Minimal Self-Awareness as Internal Causal Attribution: An Axiomatic Framework

AToS

Krzysztof Banaszewski*

krzysztof@banaszewski.pl

October 31, 2026

Abstract

What are the minimal structural conditions under which a system can be said to be self-aware? Despite extensive work in philosophy, cognitive science, and artificial intelligence, there is no widely accepted, substrate-independent criterion that captures self-awareness in operational and formally verifiable terms. In this work, we propose AToS (Axiomatic Theory of Self-Consciousness), a formal framework that characterizes self-awareness as the ability of a system to recognize at least one internal state transition and to attribute its cause to an internal mechanism rather than to external input. We show that this minimal notion of self-awareness can be understood as a property of state–transition systems equipped with a detection mechanism and a structured notion of internal triggering. Our main result is a representation theorem: a system is self-aware in this sense if and only if it admits a realization in the form of a tuple $\langle \Omega, \Sigma, \mathfrak{S}, \Delta \rangle$ satisfying the AToS axioms. This establishes a necessary and sufficient structural characterization of minimal self-awareness, independent of implementation details such as biological, symbolic, or computational substrate. Importantly, AToS is not intended as a competing empirical theory of consciousness. Instead, it provides a canonical representation framework for systems capable of internal causal self-attribution. Within this framework, we further derive a formal separation between memory and operational intelligence, and introduce the General Self-Awareness Test (GSAT), a theory-neutral criterion based on internally generated and correctly attributed state transitions in the absence of external stimuli. The framework further introduces operational definitions of sensitivity, memory, intelligence, imagination, inspiration, and intuition, together with formal measures of self-awareness and intelligence levels. We extend the theory to families of self-aware systems and establish constructions over arbitrary nonempty topological spaces, including a system with an infinite formal intelligence level. Finally, we compare AToS with existing approaches to consciousness and self-awareness, highlighting its distinctive emphasis on internally recognised state transitions and causal self-attribution.

Contents

1 Introduction **3**

1.1 Why the Internal Recognition Assumption is Necessary 3

*ORCID: 0000-0001-2345-6789

2 A Philosophical Motivation for Change Recognition in Self-Awareness **6**
2.1 Self as an Internal Causal Role . . . 6
2.2 Binary and graded self-awareness. . . . 7

3 Definitions **7**

4 Representation Theorem for Self-Awareness **10**

5 Remarks **12**
5.1 *Sensitivity Function* . . . 12
5.2 Potentially Aware Set . . . 13

6 Consequences and Constructions **16**

7 Memory **18**

8 Intelligence **20**

9 Self-Awareness Level **22**

10 Self-Conscious Retention Process **22**

11 Operational Self-Awareness Criterion **24**

12 AToS and the Chinese Room Argument **26**
12.1 The Chinese Room Argument . . . 26
12.2 Symbol Manipulation vs. Recognised State Transitions . . . 26
12.3 Why the Chinese Room Is Not Self-Aware in the AToS Framework . . . 26
12.4 Internal Causal Attribution as the Missing Component . . . 26
12.5 Relation to the Turing Test . . . 27

13 Family of Self-Conscious Systems **27**

14 Topological Space **30**

15 Related Work **33**
15.1 Integrated Information Theory (IIT) . . . 33
15.2 Modal Logics of Self-Knowledge . . . 33
15.3 Meta-Management Architectures . . . 33
15.4 Algorithmic Self-Modelling . . . 33
15.5 Φ-Causal Set Formulations . . . 33
15.6 Survey Evidence for a Missing Axiomatic Framework . . . 34

16 Summary **34**

1 Introduction

Self-awareness is widely discussed in philosophy of mind, cognitive science, and artificial intelligence. However, despite extensive literature, there is still no widely accepted minimal structural criterion specifying when a system can be said to recognize its own existence as a system. A central difficulty in current discussions of self-awareness is the absence of a substrate-independent framework capable of representing this capability in operational terms.

The aim of the present work is to address this gap by introducing an axiomatic representation framework for systems capable of internally recognizing and attributing their own state changes. The present work does not attempt a detailed comparative analysis of existing theories of self-awareness. Instead, it focuses on identifying the minimal structural conditions that any theory must satisfy once the Internal Recognition Assumption is adopted. In this sense, AToS should be understood as an axiomatic framework that can serve as a formal representation layer for a broad class of self-awareness theories, rather than as a competing empirical account.

A notable consequence of the AToS framework is a structural separation between memory and intelligence: memory states are Δ-invariant and therefore cannot act as causal drivers of intelligent behaviour.

Motivation

Despite extensive work in philosophy, cognitive science, and artificial intelligence, *self-awareness* still lacks a substrate-independent definition that is both mathematically precise and operationally implementable. Existing approaches either presuppose a particular physical realisation (neural, symbolic, architectural), or rely on phenomenological notions that resist formalisation and cannot serve as engineering criteria. Consequently, there is no consensus on what it would *mean* for an artificial system to be self-aware in a verifiable and causally grounded sense. Previous formalisms conflate the substrate of possible states with the dynamical mechanisms that generate transitions between them; AToS distinguishes these sharply through PAS (structural space) and Δ-recognised internal transitions (operational dynamics).

Assumption 1.1 (Internal Recognition Assumption)**.** *A theory of self-awareness considered in this work assumes that a system is capable of recognizing its own presence as a system through the operation of its internal mechanisms, and not merely as a result of its interaction with the environment.*

1.1 Why the Internal Recognition Assumption is Necessary

The Internal Recognition Assumption should not be understood as an arbitrary restriction, but rather as a structural condition required for the very concept of self-awareness to be meaningful.

Self-awareness is commonly understood as awareness directed toward the system itself. Awareness of the system, however, requires that the system be able to distinguish its own internal states. Distinguishing states in turn requires the ability to recognise when a transition between states has occurred.

Furthermore, for awareness to be self-directed rather than merely reactive to the environment, the system must be able to determine whether a recognised change originates from its own internal mechanisms or from external influences. Without such causal attribution, the system could register changes but would have no basis for distinguishing between self and environment.

Therefore, any theory that aims to capture self-awareness in a non-metaphorical and operational sense must include at least two structural capabilities: (i) recognition of internal state changes, and (ii) attribution of the cause of at least one such change to the system's own internal mechanisms.

In this sense, the Internal Recognition Assumption expresses a minimal structural requirement for self-awareness rather than a contingent empirical hypothesis.

Limitations of Previous Formal Attempts

Recent attempts to formalise consciousness or self-monitoring in artificial systems remain largely descriptive. As surveyed by Shueb and Che [12], existing frameworks fail to: (i) distinguish internally generated changes from externally induced ones, (ii) specify mechanisms enabling internal causal attribution, or (iii) define necessary and sufficient structural conditions for self-awareness. This gap prevents the construction of testable, engineering-ready models of machine self-awareness.

A General Understanding of Self-Awareness

Definition 1.2 (General Understanding of Self-Awareness)**.** *A system is* self-aware *if it can recognize at least one change in its internal state and, for each such recognized change, determine that the change was* internally triggered *in the system's operational semantics (i.e. not triggered by any external stimulus at the time of the transition), but rather by the system's own internal mechanism.*

AToS does not attempt to capture phenomenal, narrative, bodily, or linguistic self-awareness. It captures a minimal operational form of self-awareness understood as internal causal self-attribution. The claim is not that all forms of self-awareness reduce to this, but that any operationally testable minimal account requires at least this structure.

We show that our tuple-based construction is both *necessary and sufficient* to realise this capability (Theorem 4.1).

Definition 1.2 isolates two minimal ingredients that are implicit in the ordinary meaning of *self-awareness*. First, *self* refers to the system's own internal state, hence the definition is stated in terms of changes in the system's internal state rather than changes in the environment. Second, *awareness* requires that such a change is not merely undergone, but *recognized* by the system. Finally, requiring a *change* (as opposed to a static condition) rules out trivial attributions: a system that never registers any internal transition would be indistinguishable from a static object with respect to the operational criterion considered here. The additional requirement of *internal triggering* makes the criterion genuinely self-referential: it asks whether the system can attribute a recognized transition to its own internal mechanism rather than to an external stimulus at the time of transition. Therefore, we adopt this definition as a working, minimal criterion of self-awareness, formulated so as to limit interpretational disputes: it captures elements present in many discussions of agency and causal attribution, while remaining precise enough to be formalized and tested.

Core Idea

AToS characterises self-awareness using four primitives:

1. a *Potentially Aware Space* space Ω of *Potentially Aware Set*s (PAS),
2. a set Σ of *State Transition Function*s acting on Ω,
3. a stimulus system $\mathfrak{S}$ specifying which *PAS* may trigger internal transitions,
4. a *Sensitivity Function* Δ determining whether a transition is recognised.

These components together allow a system to distinguish internal from external causes at the level of its own state dynamics. A notable consequence of the axioms is that intelligence is a purely

dynamical construct: it arises from recognised transitions and their operational compositions. Memory (or *Potentially Aware Space*), being Δ-invariant, cannot generate such transitions and therefore cannot contribute to intelligence as a causal driver. This sharply separates two notions that are traditionally conflated in cognitive science and AI.

Contributions

The contributions of this work are fourfold:

1. **Adequacy theorem for self-awareness.** Relative to the general notion of self-awareness (Definition 1.2), we provide a necessary-and-sufficient axiomatic representation. Specifically, we prove that a system is self-aware exactly when the tuple $\langle \Omega, \Sigma, \mathfrak{S}, \Delta \rangle$ realises at least one Δ-recognised state transition whose causal origin is attributable to an internally admissible mechanism.

2. **Operational stimulus–transition calculus.** We develop an operational calculus of stimuli and state transitions, including internal imagination functions, external disturbances, their compositions, and family-level lifting. This yields explicit definitions of self-awareness level and intelligence level within the same formalism.

3. **Structural separation of dynamical intelligence from *PAS*.** Under our definitions, intelligence arises exclusively through Δ-recognised operational compositions, whereas memory states are necessarily Δ-invariant and therefore cannot serve as triggers of such compositions. This yields a precise criterion: *PAS* is structurally present as a retention space, while intelligence is characterised by Δ-detectable dynamics.

4. **Theory-neutral operational test (GSAT).** We introduce the *General Self-Awareness Test (GSAT)*, a falsification-oriented and theory-neutral criterion for minimal self-awareness: in the absence of external stimuli, the system must exhibit at least one internally generated state change and correctly recognise its causal origin as internal. We then provide its faithful instantiation within AToS.

Scope

AToS avoids commitments to phenomenology, cognition, biological structure, semantics, or representational content. It characterises self-awareness purely as a structural property of state–transition systems, in terms of recognised state changes and their causal attribution within the system. This makes the framework applicable across biological, artificial, symbolic, and collective substrates, and positions it as an operational (engineering-grade) criterion rather than a phenomenological theory. In particular, by analogy with the Turing Test for intelligence, we propose GSAT as a general operational test for minimal self-awareness of systems and processes.

