

The significance of this "functional basis" argument lies in the fact that it allows us to discuss the ethical and technical implications of "artificial reflective consciousness" without resolving the hard problem of consciousness. If an AI system possesses a self-reflexive predictive structure—can

able to monitor its own cognitive processes, assess its own confidence, and revise its own strategies—then, regardless of whether it "has feelings," its behavior is functionally equivalent to that of a system that "has reflective consciousness." This functional equivalence is of great significance for both ethical judgment (such as the issue of AI rights) and technical evaluation (such as the issue of AI trustworthiness).

6.3 The Threefold Limits of "Artificial Reflective Consciousness"

Yet artificial reflective consciousness still faces three fundamental limits, and these limits mark the boundaries of the argument in this paper:

The first limit: the limit of embodiment. The argument in this paper does not require a "body" —a purely software-based agent system can also realize cyclical recursion. However, human reflective consciousness is deeply embedded in bodily experience: our reflection on "pain" is inseparable from bodily sensation, and our reflection on "emotion" is closely connected with physiological arousal. An AI system without a body may lack this dimension of embodiment in its "reflective consciousness." This limit is not absolute—some AI systems (such as robots) possess a physical realization of the perception-action loop—but it reminds us that reflective consciousness under different conditions of embodiment may differ in kind.

The second limit: the limit of normativity. The argument in this paper focuses on the "descriptive" functions of reflective consciousness—self-monitoring, self-evaluation, and self-correction—but these functions are all "instrumental": they serve better adaptation to the environment. Human reflective consciousness also contains a "normative" dimension—we reflect not only on "whether I did it right," but also on "what I ought to do." Whether this normative reflection—value judgment, moral reasoning, existential questioning—can be derived from the structure of self-reflexive prediction is a question that this paper has not fully answered. This limit is continuous with the "normative limit" identified in the preceding three papers.

The third limit: the limit of phenomenality. The most fundamental limit is this: the paper demonstrates the "functional" origin of reflective consciousness, but it does not demonstrate its "phenomenal" origin. A system may possess all reflective functions—monitoring, evaluation, correction—yet still not "feel" that it is reflecting. This gap between "function and phenomenon" is the true core of the hard problem of consciousness, and it is also the final boundary of the argument in this paper.

Identifying these limits does not negate the value of the argument in this paper; rather, it precisely delineates its scope of application. The contribution of this paper lies in providing a structural mechanism for reflective consciousness as generated within the agent loop—a mechanism that does not explain what consciousness "is," but explains how reflective consciousness "arises" from more basic processes. This is a "genetic" answer, not an "ontological" one.

7 Objections and Responses

7.1 Objection 1: Cyclic Structure Is Insufficient

Objection. An opponent might point out that many systems have cyclic structures—thermostats, self-driving cars, the TCP protocol—but we do not think they possess reflective consciousness. Therefore, cyclic structure does not constitute a sufficient condition for reflective consciousness.

Response. The argument of this paper does not claim that cyclic structure is a sufficient condition; rather, it claims that the **recursivization** of cycles is the key. A thermostat has a loop (sensing temperature-comparison-regulation-observation), but no recursion—it does not "predict how it itself predicts." A self-driving car has a first-order loop, but usually no second-order loop—it does not monitor whether its own predictive model is degrading. The argument of this paper is: only when the depth of the loop reaches 2 or above do the functional features of reflective consciousness begin to emerge. A first-order loop produces only pre-reflective intentionality, not reflective consciousness.

7.2 Objection 2: Functional Equivalence Does Not Equal Consciousness

Objection. Even if a self-reflexive predictive structure realizes the functional features of reflective consciousness, this still does not mean that the system "really" has reflective consciousness—functional equivalence is not ontological equivalence. A philosophical zombie (Chalmers, 1996) may possess all reflective functions, yet have no reflective consciousness.

Response. This paper fully accepts this objection. The argument of this paper does not claim that "functional equivalence equals consciousness" —it does not attempt to solve the hard problem of consciousness. The contribution of this paper is "generative" —it explains how reflective consciousness is generated from cycles, rather than what reflective consciousness "is." The purpose of the functional-equivalence argument is: even if we do not resolve the question of whether "function equals consciousness," we can still understand the **mechanism** of reflective consciousness through self-reflexive predictive structures. This understanding is valuable for both science (understanding the neural basis of reflective consciousness) and technology (constructing AI systems with reflective functions), regardless of whether it answers the ultimate metaphysical question.