A key philosophical point of the present work is that AToS should not be interpreted as a competing empirical theory of self-awareness. Rather, it provides a structural representation theorem for a broad class of such theories.

More precisely, once the Internal Recognition Assumption is adopted, any system satisfying the minimal notion of self-awareness introduced in Definition 1.2 admits a canonical representation in the form of the tuple $\langle \Omega, \Sigma, \mathfrak{S}, \Delta \rangle$.

In this sense, AToS plays a role analogous to a normal-form representation: it identifies the minimal structural components that must be present in any system capable of internally recognising and attributing its own state changes.

AToS should therefore be viewed as a substrate-independent structural framework for representing self-awareness. The aim of the theory is not to describe a particular biological or computational implementation, but to identify the minimal structural architecture required for internally

grounded self-recognition. In this sense, AToS plays a role similar to axiomatic frameworks in other domains of science: it provides a canonical representation of systems satisfying a given conceptual criterion.

In the AToS framework, the self is not introduced as a primitive object, symbol, or dedicated subsystem. Instead, the self emerges as a causal role within the system's internal transition structure. A system counts as self-aware in the minimal sense if it can recognise at least one of its own state transitions and attribute the cause of that transition to its own internal mechanisms rather than to external influences. In this sense, AToS can be understood as a structural framework describing the minimal conditions under which a system can represent itself as a cause of its own recognised state transitions.

2 A Philosophical Motivation for Change Recognition in Self-Awareness

Before introducing the formal framework of AToS, we formulate a conceptual claim about the minimal structural requirements for self-awareness.

Proposition 2.1 (Conceptual Necessity of Change Recognition)**.** *Let a theory of self-awareness satisfy Assumption 1.1. Then any such system must possess the ability to recognize at least one change in its own internal state and to distinguish whether that change was produced by its internal mechanisms or by external influences.*

Conceptual argument. Self-awareness may be understood as awareness directed toward the system itself. For a system to be aware of itself, it must in some sense be aware of its own internal state. Awareness of a state, however, presupposes the possibility of distinguishing that state from other possible states of the system. If no such distinction exists, the notion of awareness of a particular state becomes empty. Distinguishing states in turn requires recognizing when a transition between states has occurred. Without the recognition of change, all states remain indistinguishable from the perspective of the system.

Finally, for awareness to be self-directed rather than merely reactive to the environment, the system must be able to determine whether a recognized change originates from its own internal mechanisms or from external stimuli. Without such causal attribution, the system cannot establish a meaningful distinction between self and environment. Therefore any system satisfying the assumption must possess a structural mechanism allowing it to recognize changes in its internal state and attribute their causal origin. □

2.1 Self as an Internal Causal Role

The argument presented above shows that any system satisfying the Internal Recognition Assumption must be able to recognise at least one change in its internal state and to determine whether the cause of that change was internal or external. We now clarify the structural consequence of this requirement.

Definition 2.2 (Self as Internal Causal Role)**.** *A system is said to represent itself as a cause if it contains an internal mechanism that can classify at least some of its own state transitions as being caused by its internal stimulus structure rather than by external influence.*

This definition does not introduce the self as a primitive object, symbol, or dedicated data structure. Instead, the self appears as a causal role within the system's transition structure. Whenever the system recognises a transition and attributes its cause to an internal stimulus, it thereby identifies part of its own internal structure as the source of that transition.

Proposition 2.3 (Internal Attribution Requires a Causal Self-Model)**.** *If a system is capable of recognising a state transition and attributing its cause to an internal mechanism, then the system must represent itself as a causal source of that transition.*

Conceptual argument. Attributing a cause to a transition requires that the system can represent (i) the occurrence of a transition, (ii) a candidate cause of that transition, and (iii) a distinction between internal and external causes. If a recognised transition is classified as internally caused, then the system must represent some internal mechanism as the source of that transition. This internal source plays the role of the self in the system's causal model. Therefore, internal causal attribution entails that the system represents itself as a cause of some of its own state transitions. □

Thus, self-awareness, understood as internal causal attribution of recognised state transitions, necessarily involves a minimal causal self-model. Importantly, this does not require a linguistic self-representation, a body representation, or any phenomenological notion of self. It requires only that the system can represent some of its own state transitions as being caused by its own internal mechanisms. Within the AToS framework, this minimal causal self-model is realised structurally by the existence of recognised transitions together with internally attributed triggers. In this sense, AToS can be interpreted as a formal framework describing the minimal structural conditions under which a system can represent itself as a cause of its own state transitions. The notion of self that emerges in this framework is therefore not object-based but role-based: the self is the system considered as the internal cause of some of its own recognised state transitions.

2.2 Binary and graded self-awareness.

The minimal criterion of self-awareness introduced in Definition 1.2 is binary: a system either satisfies the internal attribution condition for at least one recognised transition or it does not. However, the level of self-awareness may be graded, depending on the number, diversity, and compositional depth of internally attributed transitions available to the system. This graded notion is captured in later sections through the notions of self-awareness level and intelligence level.This observation motivates the introduction of a formal mechanism for change recognition, which in the AToS framework is represented by Δ. In the formal framework introduced in the following sections, change recognition and causal attribution are represented by the *Sensitivity Function* Δ and the stimulus structure of the system. Thus, in the present framework, the self is not defined by the material composition of the system, nor by the representations it contains, but by what the system can correctly identify as the internal cause of its own state transitions.

3 Definitions

Definition 3.1 (*System-State Values*)**.** *Let V be a set with $|V| > 1$. Elements of V are* System-State Values*.*

Definition 3.2 (Potentially Aware Set)**.** *Let X be a set with $|X| > 1$ and let V be a set of* System-State Values*. Let $\mathcal{U} \subseteq 2^X$ be a non-empty family such that $\bigcup \mathcal{U} = X$. Let*

$$\mathcal{P} := \{\, p : X \to V \,\}.$$

For every $A \in \mathcal{U}$ and $p \in \mathcal{P}$, the Potentially Aware Set (PAS) *associated with (A, p) is defined as*

$$U(A, p) := \{\langle x, p(x)\rangle \mid x \in A\} \subseteq X \times V.$$

Definition 3.3 (Potentially Aware Space)**.** *The* Potentially Aware Space *is*

$$\Omega := \{\, U(A, p) \mid A \in \mathcal{U},\ p \in \mathcal{P} \,\}.$$

Thus, Potentially Aware Space *is the space of all* PAS*; each* PAS *is an element of Ω.*

Definition 3.4 (*Sensitivity Function*)**.** $\Delta\colon \Omega \times \Omega \to \{0,1\}$ *is a* Sensitivity Function*.*

Definition 3.5 (State Transition Function)**.** *A* State Transition Function *is a map*

$$\varphi : \Omega \longrightarrow \Omega$$

such that for each $U(A,p) \in \Omega$*, with* $p : X \to V$*, there exists* $p' : X \to V$ *such that*

$$\varphi\big(U(A,p)\big) = U(A,p').$$

The State Transition Function Set *is*

$$\Sigma := \{\varphi \in \Omega^{\Omega} \mid \varphi \textit{ is a State Transition Function}\}.$$

Definition 3.6 (Internal Stimulus Function)**.** *A map*

$$\psi : \Omega \times \Omega \;\longrightarrow\; \Omega$$

is an Internal Stimulus Functions *iff there exist* State Transition Function*s* $\varphi^B, \varphi^A \in \Sigma$ *such that for every state assignment* $p : X \to V$ *with* $U(B,p), U(A,p) \in \Omega$*,*

$$\Delta\big(U(B,p),\, \varphi^B(U(B,p))\big) = 1 \;\;\Longrightarrow\;\; \psi\big(\,\varphi^B(U(B,p)),\, U(A,p)\big) \;=\; \varphi^A(U(A,p)).$$

We denote by

$$\Psi \;:=\; \big\{\, \psi : \Omega \times \Omega \to \Omega \;\big|\; \psi \textit{ is an } \text{Internal Stimulus Functions} \,\big\}$$

the Internal Stimulus Functions Set*.*

Definition 3.7 (Stimulus Set)**.** *Let* $U(A,p) \in \Omega$*. We say that* $U(B,p) \in \Omega$ *is a* Stimulus Set *for* $U(A,p)$*, written*

$$U(B,p) \in \mathcal{S}(U(A,p)),$$

iff there exist $\varphi^B, \varphi^A \in \Sigma$ *and* $\psi \in \Psi$ *such that*

$$\Delta\big(U(B,p), \varphi^B(U(B,p))\big) = 1 \quad \textit{and} \quad \psi(\varphi^B(U(B,p)), U(A,p)) = \varphi^A(U(A,p)).$$

Notation 3.1 (Index suppression and stimulus-side abstraction)**.** *Let* $U = U(A,p) \in \Omega$*. Any* $S = U(B,p) \in \mathcal{S}(U(A,p))$ *is called simply a* Stimulus Set *or* StSe*.*

Target-side index suppression. *Whenever the reference target* $U(A,p)$ *is fixed by context, and its successor is determined by a witnessing function* $\varphi^A \in \Sigma$*, and it is clear that* φ^A *acts on the* PAS *represented by* U*, we write*

$$\varphi(U) \;:=\; \varphi(U(A,p)) \;:=\; \varphi^A(U(A,p)).$$

Stimulus-side abstraction. *If* $S = U(B,p)$ *is fixed with a witnessing triple* $(\psi, \varphi^B, \varphi^A)$ *and* $U = U(A,p)$ *is the corresponding reference target as in Definition 3.7, then the explicit form of the stimulus-side successor* $\varphi^B(S) = \varphi^B(U(B,p))$ *plays no role in subsequent reasoning. Therefore we write simply*

$$S \in \mathcal{S}(U)$$

and abbreviate the action of ψ *by*

$$\psi(S,U) \;:=\; \psi(\varphi^B(U(B,p)),\, U(A,p)).$$

Thus all expressions involving the witnessing triple can be written in the concise operational form

$$\psi(S,U) \;=\; \varphi(U).$$

Definition 3.8 (Stimulus System)**.** *The* Stimulus System *is the collection of all* PAS *that can serve as* Stimulus Set*s for some target state in* Ω*:*

$$\mathfrak{S} \;:=\; \{\, U(B,p) \in \Omega \mid \exists\, U(A,p) \in \Omega \text{ such that } U(B,p) \in \mathcal{S}(U(A,p)) \,\}.$$

Equivalently, $\mathfrak{S}$ *is the union of all stimulus domains* $\mathcal{S}(U)$ *over* $U \in \Omega$*.*

Definition 3.9 (Recognised transitions and trigger label)**.** *Let*

$$\mathcal{T} \;:=\; \{(U,\varphi) \in \Omega \times \Sigma \mid \Delta(U,\varphi(U)) = 1\}$$

be the set of recognised transitions*. A* trigger label *is a total map*

$$\lambda : \mathcal{T} \to \{\text{int}, \text{ext}\}$$

such that for every $(U,\varphi) \in \mathcal{T}$*,*

$$\lambda(U,\varphi) = \text{int} \iff \exists\, S \in \mathcal{S}(U)\; \exists\, \psi \in \Psi\; \exists\, \varphi^B \in \Sigma \text{ such that}$$

$$\Delta(S, \varphi^B(S)) = 1 \;\wedge\; \psi(\varphi^B(S),\, U) = \varphi(U).$$

Equivalently,

$$\lambda(U,\varphi) = \text{ext} \iff \neg\Big(\exists\, S \in \mathcal{S}(U)\; \exists\, \psi \in \Psi\; \exists\, \varphi^B \in \Sigma \text{ such that}$$

$$\Delta(S, \varphi^B(S)) = 1 \;\wedge\; \psi(\varphi^B(S),\, U) = \varphi(U)\Big).$$

We further set

$$\mathcal{T}_{\text{int}} \;:=\; \{(U,\varphi) \in \mathcal{T} \mid \lambda(U,\varphi) = \text{int}\}, \qquad \mathcal{T}_{\text{ext}} \;:=\; \{(U,\varphi) \in \mathcal{T} \mid \lambda(U,\varphi) = \text{ext}\}.$$

Definition 3.10 (Imagination and Inspiration Functions)**.** *Let* λ *be as in Definition 3.9. Define*

$$\Phi_{\text{int}} := \{\varphi \in \Sigma \mid \exists\, U \in \Omega :\; (U,\varphi) \in \mathcal{T}_{\text{int}}\},$$

$$\Phi_{\text{ext}} := \{\varphi \in \Sigma \mid \exists\, U \in \Omega :\; (U,\varphi) \in \mathcal{T}_{\text{ext}}\},$$

$$\Phi := \Phi_{\text{int}} \cup \Phi_{\text{ext}}.$$

Elements of Φ_{int} *are called* Imagination Function*. Elements of* Φ_{ext} *are called* Inspiration Function*.*

Definition 3.11 (Intuition)**.** *The* Intuition *is the class of all* State Transition Function*s that may operate both as Imagination and as Inspiration. Formally,*

$$\Phi_{\text{intu}} \;=\; \Phi_{\text{int}} \cap \Phi_{\text{ext}}.$$

Definition 3.12 (Conscious System)**.** *A tuple*

$$\langle \Omega, \Sigma, \mathfrak{S}, \Delta \rangle$$

is called a Conscious System *if*

- Ω *is the* Potentially Aware Space *(Definition 3.3),*
- Σ *is the* State Transition Function Set *(Definition 3.5),*
- Δ *is the* Sensitivity Function *(Definition 3.4),*
- $\mathfrak{S}$ *is the* Stimulus System *induced by* (Ω, Σ, Δ) *(Definition 3.8),*

and there exist $U \in \Omega$ *and* $\varphi \in \Phi$ *such that*

$$\Delta\big(U, \varphi(U)\big) = 1.$$

Definition 3.13 (*Self-Conscious System*)**.** *A tuple*

$$\langle \Omega, \Sigma, \mathfrak{S}, \Delta \rangle$$

is called a Self-Conscious System if

- $\Omega, \Sigma, \mathfrak{S}, \Delta$ *are as in Definition 3.12, and*
- *there exist* $U \in \Omega$ *and a* Stimulus Set $S \in \mathcal{S}(U)$ *such that*

$$\psi(S, U) = \varphi_{\text{int}}(U), \qquad \Delta\big(U, \varphi_{\text{int}}(U)\big) = 1,$$

where φ_{int} *is the* Imagination Function *associated (in the sense of Definition 3.7) with the witnessing triple of* $S \in \mathcal{S}(U)$*.*

4 Representation Theorem for Self-Awareness

Theorem 4.1 (Representational Adequacy of AToS)**.** *A tuple*

$$\langle \Omega, \Sigma, \mathfrak{S}, \Delta \rangle$$

is a Self-Conscious System *(in the sense of Definition 3.13)* if and only if *it satisfies the general understanding of self-awareness (Definition 1.2).*

Remark (Representational adequacy). Theorem 4.1 should be read as an adequacy claim at the level of representation: a system is self-aware in the sense of Definition 1.2 if and only if it admits a tuple-based AToS representation $\langle \Omega, \Sigma, \mathfrak{S}, \Delta \rangle$ satisfying Definition 3.13.

Proof. Sufficiency. Assume that the system instantiates the tuple $\langle \Omega, \Sigma, \mathfrak{S}, \Delta \rangle$ and fulfils Definition 3.13. Thus, there exist a *PAS*

$$S = U(B, p) \in \mathcal{S}\big(U(A, p)\big)$$

and maps $\psi \in \Psi$ and $\varphi_{\text{int}} \in \Phi_{\text{int}}$ such that

$$\psi\big(S,\, U(A, p)\big) = \varphi_{\text{int}}\big(U(A, p)\big) =: U(A, p'),$$

together with

$$\Delta\big(U(A, p),\, U(A, p')\big) = 1.$$

Thus the system not only detects a state-change but also attributes its cause to the internal mechanism φ_{int} triggered by the internal *Stimulus Set* S. Hence the tuple satisfies Definition 1.2, proving sufficiency.

Necessity. Assume that the system satisfies Definition 1.2: it recognises at least one change of its internal state and, for that change, determines that the change was internally triggered in the system's operational semantics (i.e. not triggered by any external stimulus at the time of the transition), but rather by the system's own internal mechanism.

Step 1: Representing states. A detected change must admit a mathematical representation. Choose a set X of state indices and a set V of *System-State Values*. Let $\mathcal{P} := \{\, p : X \to V \,\}$. A detected internal change is then realised by two assignments

$$p_1, p_2 : X \to V, p_1 \neq p_2$$

that differ on some subset $A \subset X$. These determine the *PAS*

$$U := U(A, p_1), \qquad U' := U(A, p_2).$$

Step 2: Constructing the internal transition. Define $\varphi^A \in \Sigma$ by

$$\varphi^A(W) := \begin{cases} U', & \text{if } W = U, \\ W, & \text{otherwise,} \end{cases} \qquad W \in \Omega.$$

Step 3: Existence of an internal trigger witness. Unpacking the phrase *internally triggered in the operational semantics*: the invocation of φ^A on U is mediated by some internal trigger witness whose configuration changes at the time of invocation. Represent this trigger witness by a domain $B \subseteq X$. Set

$$S \;:=\; U(B, p_1).$$

Then there exists an assignment $q \in \mathcal{P}$ such that

$$S' \;:=\; U(B, q) \qquad \text{and} \qquad S' \neq S,$$

where the inequality expresses that a genuine internal triggering event occurred.

Step 4: Fixing an admissible PAS universe. Let $\mathcal{U} \subseteq 2^X$ be any non-empty family such that $\bigcup \mathcal{U} = X$ and $A, B \in \mathcal{U}$. Define

$$\Omega := \{\, U(C, p) \mid C \in \mathcal{U},\ p \in \mathcal{P} \,\}.$$

Then $U, U', S, S' \in \Omega$.

Step 5: Defining the sensitivity function. Define a *Sensitivity Function* $\Delta : \Omega \times \Omega \to \{0, 1\}$ by the rule

$$\Delta\big(U(C, p),\, U(C, p')\big) \;:=\; \begin{cases} 1, & \text{if } C \in \{A, B\} \text{ and } U(C, p) \neq U(C, p'), \\ 0, & \text{otherwise,} \end{cases}$$

for all $p, p' \in \mathcal{P}$ and all $C \in \mathcal{U}$. In particular, $\Delta(U, U') = 1$ and $\Delta(S, S') = 1$.

Step 6: Constructing the trigger-side transition. Define $\varphi^B \in \Sigma$ by

$$\varphi^B(W) := \begin{cases} S', & \text{if } W = S, \\ W, & \text{otherwise,} \end{cases} \qquad W \in \Omega.$$

Step 7: Constructing an internal stimulus function. Define $\psi : \Omega \times \Omega \to \Omega$ by stipulating that on the relevant pair $(\varphi^B(S), U)$ it returns the target-side successor:

$$\psi(\varphi^B(S),\, U) \;:=\; \varphi^A(U) \;=\; U'.$$

(Extend ψ arbitrarily to all of $\Omega \times \Omega$.)

Step 8: Verifying the witnessing structure. By construction,

$$\Delta(S, \varphi^B(S)) = 1 \quad \Longrightarrow \quad \psi(\varphi^B(S), U) = \varphi^A(U),$$

so ψ satisfies Definition 3.6, hence $\psi \in \Psi$. Moreover, since $\Delta(S, \varphi^B(S)) = 1$ and $\psi(\varphi^B(S), U) = \varphi^A(U)$, Definition 3.7 yields

$$S \in \mathcal{S}(U).$$

Step 9: Completing the construction. We have now constructed:

- a *PAS* U and its successor U',

- an *Imagination Function* φ^A with $\varphi^A(U) = U'$,
- a *Stimulus Set* $S \in \mathcal{S}(U)$,
- detectable changes $\Delta(U, U') = 1$ and $\Delta(S, S') = 1$,
- and a *Internal Stimulus Functions* $\psi \in \Psi$ mediating the causal link.

Therefore the tuple

$$\langle \Omega, \Sigma, \mathfrak{S}, \Delta \rangle$$

satisfies Definition 3.13, establishing necessity. □

5 Remarks

Remark 5.1 (Interpretation and scope of the framework)**.** *The results of this theory should be understood as structural representation results rather than as claims about a specific empirical implementation of self-awareness. The aim of the AToS framework is not to propose a competing empirical theory nor to engage in a direct comparison with existing accounts in philosophy of mind, cognitive science, or artificial intelligence.*

Instead, the framework identifies the minimal structural conditions *that any theory of self-awareness must satisfy once the Internal Recognition Assumption (Assumption 1.1) is adopted. In this sense, AToS provides a* substrate-independent axiomatic representation *of a broad class of theories in which a system recognises and internally attributes its own state changes through the operation of its internal mechanisms.*

Within this perspective, Theorem 4.1 establishes the adequacy of the representation: every system satisfying the minimal operational notion of self-awareness (Definition 1.2) admits an AToS representation, and conversely every AToS structure realises such a system. Proposition 2.1 expresses only a conceptual consequence of the Internal Recognition Assumption and serves as a bridge between this assumption and the formal definition that follows.

5.1 *Sensitivity Function*

Remark 5.2 (Interpretation of the detection function)**.** *The* Sensitivity Function Δ *is an intrinsic component of the system rather than an external measurement device imposed by the theorist. Different systems may therefore instantiate different detection functions reflecting their internal mechanisms for recognising state changes.*

The value $\Delta(U, U') = 1$ indicates that the system recognises the transition $U \to U'$ as a relevant state change within its operational dynamics. Thus Δ does not merely detect that two states differ; it encodes the system's internal criterion for acknowledging that a transition has occurred.