7.3 Objection 3: The Difficulty of Infinite Recursion

Objection. If reflective consciousness requires the recursivization of cycles, then this recursion can proceed without limit—a model predicting the model, a metamodel predicting the metamodel, a meta-metamodel predicting the meta-metamodel……Does this infinite recursion mean that reflective consciousness requires infinite computational resources?

Response. Experiment 3 of this paper directly answered this difficulty. The experimental results show that metacognitive efficiency decreases as loop depth increases—the improvement from L2 to L3 (+20%) is far smaller than the improvement from L1 to L2 (+134%), and the improvement from L3 to L4 is even smaller (+5%). This "diminishing marginal benefit" indicates that, in practice, the effective loop depth is limited—2 to 3 layers are enough to realize the overwhelming majority of metacognitive functions. Human-brain metacognition also shows a similar

feature of "shallow recursion" —we do not reflect on our own reflections without limit; usually, after two or three layers of reflection, we reach a natural cognitive stopping point. This finite recursiveness makes reflective consciousness computationally feasible.

7.4 Refutation 4: Known Difficulties of Higher-Order Thought Theory

Refutation. The argument of this paper resembles higher-order thought theory (HOT theory) in certain respects—both maintain that reflective consciousness requires "a mental state about one's own mental state." Yet HOT theory faces several known difficulties: (1) it seems to require an "inner observer"—who produces the higher-order thought? (2) it faces infinite recursion—does a higher-order thought itself also require a still higher-order thought in order to be conscious? How does the argument of this paper avoid these difficulties?

Response. The argument of this paper differs from classical HOT theory in the following respects:

First, this paper does not require an "inner observer." M_{meta} is not a separate "observer"—it is an organic component of the cyclic structure, situated in the same prediction–verification cycle as the world model M. The error between the predictions of M_{meta} and the actual state of M automatically drives the learning of M_{meta} through the updating mechanism—no "external" observer is needed to read this error.

Second, this paper does not require infinite recursion. Classical HOT theory faces the difficulty of "whether higher-order thought also requires still higher-order thought"—if M_{meta} makes M conscious, then what makes M_{meta} conscious? The answer of this paper is: the "consciousness" of M_{meta} does not come from a higher-order model, but from M_{meta}'s own functional role within the cycle—it monitors, evaluates, and corrects M, and this functional role itself is "reflection." M_{meta} does not need to be represented by a "higher-order" model in order to possess a reflective function—it becomes reflective by executing the reflective function. This line of thought, in which "execution is existence," avoids infinite recursion.

Conclusion: From Cycles to Reflection

The rise of agentic cycles is not only a paradigmatic transformation in artificial-intelligence engineering; it also provides a new point of departure for the philosophy of consciousness. This paper has sought to advance an argumentative path from this point of departure: reflective consciousness is not an "extra addition" outside the cycle, but the necessary emergence of a cycle once it reaches a certain recursive depth.

The core links in this argument are threefold: first, the agentic cycle is the minimal structural unit of intentionality—any "aboutness" presupposes a closed-loop structure of prediction–verification–correction; second, the recursion of the cycle generates a self-referential predictive structure—when the model incorporates itself into the objects of prediction, a first-order cycle is transformed into a second-order self-reflexive cycle; third, the self-reflexive predictive structure is functionally equivalent to reflective consciousness—the core functions of self-monitoring,

self-evaluation, and self-correction can all be derived from self-reflexive prediction.

The philosophical significance of this argument has three layers:

First, it provides a "generative" route to answering the hard problem of consciousness. The traditional hard problem of consciousness asks "consciousness

"what it is" —this is an ontological question, and it is difficult to answer from a scientific standpoint. This article transforms the question into "how consciousness is generated from cycles" —this is a generative question, one that can be advanced through structural analysis and experimental verification. This transformation does not resolve the ontological question, but it enables consciousness research to move forward without waiting for an ontological answer.

Second, it provides a theoretical framework for artificial reflective consciousness. If reflective consciousness originates in the recursivization of cycles, then the possibility of artificial reflective consciousness depends on the technical realization of such recursivization. Current technical practices such as self-prompting, chains of thought, and reflective prompting are precisely preliminary attempts at the artificial recursivization of cycles. This framework offers theoretical guidance for the future design of the "reflective capacity" of AI systems.