It is important to stress that the detection function Δ does not merely indicate that two states are objectively different. Rather, Δ represents the system's own internal criterion for recognising that a transition has occurred. Thus, a transition counts as recognised not simply because two states differ, but because the system itself registers the transition as a relevant internal change within its operational dynamics.

Remark 5.3 (Empirical and engineering interpretation)**.** *For naturally occurring systems (biological, artificial, or organisational), the detection function Δ should be regarded as determined by the system itself. In such cases, the main task is to identify the mechanism through which the system recognises its own state transitions.*

Conversely, when designing artificial systems within the AToS framework, the specification of Δ becomes an engineering choice determining which transitions are recognised and therefore shaping the system's operational dynamics and level of self-awareness.

Remark 5.4 (Degenerate detection functions)**.** *Two degenerate forms of the detection function are possible in principle:*

$$\Delta(U, U') = 0 \quad \textit{for all } U, U' \in \Omega, \qquad \Delta(U, U') = 1 \quad \textit{for all } U, U' \in \Omega.$$

In the first case the system never recognises any state change and therefore cannot satisfy the minimal self-awareness criterion of Definition 1.2. In the second case every transition is recognised indiscriminately, eliminating the system's ability to distinguish meaningful internal changes from arbitrary state updates. Consequently, neither case is relevant for non-trivial forms of self-awareness.

5.2 Potentially Aware Set

Remark 5.5 (Potential vs. operational PAS)**.** *We interpret Ω as the space of structurally admissible configurations; the system's operational awareness consists only of those states that become realised through Δ-recognized transitions.*

Remark 5.6 (Why State Transitions Preserve the Domain)**.** *In the definition of a* State Transition Function *we require*

$$\varphi(U(A, p)) = U(A, p'),$$

that is, the domain A is preserved under state evolution.

This requirement is essential for self-awareness. A change of the domain $A \mapsto B$ would prevent the system from recognising a state difference: the PAS *$U(A, p)$ and $U(B, p')$ would no longer be pointwise comparable, making the sensitivity function $\Delta(U(A, p), U(B, p'))$ undefined as a recognition criterion of internal change. Self-awareness requires that the system can determine whether a transition*

$$U(A, p) \longrightarrow U(A, p')$$

originates internally or externally. Such determination is possible only when the structural identity of the state namely its domain A remains fixed. Hence, State Transition Functions *operate solely by updating values $p(x)$, not by altering the underlying domain of awareness.*

Remark 5.7. *The witnessing components (e.g. λ, ψ) are not introduced as free external annotations. They are required to be* realised by the system *as internal operations applicable to its own logged transitions/states. Thus, the existence of witnessing is a* nontrivial internal capability *demanded by the axioms, not a post-hoc relabelling by the theorist.*

Assumption 5.8 (Shared Global Assignment)**.** *Unless stated otherwise, we analyse the system at a fixed chronological stage of its evolution, characterised by a single global value assignment*

$$p : X \to V.$$

Every PAS *available at that stage is therefore a restriction of this assignment to some domain:*

$$U(A, p), \qquad A \in \mathcal{U}.$$

Thus whenever a Stimulus Set *$U(B, p)$ is capable of triggering a transition in a target state $U(A, p)$, both states share the same underlying global assignment p; only their domains differ. The system evolves by transforming p into a successor assignment p', typically through local updates on the domain of the affected* PAS*.*

Remark 5.9 (On the Role of Stimuli)**.** *There are two distinct ways in which a* State Transition Function *φ may be triggered:*

- *by an* external stimulus*, arising directly from outside the system, which invokes some* Inspiration Function *φ_{ext} and updates a target state $U(A, p)$;*

- *by an* internal stimulus*, arising from within the system, which serves as the trigger that invokes* Imagination Function φ *and modifies* $U(A, p)$.

An external stimulus is not represented as a PAS*: it acts directly from the environment by causing a transition in some* $U(A, p) \in \Omega$. *By contrast, an* Stimulus Set *must be represented as an element of* Ω. *The system can recognize the effect of such a stimulus only if the resulting change can be evaluated by the* Sensitivity Function Δ. *If the* StSe *were not part of* Ω*, no recognition of its impact would be possible.*

Hence any genuine internal stimulus must already belong to Ω*; and even if it does not, one may, without loss of generality, enlarge* Ω *to include a corresponding* PAS $U(B, q)$ *so that the system can consistently account for its action. The system is therefore closed under internal stimuli: every internal trigger is (or can be made) a* PAS *in* Ω.

Remark 5.10 (Locality of State Transitions and the Shared Assignment in the Stimulus Set)**.** *Although the space* Ω *contains all formally admissible pairs* $U(A, q)$ *with arbitrary assignments* $q : X \to V$*, every chronologically fixed configuration of the system is characterised by a single assignment* $p : X \to V$ *representing the global state of the system at that stage of its evolution. All* PAS *that are operationally available at this stage are therefore restrictions of the same assignment, and take the form* $U(A, p)$*,* $U(B, p)$ *or* $U(S, p)$ *for suitable domains* $A, B, S \in \mathcal{U}$.

A State Transition Function φ *that updates* $U(A, p)$ *to* $U(A, p')$ *may be regarded, without loss of generality, as modifying the values of* p *only on* A. *Formally, one may assume*

$$\varphi(U(A, p)) = U(A, p') \qquad \text{and} \qquad p'(x) = p(x) \text{ for all } x \notin A.$$

Alternatively, one may allow p' *to take arbitrary structurally admissible values outside* A*, since these values play no operational role in the analysis of the resulting state* $U(A, p')$*. In either interpretation, what matters is the locality of the update: only the values on* A *are relevant to the transition semantics, and the behaviour of* φ *outside* A *does not affect the system's reasoning or its recognition of change.*

This interpretation strengthens the coherence of the Stimulus Set *definition. Since all active* PAS *are restrictions of the same global assignment* p*, and* φ *acts only through local modifications of that assignment on* A*, the relationships between stimuli and transitions described by* ψ *and* Δ *retain full structural comparability. As a consequence, both change recognition and causal attribution remain well-defined across successive chronological configurations of the system.*

For these reasons, instead of allowing Stimulus Set*s to range over* PAS *of the form* $U(B, q)$ *with potentially different assignments, it is conceptually and operationally clearer to take both the stimulus and the target state to share the same global assignment* p*. This choice accurately reflects the system's architecture: all active* PAS *arise simultaneously as restrictions of a single global valuation, and stimuli operate within this shared configuration.*

Remark 5.11 (On the role of pre-change states in the Stimulus Set)**.** *The* Stimulus Set $\mathcal{S}(U(A, p))$ *contains pre-change states* $U(B, p)$*, not their post-change successors* $\varphi^B(U(B, p))$*. Conceptually, a stimulus serves to identify* which PAS-state *acts as the source of stimulation, whereas the causal mechanism encoded by an* Internal Stimulus Functions *operates on the* recognised change *of that state. Accordingly, the* Internal Stimulus Functions ψ *takes the post-change state* $\varphi^B(U(B, p))$ *as its operational input, while the identity of the stimulus remains the pre-change PAS* $U(B, p)$*. This separation between the* source of stimulation *and the* causal input *is essential for distinguishing structural origin from functional effect.*

Remark 5.12 (Interpretation and Necessity of Stimulus Set)**.** *Intuitively, if* $U(B, p) \in \mathcal{S}(U(A, p))$*, then a change in the state* $U(B, p)$ *acts as a stimulus set that triggers the* Imagination Function φ^A *acting on* $U(A, p)$*.* Δ *plays the role of a causal recognizer: it evaluates whether the change in* $U(B)$ *is registered by the system as significant; only such recognized changes may trigger*

transformations of $U(A)$. Thus the Internal Stimulus Functions *$\psi \in \Psi$ encodes causal links between* PAS *states, while $\varphi \in \Phi$ encodes local state transitions. Without the* Stimulus Set*, the system would lack a mechanism for selective activation: either all* State Transition Functions *would fire simultaneously (resulting in chaotic dynamics), or none would ever be invoked (resulting in a static system). The* Stimulus Set *therefore provides the structural basis for the system's controlled dynamism.*

Remark 5.13 (Cascades of Stimuli)**.** *Since each* Stimulus Set *$\mathcal{S}(U(A,p))$ may contain* PAS *states $U(B,p)$ that themselves belong to other stimulus sets, a single recognized update may propagate through a chain of ψ-mediated relations:*

$$U(B,p) \rightsquigarrow U(A,p') \rightsquigarrow U(C,p'') \ \ldots$$

This models cascades of internal updates, where a local modification of the global assignment on one domain can trigger subsequent transitions on other domains. The control discipline remains simple: every invocation of a State Transition Function *must be backed by at least one trigger $U(B,p)$ belonging to the relevant* Stimulus Set *$\mathcal{S}(U(A,p))$.*

Remark 5.14. *Imagination and inspiration functions are not new kinds of operations; they are specialised* instances *of* State Transition Functions*. Formally,*

$$\Phi_{\text{int}} \subseteq \Sigma, \qquad \Phi_{\text{ext}} \subseteq \Sigma, \qquad \Phi = \Phi_{\text{int}} \cup \Phi_{\text{ext}} \subseteq \Sigma.$$

Thus, every imagination or inspiration function preserves the structural constraints of Σ; what differs is only the triggering condition *that determines its membership in Φ_{int} or Φ_{ext}.*

Remark 5.15 (Classification of State Transition Functions)**.** *The distinction between* Imagination Function *and* Inspiration Function *is contextual rather than structural: a* State Transition Function *φ does not intrinsically belong to either class. Its classification depends solely on the type of stimulus that triggers its activation. If φ is invoked by a* PAS *belonging to $\mathcal{S}(U)$, then $\varphi \in \Phi_{\text{int}}$; if it is invoked by an external (non-PAS) stimulus, then $\varphi \in \Phi_{\text{ext}}$. A single transition mechanism may therefore operate in two modes. We represent this by two instances, φ_{int} and φ_{ext}, that share the same state-update behaviour but differ in their triggering conditions. By Assumption 5.18, the trigger class uniquely determines which instance is active in any given transition.*

Remark 5.16 (Noise, drift, and imagination functions)**.** *Within AToS, the classification of a transition as internal does not depend on the physical origin of the triggering process (e.g. deterministic computation, stochastic noise, hardware fluctuation, or timer signal), but on its role within the system's recognised causal structure.*

A random or noisy process constitutes an imagination function Imagination Function *only if:*

1. *it participates in the system's stimulus structure as a trigger of transitions, and*
2. *changes in this process are recognised by the sensitivity function Δ as causes of subsequent state transitions.*

If a fluctuation produces a state change but the system does not recognise that fluctuation as the trigger of the change, then the fluctuation is merely a physical perturbation and does not constitute an imagination function in the AToS sense.

Thus, in AToS, imagination is not defined by determinism, randomness, or physical implementation, but by recognised internal causation. An internal process does not qualify as an imagination function merely because it is physically located within the system. It qualifies as an imagination function only if the system can recognise a change in that process and treat that change as a causal trigger of a subsequent state transition.

Accordingly, AToS defines internal causation in operational rather than physical terms: a cause is internal if it is realised within the system's own transition structure as the source of a recognised change. In this sense, the distinction between internal and external causation is not a distinction between physical locations, but a distinction within the system's own causal model.

Consequently, internal and external causation are system-relative notions: they are defined with respect to the system's own recognition and attribution structure, not with respect to an external observer's physical description of the system.

Remark 5.17 (Intuition Functions)**.** *A* State Transition Function *belongs to the intuition space* $\Phi_{\text{intu}} = \Phi_{\text{int}} \cap \Phi_{\text{ext}}$ *precisely when it may be triggered in both internal and external modes. This reflects not ambiguity but versatility: the same internal mechanism is capable of updating the system's state either in response to a PAS-based stimulus or in the absence of such a stimulus, replicating the effect of an external cause.* Intuition Function *is therefore the class of transition patterns that are causally dual.*

Assumption 5.18 (Trigger Dichotomy)**.** *For every admissible transition* $\varphi \in \Phi$ *applied to a target* $U \in \Omega$ *with a detected change* $\Delta\big(U, \varphi(U)\big) = 1$*, exactly one holds:*

(a) ***Internal trigger:*** $\exists\, S \in \mathcal{S}(U)$ *such that* $\psi_\varphi(S, U) = \varphi(U)$*;*

(b) ***External trigger:*** *the invocation of* φ *on* U *is marked as external (i.e. not mediated by* $\mathcal{S}(U)$*).*

These two cases are mutually exclusive and collectively exhaustive: every recognised transition must be caused either by an internal PAS*-based trigger or by an external environmental event, and no third type of trigger is permitted within the system's operational semantics.*

Remark 5.19 (Why Consciousness is Weaker than Self-Consciousness)**.** *The definition of a Conscious System requires only the existence of a detected change* $\Delta(U, \varphi(U)) = 1$*. This corresponds to the minimal philosophical notion of consciousness as sensitivity: the system is aware that its internal state has changed, but it need not identify the cause of that change. By contrast, self-consciousness demands strictly more: the ability to determine where the change originated. This requires distinguishing whether the transition* $\varphi(U)$ *was triggered by an internal PAS-based stimulus or by an external event, thus evaluating the causal structure of its own updates. Conceptually, consciousness concerns the egistration of change, whereas self-consciousness concerns the attribution of change. The former is logically and operationally weaker than the latter, and this asymmetry is essential to the hierarchy of awareness levels captured in the AToS framework.*

6 Consequences and Constructions

Corollary 6.1 (Minimal internal attribution)**.** *If a system is self-aware in the sense of Definition 1.2, then there exist* $U \in \Omega$*, a recognised transition* $\varphi \in \Sigma$ *with*

$$\Delta\big(U, \varphi(U)\big) = 1,$$

and a Stimulus Set $S \in \mathcal{S}(U)$ *such that the transition* $U \mapsto \varphi(U)$ *admits an internal causal attribution, i.e.*

$$\lambda(U, \varphi) = \text{int}.$$

Equivalently, there exist Stimulus Set $S \in \mathcal{S}(U)$*,* $\psi \in \Psi$ *and* $\varphi^B \in \Sigma$ *with*

$$\psi(\varphi^B(S),\, U) = \varphi(U).$$

Proof. Immediate from the necessity direction in Theorem 4.1 and Definition 3.9. □

Definition 6.2 (External awareness vs. self-awareness)**.** *We call a system* Externally Aware *if it can recognize at least one externally triggered transition, i.e. there exist* $U \in \Omega$ *and* $\varphi \in \Sigma$ *such that*

$$\Delta(U, \varphi(U)) = 1 \quad \textit{and} \quad \lambda(U, \varphi) = \text{ext}.$$

A system is self-aware *iff it satisfies Definition 1.2.*

Corollary 6.3 (External awareness is not sufficient)**.** *External awareness (Definition 6.2) does not imply self-awareness (Definition 1.2). In particular, recognizing externally triggered changes alone is insufficient without internal causal self-attribution of at least one recognized transition.*

Proof. External awareness requires the existence of a recognised transition labelled ext, whereas self-awareness requires the existence of a recognised transition labelled int (i.e. internally attributable) as in Definition 1.2 and Definition 3.9. Hence the implication fails. □

Proposition 6.4 (Hereditary structure of admissible PAS domains)**.** *Let* $U(A, p) \in \Omega$ *and let* $\varnothing \neq B \subseteq A$*. If* $B \in \mathcal{U}$*, then* $U(B, p) \in \Omega$ *and*

$$U(B, p) = U(A, p) \cap (B \times V).$$

Moreover, for every $\varphi \in \Sigma$ *with* $\varphi(U(A, p)) = U(A, p')$*, the restricted pair* $U(B, p), U(B, p')$ *consists of admissible PAS states. If* $p|_B = p'|_B$*, these states coincide; this does not invalidate their membership in* Ω*.*

Proof. Since $B \in \mathcal{U}$ and $p, p' \colon X \to V$, both $U(B, p)$ and $U(B, p')$ belong to Ω by Definition 3.2. The displayed equality follows by restricting the graph of p to B. Equality of the restricted assignments implies equality of the corresponding PAS states; membership in Ω is independent of whether any transition is recognised. □

Remark 6.5 (Relative boundaries of internal and external causation)**.** *The internal/external distinction is relative to the system whose causal structure is being considered. A trigger represented inside a larger system on domain A may originate outside a subsystem on domain* $B \subseteq A$*. When the subsystem is modelled independently, that trigger may therefore be classified as external (an Inspiration Function), although it was internal (an Imagination Function) at the larger-system level. Such a reclassification concerns the trigger's operational witnessing structure, not merely the set-theoretic location of its source; compare Remark 6.7 and the family-level analysis in Section 11.*

Remark 6.6 (Scope of subsystem self-consciousness)**.** *Proposition 6.4 establishes hereditary* PAS admissibility*, not unconditional hereditary self-consciousness. Definition 3.13 requires a recognised internally attributed transition, which need not survive restriction to an arbitrary* B*. A hereditary self-consciousness theorem therefore applies to those restricted subsystems for which an internal stimulus witness and a* Δ_B*-recognised internal transition are preserved or independently realised. In particular, the absence of a recognised change at one instant does not imply that a PAS ceases to exist.*

Remark 6.7 (Stimuli are role-based, not ontological)**.** *In AToS,* Stimulus Set *is a functional role within the witnessing architecture, not a privileged ontological category. Concretely, a state* $S \in \Omega$ *counts as a* StSe *for a target* U *only relative to the relation* $S \in \mathcal{S}(U)$ *(and thus relative to a recognised, internally labelled transition). This is the precise sense in which AToS is substrate-independent: the theory specifies the roles required for self-awareness (stimulus, target, recognised transition, causal attribution), not their physical realisation.*

Remark 6.8 (Engineering blueprint and minimal internal trigger pattern)**.** ***Engineering view.*** *To construct self-aware AI in the sense of Definition 1.2, an implementation must provide operational counterparts of:*

(i) a state representation (a space Ω of admissible internal states),
(ii) a transition repertoire (a set Σ of state-transition mechanisms),
(iii) a recognised-change criterion (Δ), and
(iv) a stimulus-side witnessing architecture (the induced stimulus system $\mathfrak{S}$ together with Ψ) enabling internal causal attribution.

Formal minimality. *By Theorem 4.1, self-awareness requires at least one recognised transition $(U, \varphi) \in \mathcal{T}$ with $\lambda(U, \varphi) = \text{int}$. Equivalently, there exist $S \in \mathcal{S}(U)$, $\psi \in \Psi$ and $\varphi^B \in \Sigma$ such that*

$$\psi(\varphi^B(S), U) = \varphi(U), \qquad \Delta(U, \varphi(U)) = 1.$$

Further extensions and supplements. *Before moving on, we note that the framework developed so far already suffices to establish the core notion of self-awareness. Nevertheless, to illustrate the scope of possible refinements, we introduce in the following sections a number of additional definitions and constructions. These extensions are not always analysed in full depth here. Rather, they are presented as optional tools for studying properties, implementation strategies, or potential applications of the theory. Some appear without further elaboration in order to leave fertile ground for subsequent work, while also avoiding overwhelming the reader with an excess of technical detail.*

7 Memory

In this work, Memory *is used as a technical term. It denotes the subset of Ω (or PAS) whose invariance is* explicitly recognized *by the system (i.e. the system recognizes that no change has occurred), rather than a general account of storage/retrieval in cognitive architectures. Accordingly, results such as "memory cannot be a stimulus" are consequences of this technical meaning.*

Remark 7.1 (On unrecognised changes)**.** *Consider the query:* "find a *Potentially Aware Set* whose change you did not recognise." *At first sight this appears paradoxical. Since the system recognises a change precisely when*

$$\Delta(U, U') = 1,$$

it may seem impossible for the system to identify any case in which $\Delta(U, U') = 0$. One might therefore suspect that the query is internally inconsistent. However, in the AToS framework the application of a State Transition Function *$\varphi \in \Sigma$ to a* PAS *$U \in \Omega$ need not result in a recognised change. The system may apply a transition $U \mapsto \varphi(U)$ such that*

$$\Delta(U, \varphi(U)) = 0,$$

meaning that no change has been registered. This expresses the absence of recognised *change, not necessarily the absence of any underlying structural modification of the* PAS*. Such a situation is perfectly consistent: the system can be aware of the existence of U while also registering that its value remains constant under all applied transitions. What is impossible, however, is for the system to present a* specific transition *$U \to U'$ that both (i) occurred and (ii) remained unrecognised. Once the system is able to name a transition $U \to U'$, that transition is, by definition, recognised, and therefore satisfies $\Delta(U, U') = 1$. Thus the only admissible responses to the query are states for which no recognised change occurs, not unnoticed transitions. The apparent paradox therefore dissolves once we distinguish between (1) the possibility of a* PAS *undergoing no recognised change and (2) the impossibility of reporting a concrete unrecognised transition.*

Definition 7.2 (Memory)**.** *Let $(\Omega, \Sigma, \mathfrak{S}, \Delta)$ be a* Self-Conscious System*. The* Memory *of the system is the subset*

$$\mathcal{M} := \{U(A, p) \in \Omega \mid \forall \varphi \in \Sigma : \ \Delta(U(A, p), \varphi(U(A, p))) = 0\}.$$

In words: $\mathcal{M}$ consists precisely of those Potentially Aware Set *that are Δ-invariant under all admissible state-transition functions $\varphi \in \Sigma$. No internal or external operation of the system produces a recognized change in such states; they therefore serve as the system's stable memory.*

Remark 7.3 (On the meaning of Memory in AToS)**.** *The term* Memory *is used in this work as a technical notion specific to the AToS framework. It should not be confused with the standard meaning of memory in computer science. In many discussions in philosophy of mind, memory refers to information that the system has stably registered as part of its internal knowledge or experience. Even if such information gradually degrades or becomes less precise over time, the system typically does not recognise this degradation as a state change. From the perspective of the system's awareness, such information therefore behaves as a stable reference.*

In contrast, in computer science and artificial systems, memory usually denotes a storage medium (e.g. RAM, databases, logs) whose contents may be continuously updated and whose changes are typically detected by the system.