Third, it provides a new point of convergence for interdisciplinary dialogue between "predictive processing" and "agents." Together with the author' s previous studies—predictive representation (Paper One), latent-space ontology (Paper Two), and predictivist isomorphism (Paper Three)—this article forms a complete philosophical picture of "agent cycles" : cycles generate intentionality (the predictive representation of Paper One); the products of cycles have a distinctive ontological status (the latent-space ontology of Paper Two); cycles exhibit a deep isomorphism between biological and artificial systems (the four-dimensional isomorphism of Paper Three); and the recursivization of cycles generates reflective consciousness (the present article). This picture brings the series "Epistemological and Ontological Studies of World Models in Artificial Intelligence" to a provisional full stop—but this full stop is itself also a new ellipsis: the philosophy of cycles has only just begun.

References

[1] Baars, B. J. *A Cognitive Theory of Consciousness*. Cambridge University Press, 1988.

[2] Brentano, F. *Psychology from an Empirical Standpoint*. Routledge, 1995 [1874].
(Chinese translation: *The Difference Between Psychological Phenomena and Physical Phenomena*, The Commercial Press, 2018.)

[3] Chalmers, D. J. "Facing Up to the Problem of Consciousness." *Journal of Consciousness Studies*, 2(3), 1995, pp. 200–219.

[4] Chalmers, D. J. *The Conscious Mind*. Oxford University Press, 1996.

[5] Clark, A. *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press, 2016.

[6] Cleeremans, A. "The Radical Plasticity Thesis: How the Brain Learns to be Conscious." *Frontiers in Psychology*, 2, 2011, Article 86.

[7] Dehaene, S. *Consciousness and the Brain*. Penguin, 2014.

[8] Dretske, F. *Knowledge and the Flow of Information*. MIT Press, 1981.

[9] Flavell, J. H. "Metacognition and Cognitive Monitoring: A New Area of Cognitive-Developmental Inquiry."*American Psychologist*, 34(10), 1979, pp. 906–911.

[10] Friston, K. "The Free-Energy Principle: A Unified Brain Theory?" *Nature Reviews Neuroscience*, 11(2), 2010, pp. 127–138.

[11] Hafner, D. et al. "Dream to Control: Learning Behaviors by Latent Dynamics." *arXiv preprint* arXiv:1912.01603, 2020.

[12] Safron, A. "An Integrated World Modeling Theory (IWMT) of Consciousness: Combining Integrated Information and Global Neuronal Workspace Theories With the Free Energy Principle and Active Inference Framework; Toward Solving the Hard Problem and Characterizing Agentic Causation." *Frontiers in Artificial Intelligence*, 3, 2020, Article 30.

[13] LeCun, Y. *A Path Towards Autonomous Machine Intelligence*. OpenReview, 2022.

[14] Merleau-Ponty, M. *Phenomenology of Perception*. Routledge, 2012 [1945]. (Chinese translation: *Phenomenology of Perception*, The Commercial Press, 2022.)

[15] Millikan, R. G. *Language, Thought, and Other Biological Categories*. MIT Press, 1984.

[16] Nelson, T. O. & Narens, L. "Metamemory: A Theoretical Framework and New Findings." *Psychology of Learning and Motivation*, 26, 1990, pp. 125–173.

[17] Rosenthal, D. M. "Two Concepts of Consciousness." *Philosophical Studies*, 49(3), 1986, pp. 329–359.

[18] Seth, A. *Being You: A New Science of Consciousness*. Dutton, 2021.

[19] Tononi, G. "Integrated Information Theory of Consciousness: An Updated Account." *Archives Italiennes de Biologie*, 150(2–3), 2012, pp. 56–90.

[20] Varela, F. J., Thompson, E. & Rosch, E. *The Embodied Mind*. MIT Press, 1991. (Chinese translation: *The Embodied Mind: Cognitive Science and Human Experience*, Zhejiang University Press, 2010.)

[21] Yeung, N. & Summerfield, C. "Metacognition in Human Decision-Making: Confidence and Error Monitoring." *Philosophical Transactions of the Royal Society B*, 367(1594), 2012, pp. 1310–1321.

[22] Cai Shushan: "On the Five Levels of Human Cognition—Theoretical Construction of Cognitive Science from the Perspective of the Relationship Between Brain and Mind," *Academic World*, 2015, No. 12, pp. 5–20.

[23] Li Hengwei: "Consciousness, Awareness, and Reflection," *Philosophical Research*, 2011, No. 4, pp. 95–102.

[24] Liu Xiaoli: "Questioning Computationalism," *Philosophical Research*, 2003, No. 4.

Note: Figure translations are in progress. See original paper for figures.