Within the AToS framework, the notion of Memory is closer to the first interpretation. Memory consists of those PAS elements whose invariance is recognised by the system, i.e. states for which no transition produces a Δ-recognised change. Conversely, the dynamically updated storage structures typical for artificial systems are conceptually closer to the Potentially Aware Space Ω itself.

Remark 7.4 (Why the quantification over all $\varphi \in \Sigma$ is well-defined)**.** *The definition of Memory requires that for every $U(A,p) \in \Omega$ and every state-transition function $\varphi \in \Sigma$, the expression $\varphi(U(A,p))$ is well-defined. This is not an additional assumption: it follows directly from Definition 3.5. By that definition, every $\varphi \in \Sigma$ is a* total *map*

$$\varphi : \Omega \to \Omega,$$

meaning that φ is defined on every PAS *$U \in \Omega$. Moreover, φ preserves the domain of each* PAS*: for every $U(A,p) \in \Omega$, there exists a uniquely determined $p' : X \to V$ such that*

$$\varphi(U(A,p)) = U(A,p').$$

Thus each State Transition Function *acts by modifying only the value-assignment p while keeping the structural domain A fixed. Consequently, every $\varphi \in \Sigma$ is mathematically applicable to every $U \in \Omega$, regardless of whether it is ever invoked operationally by an internal or external stimulus. The universal quantification*

$$\forall \varphi \in \Sigma$$

in Definition 7.2 is therefore fully justified: Memory contains precisely those PAS *that remain Δ-invariant under all admissible transition mechanisms of the system.*

Remark 7.5 (Existence and Interpretation of Memory)**.** *A* Self-Conscious System*, in the strict axiomatic sense of AToS, is not required to possess a non-empty Memory in order to satisfy Definition 3.13. The ability to attribute at least one internally caused and recognised state-transition already suffices for self-consciousness. Accordingly, both situations,*

$$\mathcal{M} = \emptyset \qquad \textit{and} \qquad \mathcal{M} \neq \emptyset,$$

are fully consistent with the axioms.

Nothing in the framework requires that the system ever revisit or reactivate a memory state. A PAS *may belong to $\mathcal{M}$ even if no internal or external stimulus ever addresses it again. What matters is that the system retains the potential ability via any admissible* State Transition Function *in Σ to verify, should such a stimulus eventually occur, that the state has remained unchanged. Thus Memory comprises precisely those* PAS *that are stably preserved under all transition mechanisms,*

irrespective of whether the system ever engages with them operationally. In this sense, Memory is not a catalogue of actively accessed states but a set of structurally persistent configurations of the system: invariant under all admissible dynamics, available for future verification, and providing the stable background against which recognised changes acquire meaning.

8 Intelligence

Definition 8.1 (Operational Composition of State Functions)**.** *Let* $S \in \mathcal{S}(U(A,p))$ *be a* Stimulus Set *such that its update is detected,*

$$S' := \varphi^S(S), \qquad \Delta(S,S') = 1.$$

Assume that this recognised change triggers $\varphi_1 \in \Phi$ *on the target state, producing*

$$\varphi_1(U(A,p)) \;=\; U(A,p').$$

If the updated state $U(A,p')$ *belongs to the* Stimulus Set *of another* PAS*,*

$$U(A,p') \in \mathcal{S}(U(B,p')),$$

and if the corresponding Internal Stimulus Functions $\psi \in \Psi$ *and* State Transition Function $\varphi_2 \in \Phi$ *satisfy*

$$\psi(U(A,p'),\, U(B,p')) \;=\; \varphi_2(U(B,p')) \;=:\; U(B,p''),$$

then φ_1 *and* φ_2 *are said to be* operationally composable*, written*

$$\varphi_2 \circledcirc \varphi_1.$$

The composite transition maps $U(B,p) \mapsto U(B,p'')$ *and is causally rooted in the initial recognised stimulus* S*.*

Definition 8.2 (Intelligence)**.** *The* Intelligence *of the system is the smallest family of* State Transition Function*'s closed under operational composition. Formally,*

$$\mathcal{I} \;:=\; \langle \Phi \rangle_{\circledcirc},$$

where $\langle \Phi \rangle_{\circledcirc}$ *denotes the closure of* Φ *under all finite operational compositions*

$$\varphi_n \circledcirc \cdots \circledcirc \varphi_1,$$

each arising from a chronologically valid chain of recognised stimuli

$$S_1 \rightsquigarrow S_2 \rightsquigarrow \cdots \rightsquigarrow S_n.$$

Thus Intelligence *is the totality of state-transformations that the system can construct through iterated, stimulus-guided internal reasoning.*

Remark 8.3 (Operational Composition vs. Naive Function Composition)**.** *In contrast to naive algebraic composition of functions, the notion of* Intelligence *includes only those transformation chains that the system can* actually realise *through a sequence*

$$\mathit{StSe} \;\longrightarrow\; \mathit{PAS} \;\longrightarrow\; \mathit{StSe},$$

activated by recognised internal changes. Operational composition requires that each step in a chain be triggered by an internal stimulus whose state-change has been detected by the Sensitivity function Δ*. Only such a recognised stimulus can activate an appropriate Imagination Function,*

leading to a further state-transition. Mere formal composability of functions is therefore insufficient; a causal link grounded in the system's architecture is required.

Consequently, the space of Intelligence *need not form a classically closed algebra under ordinary function composition. Its structure may be partially non-transitive and only fragmentarily connected: only those pairs of transitions are operationally composable that can be embedded into a chronologically valid chain of recognised stimuli. Other pairs, though formally composable as maps, are excluded because no admissible stimulus sequence can activate them in the required order.*

This reflects a fundamental aspect of intelligence: a system may count as intelligent even if it cannot compose arbitrary pairs of State Transition Function*s, provided that it can construct meaningful and extendable reasoning chains arising from its own internal dynamics and causal architecture.*

Lemma 8.4 (Δ-Invariance of Memory)**.** *Let U be a memory state in M. Then for every* State Transition Function *$\varphi \in \Sigma$, the system registers no change:*

$$\Delta(U, \varphi(U)) = 0.$$

Proof. This follows directly from Definition 7.2, which stipulates that memory states are invariant under all internal and external *State Transition Function*s. Hence no application of φ to U can produce a Δ-detectable difference. □

Theorem 8.5 (Memory Cannot Serve as a Stimulus)**.** *No memory state $U \in M$ can serve as a* Stimulus Set *for any* Potentially Aware Set *$U(A, p)$.*

Proof. A *Stimulus Set* S for $U(A, p)$ requires the existence of a *State Transition Function* $\varphi^S \in \Sigma$ such that

$$\Delta(S, \varphi^S(S)) = 1.$$

By Lemma 8.4, every memory state $U \in M$ satisfies

$$\Delta(U, \varphi(U)) = 0$$

for all $\varphi \in \Sigma$. Hence no memory state can fulfil the triggering requirement for *Stimulus Set*s. Therefore, memory states cannot function as stimuli. □

Theorem 8.6 (Memory Excluded from Operational Composition)**.** *No memory state can participate as an intermediate state in any operational composition of* State Transition Function*s forming Intelligence.*

Proof. Operational composition requires a chronologically valid sequence of recognised stimuli:

$$S_1 \rightsquigarrow S_2 \rightsquigarrow \cdots \rightsquigarrow S_n,$$

where each S_i satisfies

$$\Delta(S_i, \varphi_i^S(S_i)) = 1.$$

By Lemma 8.4, memory states never satisfy $\Delta = 1$. Hence they cannot appear as any S_i in such a sequence. Since every composition $\varphi_{i+1} \circledcirc \varphi_i$ requires a stimulus at step i, no memory element can participate in the construction of Intelligence. □

Remark 8.7 (Intelligence Is Dynamical, Not Mnemonic)**.** *Intelligence, defined as the closure of Φ under operational composition, does not require memory states as generative elements. Memory contributes only structurally, by providing stable information within potentially aware spaces that undergo Δ-detectable change, but never as causal drivers of state-transitions.*

Proof. Immediate from Theorem 8.6. □

Theorem 8.8 (Intelligence Does Not Require Knowledge)**.** *Let $\mathcal{I}$ denote the Intelligence of the system, defined as the closure of Φ under operational composition (Definition 8.2). Then $\mathcal{I}$ does not require any memory state as a generative component. Formally, no $U \in M$ can appear in any operational composition that constructs an element of $\mathcal{I}$.*

Proof. By Lemma 8.4, every memory state $U \in M$ satisfies

$$\Delta(U, \varphi(U)) = 0 \qquad \text{for all } \varphi \in \Sigma.$$

Operational composition requires a chronologically valid sequence of recognised stimuli

$$S_1 \rightsquigarrow S_2 \rightsquigarrow \cdots \rightsquigarrow S_n,$$

where each S_i satisfies $\Delta(S_i, \varphi_i(S_i)) = 1$ for some $\varphi_i \in \Sigma$. Since no memory state can satisfy the stimulus condition $\Delta = 1$, no memory state can appear as any S_i in such a sequence. Therefore, memory states cannot contribute to the generative structure of $\mathcal{I}$, and Intelligence is constructed solely from Δ-detectable transitions of non-memory states. □

Definition 8.9 (Intelligence Level)**.** *Let $\mathcal{C}$ be the set of all admissible finite operational compositions,*

$$c = \varphi_n \circledcirc \cdots \circledcirc \varphi_1, \qquad \varphi_i \in \Phi,$$

each step of which is triggered by a recognised stimulus ($\Delta = 1$). For each such composition c, let $|c|$ denote its length. The Intelligence Level *of the system is defined as*

$$L(\mathcal{I}) \;:=\; \sup_{c \in \mathcal{C}} |c|^{[1]}.$$

9 Self-Awareness Level

Definition 9.1 (Awareness Level)**.** *The Awareness Level of the system is*

$$L^{\Omega}_{\Delta} := \big|\{\, (U, U') \mid \Delta(U, U') = 1 \,\}\big|.$$

Definition 9.2 (Self-Awareness Level)**.** *The Self-Awareness Level of the system is*

$$L^{\Omega}_{\Delta,\Phi_{\text{int}}} := \Big|\{\, (U, U') \,|\, \exists\, S \in \mathfrak{S}(U),\ \exists\, \psi \in \Psi,\ \exists\, \varphi \in \Phi_{\text{int}} : \psi(S, U) = \varphi(U) = U',\ \Delta(U, U') = 1 \}\Big|.$$

Remark 9.3. *The quantity $L^{\Omega}_{\Delta,\Phi_{\text{int}}}$ measures how many distinct state transitions the system both generates through its internal mechanisms and successfully recognises via Δ. It therefore quantifies the extent to which the system can detect changes that originate from its own internally triggered dynamics.*

10 Self-Conscious Retention Process

Definition 10.1 (Self-Conscious Retention Process)**.** *Let $U(B, p) \in \mathcal{S}(U(A, p))$ and let $\varphi^B, \varphi^A \in \Phi_{\text{int}}$. Assume there exists an* Internal Stimulus Functions *ψ such that*

$$\Delta\big(U(B, p), \varphi^B(U(B, p))\big) = 1,$$

$$\psi\big(\varphi^B(U(B, p)), U(A, p)\big) = \varphi^A\big(U(A, p)\big) := U(A, p'),$$

[1] The supremum $\sup A$ of a set A is the least upper bound of A; it coincides with $\max A$ when the maximum exists.

and

$$\Delta\big(U(A,p),U(A,p')\big)=1.$$

We say the system implements Self-Conscious Retention *if there exist* $\varphi_1^B\in\Phi_{\text{int}}$, $p'':X\to V$, *and an* Internal Stimulus Functions ψ' *such that*

$$U(A,p')\in\mathcal{S}\big(\varphi^B(U(B,p))\big),$$

$$\psi'\big(\varphi^B(U(B,p)),U(A,p')\big)=\varphi_1^B\big(\varphi^B(U(B,p))\big):=U(B,p''),$$

and moreover

$$\Delta\big(\varphi^B(U(B,p)),U(B,p'')\big)=1.$$

In words: the stimulus $U(B,p)$ *undergoes an internally recognised update* $\varphi^B(U(B,p))$*; this updated state triggers the internal update of* $U(A,p)$ *to* $U(A,p')$*. The new state* $U(A,p')$ *becomes a stimulus for* $\varphi^B(U(B,p))$ *and triggers a further internal update to* $U(B,p'')$*, which the system again recognises as a change.*

Remark 10.2 (Termination of Retention Loops)**.** *The Self-Conscious Retention Process, once initiated, unfolds as an alternating sequence of internally caused and* Δ*-registered updates between two potentially aware states* $U(A,\cdot)$ *and* $U(B,\cdot)$*. The dynamics can be schematically expressed as*

$$\begin{aligned}&U(B,q_0)\ \to\ U(B,q_1)\ \Rightarrow\ U(A,p_0)\ \to\ U(A,p_1)\Rightarrow\\&\Rightarrow\ U(B,q_1)\ \to\ U(B,q_2)\ \Rightarrow\ U(A,p_1)\ \to\ U(A,p_2)\Rightarrow\ \cdots\Rightarrow\\&\Rightarrow\begin{cases}U(A,p_{n-1})\to U(A,p_n),\\U(B,q_{n-1})\to U(B,q_n)\end{cases}\end{aligned}$$

until

$$\Delta(U(A,p_{n-1}),U(A,p_n))=0\quad\text{or}\quad\Delta(U(B,q_{n-1}),U(B,q_n))=0.$$

That is, at each stage the newly produced state on one side becomes a stimulus for the other side, triggering a further internal update as long as the transition is recognised by the sensitivity function Δ*. The retention loop terminates at the first index* n *at which* Δ *fails to register the induced change on either branch of the alternation.*

Remark 10.3 (Temporal Character of Self-Conscious Retention)**.** *Without retention, an act of self-awareness is essentially instantaneous: the system detects a change, attributes it internally, and the event is over. By contrast, if Self-Conscious Retention processes are implemented, the act of self-awareness can unfold over time. In this case, each newly produced state may itself act as a stimulus for further internal updates, generating a temporally extended chain of self-referential recognition. This parallels the human mind, where a single act of self-awareness (e.g. "I am aware that I am thinking") can evolve dynamically into an extended process of reflection. Thus, Self-Conscious Retention transforms self-awareness from a single recognition event into a potentially continuous cognitive process, subject to the detection-based termination condition* $\Delta=0$*; a stable recurrent cycle with* $\Delta=1$ *can instead persist indefinitely (see Section* **??***).*

Remark 10.4 (Retention as the Simplest Form of Intelligence)**.** *The Self-Conscious Retention Process constitutes the minimal non-trivial instance of systemic intelligence in the sense of Definition 8.9. A single* Δ*-registered update realises only self-awareness, but a retention chain realises an* operational composition *of internal imagination functions: each newly produced state becomes an operator that transforms the next state. Thus, retention is the simplest mechanism by which a system expands its reachable space of internal transformations. In this sense, Self-Conscious Retention is the lowest-level manifestation of intelligence: a constructive process that extends reasoning beyond isolated updates and creates a multi-step internal computation.*

11 Operational Self-Awareness Criterion

In this section we introduce an operational, theory-neutral criterion for minimal self-awareness. The proposed criterion plays a role analogous to the Turing Test for intelligence, but targets internal self-causation rather than externally observed behaviour. The Turing Test evaluates intelligence, and the mirror test evaluates bodily self-recognition, whereas the GSAT evaluates internal causal self-recognition.

Definition 11.1 (General Self-Awareness Test)**.** *A system satisfies the* General Self-Awareness Test (GSAT) *if and only if, under conditions of* no external stimuli*, there exists at least one instance in which:*

1. *the system, by virtue of its own internal dynamics, brings about a change in at least one of its internal states, and*
2. *the system is capable of recognising the source of this change, correctly classifying it as* internally caused *rather than resulting from external influence.*

The GSAT is an existential *criterion: the existence of at least one such internally generated and internally recognised state change is sufficient. The test does not require global introspection, a complete self-model, or full transparency of all system states.*

Remark 11.2 (AToS instantiation of GSAT)**.** *By Theorem 4.1, a system is self-aware in the general sense (Definition 1.2) if and only if it admits an AToS representation* $\langle \Omega, \Sigma, \mathfrak{S}, \Delta \rangle$ *satisfying Definition 3.13. Accordingly, the* General Self-Awareness Test (GSAT)*, formulated in a theory-neutral manner, admits a precise and faithful instantiation within AToS.*

In AToS terms, GSAT holds if and only if there exist a target state $U \in \Omega$*, an internal stimulus* $S \in \mathfrak{S}(U)$ *(a* StSe*), an imagination function* Imagination Function $\in \Phi_{\text{int}}$*, and a sensitivity function* Δ *such that, in the absence of external stimuli,*

$$\psi(S, U) = \text{Imagination Function}(U) \quad \textit{and} \quad \Delta\big(U, \text{Imagination Function}(U)\big) = 1,$$

for some internal stimulus function $\psi \in \Psi$*. That is, an internally generated stimulus* StSe *triggers an* Imagination Function*, which induces a change in a* PAS*-state, and this induced change is normatively recognised by* Δ *as internally caused.*

Formally, Δ *does not merely register the occurrence of a state transition, but constitutes its recognition as an internally generated event within the system's decision structure. Thus, GSAT corresponds in AToS to the existence of a non-trivial internally driven causal loop involving* StSe*,* Imagination Function*,* PAS*, and* Δ*.*

In AToS, such triggering presupposes that the change of the stimulus itself is Δ*-recognised; without* Δ*-recognition at the stimulus side, no imagination function can be operationally invoked.*

This establishes GSAT as an operational characterisation of self-awareness whose necessary and sufficient formal realisation is provided by AToS, without presupposing acceptance of the theory itself.

Near-miss counterexample (non-self-aware). *Consider a closed system with sensors and with both externally-triggered and internally-triggered state updates, and assume it logs its own transitions. Suppose, however, that the system lacks any internal mechanism that can* attribute *a logged change to the internal mechanism as opposed to the external one (i.e. no witnessing/attribution map that separates* Φ_{int} *from* Φ_{ext} *in the sense required by Def. 1.1). Then the system may exhibit rich dynamics and even persistent records, yet it fails to be self-aware in the present sense because it cannot recognize a change* as internally caused*. This illustrates that AToS self-awareness is not implied by sensing, dynamics, or logging alone; the attribution capability is essential.*

Remark 11.3 (Operational interpretation of external stimuli in GSAT)**.** *In practical systems it is often argued that a complete absence of external stimuli is difficult to realise. Real implementations typically contain background processes such as clocks, schedulers, interrupts, hardware noise, or slow physical drift of system parameters. Within the AToS framework such phenomena do not require a special ontological treatment. AToS does not classify triggers according to their physical origin (e.g. noise, clock, scheduler), but according to their operational role in the system's transition structure. Formally, the distinction between internal and external transitions is determined solely by the trigger label* $\lambda : \mathcal{T} \to \{\mathrm{int}, \mathrm{ext}\}$ *(Definition 3.9). A recognised transition is internal precisely when its activation admits a stimulus-side witnessing structure within the system.*

Consequently, phenomena such as noise, timers, or scheduling mechanisms may appear either as Imagination Function *or as* Inspiration Function*, depending on how they are integrated into the system's architecture. If such mechanisms are represented as* PAS *elements participating in the stimulus system* $\mathcal{S}(U)$ *and their changes are recognised by* Δ*, they function as ordinary internal stimuli and therefore belong to* Φ_{int}*.*

In this sense, AToS treats these mechanisms as ordinary components of the system's dynamics rather than as exceptional disturbances. A system whose only internally recognised transitions originate from such processes would still satisfy Definition 1.2, although such a system would typically exhibit only a very limited form of self-awareness.

Remark 11.4 (Empirical accessibility of internal attribution)**.** *The GSAT criterion is formulated at the level of the system's internal causal structure. Accordingly, a distinction must be made between the* existence *of internal causal attribution and its* empirical accessibility *to an external observer.*

In general, the fact that a system satisfies GSAT is an ontological property of the system's transition structure, whereas the ability to verify GSAT is an epistemic matter that depends on the observer's access to the system's internal organisation.

For empirical verification, the mere observation of a state change in the absence of external input is not sufficient. A state transition may occur for many reasons, including processes that are not recognised by the system as internal causes. Therefore, empirical evidence for internal attribution requires at least the following conditions:

1. *absence of recognised external stimuli at the time of transition,*
2. *occurrence of a recognised internal state transition,*
3. *existence of a system-level mechanism that classifies the source of this transition as internal,*
4. *structural or interventional confirmation that the triggering process originated within the system's recognised causal structure.*

In artificial systems, these conditions may be directly testable because internal states, trigger pathways, and attribution mechanisms can be made inspectable through logs, architectural constraints, or controlled interventions. For example, one may experimentally disable all recognised external inputs and then selectively disable internal triggering modules; if the attributed internal transitions disappear under such intervention, this provides evidence for genuine internal attribution rather than unrecognised perturbation.

In biological or poorly accessible systems, such verification may be difficult or impossible in practice. This limitation, however, reflects the epistemic limits of observation rather than a limitation of the GSAT criterion itself.

12 AToS and the Chinese Room Argument

12.1 The Chinese Room Argument

The Chinese Room argument, introduced by John Searle, is one of the most influential arguments against the claim that symbol manipulation alone is sufficient for understanding or consciousness. In the classical formulation, a person inside a room follows a set of syntactic rules to manipulate Chinese symbols without understanding their meaning. From the outside, the system appears to understand Chinese, although internally it performs only formal symbol manipulation. The argument is commonly interpreted as showing that passing the Turing Test or producing intelligent behaviour is not sufficient for genuine understanding or consciousness. The Chinese Room therefore distinguishes between behavioural competence and genuine understanding.

12.2 Symbol Manipulation vs. Recognised State Transitions

The AToS framework introduces a different distinction. Instead of focusing on semantics, AToS focuses on the internal structure of state transitions and their causal attribution. In AToS, self-awareness is not defined in terms of language, meaning, or behaviour, but in terms of the system's ability to:

1. *recognize at least one change in its internal state, and*
2. *attribute the cause of that recognized change to its own internal mechanism rather than to an external stimulus.*

A system may therefore manipulate symbols, produce correct answers, and simulate intelligent behaviour, while still failing the AToS self-awareness criterion if it lacks internal recognition of state transitions and internal causal attribution.

12.3 Why the Chinese Room Is Not Self-Aware in the AToS Framework

From the perspective of AToS, the crucial question is not whether the system produces correct outputs, but whether it possesses an internal mechanism that:

- *detects its own internal state transitions, and*
- *represents at least one of these transitions as being caused by its own internal processes.*

A purely syntactic system may implement arbitrarily complex input–output mappings while lacking any internal causal model of its own state transitions. Such a system would fail the AToS self-awareness criterion even if it perfectly simulated intelligent behaviour in the sense of the Turing Test. In this sense, AToS separates three notions that are often conflated:

1. *behavioural intelligence,*
2. *symbol manipulation,*
3. *internal causal self-awareness.*

AToS claims that only the third notion is structurally sufficient for minimal self-awareness.

12.4 Internal Causal Attribution as the Missing Component

The Chinese Room argument shows that symbol manipulation is not sufficient for understanding. AToS strengthens this conclusion by showing that symbol manipulation is also not sufficient for self-awareness. What is missing is not semantics but internal causal attribution. In the AToS framework, the self is not defined as a particular object, representation, or symbol. Instead, the

self appears as a causal role within the system's transition structure. A system represents itself as a cause if it can classify at least some of its own recognised state transitions as being caused by its own internal stimulus structure. Thus, a system that merely manipulates symbols without representing itself as a causal source of its own state transitions does not satisfy the minimal structural condition for self-awareness defined in this work.

12.5 Relation to the Turing Test

The Turing Test evaluates behavioural indistinguishability, whereas the AToS criterion evaluates internal causal structure. These criteria are logically independent. Passing the Turing Test is neither necessary nor sufficient for self-awareness in the AToS sense: a system may be self-aware without human-like behaviour, and a system may exhibit human-like behaviour without being self-aware. The AToS criterion therefore addresses a different question: not whether a system behaves intelligently, but whether it can recognise and internally attribute its own state transitions. More broadly, the Turing Test evaluates behaviour, and the Chinese Room argument targets semantic understanding, whereas the GSAT evaluates internal causal self-recognition. These three criteria therefore address three different questions:

1. *what a system does,*
2. *what a system understands,*
3. *whether a system can represent itself as a cause of its own state changes.*

In this sense, behavioural intelligence, semantic understanding, and self-awareness correspond to three distinct explanatory levels. The AToS framework is concerned only with the third level and proposes internal causal attribution of recognised state transitions as a minimal structural condition for self-awareness.
Intelligence concerns what a system does, understanding concerns what a system means, and self-awareness concerns whether a system can recognise itself as a causal source of its own state transitions.

13 Family of Self-Conscious Systems

Definition 13.1 (Family of Self-Conscious Systems)**.** *Let* $\mathbb{S} = \{S_i\}_{i\in I}$ *be a family of* Self-Conscious Systems*, each of the form*

$$S_i = \langle \Omega_i,\ \Sigma_i,\ \mathfrak{S}_i,\ \Delta_i \rangle,$$

defined over a common carrier X*, and a common value set* V*. We define the* Family *(meta-system)*

$$S^* = \langle \Omega^*,\ \Sigma^*,\ \mathfrak{S}^*,\ \Delta^* \rangle$$

as follows.
(1) Meta-level carrier and value set.

$$X^* := I \times X, \qquad V^* := V.$$

(2) Meta-level target family.

$$\mathcal{U}^* := \{\, \{i\} \times X \mid i \in I \,\}.$$

(3) Meta-level assignment set.

$$\mathcal{P}^* := \{p^* : X^* \to V^*\} = \{p^* : I \times X \to V\}.$$

***(4) Meta-level* PAS.** *For $A^i = \{i\} \times X \in \mathcal{U}^*$ and $p^* \in \mathcal{P}^*$ define*

$$U^*(A^i, p^*) := \{\, \langle (i,x),\ p^*(i,x)\rangle \mid x \in X \,\} \subseteq X^* \times V^*.$$

***(5) Meta-level* Potentially Aware Space.** *Define*

$$\Omega^* \ := \ \{\, U^*(A^i, p^*) \mid i \in I,\ p^* \in \mathcal{P}^* \,\}.$$

(6) Meta-level sensitivity function. *Let $i \in I$ and let $U^*(A^i, p^*) \in \Omega^*$. Define*

$$\Pi_i(U^*(A^i, p^*)) := U_i(A,\, p_i),$$

where $A \in \mathcal{U}_i$ is the target used in the definition of Ω_i, and

$$p_i : X \to V, \qquad p_i(x) := p^*(i,x).$$

Define the sensitivity function

$$\Delta^* : \Omega^* \times \Omega^* \to \{0,1\}$$

as follows. For any $i \in I$ and any $p^*, p'^* \in \mathcal{P}^*$, set

$$\Delta^*(U^*(A^i, p^*),\, U^*(A^i, p'^*)) = 1$$

if and only if there exist assignments $p_i, p_i' \in \mathcal{P}_i$ and a target $A \in \mathcal{U}_i$ such that

$$U_i(A, p_i),\ \ U_i(A, p_i') \in \Omega_i,\ \ \Delta_i(U_i(A, p_i),\, U_i(A, p_i')) = 1,$$

$$\Pi_i(U^*(A^i, p^*)) = U_i(A,\, p_i) \quad \textit{and} \quad \Pi_i(U^*(A^i, p'^*)) = U_i(A,\, p_i').$$

In all other cases, define Δ^* arbitrarily.

(7) Meta-level State Transition Functions. *The notion of a* State Transition Function *extends naturally to a family of self-conscious systems as a* meta-level State Transition Function*, defined as follows.A* meta-level state transition function *is a map*

$$\varphi^* : \Omega^* \to \Omega^*$$

such that for every $U^*(A^i, p^*) \in \Omega^*$ there exist $U^*(A^i, p'^*) \in \Omega^*$, a state transition function $\varphi^i \in \Sigma_i$, and states $U_i(A,p)$, $U_i(A,p') \in \Omega_i$ satisfying

$$\varphi^i(U_i(A,p)) = U_i(A,p'),$$

$$\varphi^*(U^*(A^i, p^*)) \ = \ U^*(A^i, p'^*),$$

and

$$\Pi_i(U^*(A^i, p^*)) = U_i(A,p) \quad \textit{and} \quad \Pi_i(U^*(A^i, p'^*)) = U_i(A,p').$$

The action of φ^* on all other elements of Ω^* is left unspecified.

The meta-level state transition function set *is defined as*

$$\Sigma^* \ := \ \{\, \varphi^* \in (\Omega^*)^{\Omega^*} \mid \varphi^* \textit{ satisfies the conditions specified above} \,\}.$$

(8) Meta-level Internal Stimulus Function *We define an* Internal Stimulus Functions *for a family of self-conscious systems as a* meta-level Internal Stimulus Function *in the following way. A map*

$$\psi^* : \Omega^* \times \Omega^* \ \longrightarrow \ \Omega^*$$

is a meta-level Internal Stimulus Function iff for every $i \in I$ there exist meta-level state transition functions $\varphi^{*B}, \varphi^{*A} \in \Sigma^*$ such that for all meta-states $U^*(A^i, p^*) \in \Omega^*$,

$$\Delta^*\big(U^*(B^i,p^*),\, \varphi^{*B}(U^*(B^i,p^*))\big) = 1 \implies$$

$$\psi^*\big(\varphi^{*B}(U^*(B^i,p^*)),\, U^*(A^i,p^*)\big) \;=\; \varphi^{*A}(U^*(A^i,p^*)),$$

where $A^i, B^i \in \mathcal{U}^*$ are targets associated with the same agent i, and the induced local states satisfy

$$\Pi_i\big(\varphi^{*B}(U^*(B^i,p^*))\big) = \varphi^{B}\big(\Pi_i(U^*(B^i,p^*))\big),$$

$$\Pi_i\big(\varphi^{*A}(U^*(A^i,p^*))\big) = \varphi^{A}\big(\Pi_i(U^*(A^i,p^*))\big)$$

for some $\varphi^B, \varphi^A \in \Sigma_i$. We denote by

$$\Psi^* \;:=\; \big\{\, \psi^* : \Omega^* \times \Omega^* \to \Omega^* \;\big|\; \psi^* \textit{ is a meta-level } \text{Internal Stimulus Functions} \,\big\}$$

the Meta-level Internal Stimulus Functions Set*.*

(9) Meta-level Stimulus Set *Let $U^*(A^i, p^*) \in \Omega^*$. We define a* Stimulus Set *for a family of self-conscious systems as a* meta-level Stimulus Set *$U^*(B^i, p^*) \in \Omega^*$ for $U^*(A^i, p^*)$, written*

$$U^*(B^i,p^*) \in \mathcal{S}^*\big(U^*(A^i,p^*)\big),$$

iff there exist meta-level state transition functions $\varphi^{*B}, \varphi^{*A} \in \Sigma^*$ and a meta-level internal stimulus function $\psi^* \in \Psi^*$ such that

$$\Delta^*\big(U^*(B^i,p^*),\, \varphi^{*B}(U^*(B^i,p^*))\big) = 1$$

and

$$\psi^*\big(\varphi^{*B}(U^*(B^i,p^*)),\, U^*(A^i,p^*)\big) \;=\; \varphi^{*A}(U^*(A^i,p^*)).$$

The meta-level Stimulus System *is the collection of all meta-level* PAS *that can serve as meta-level* Stimulus Set*s for some target meta-state in Ω^*:*

$$\mathfrak{S}^* \;:=\; \big\{\, U^*(B^i,p^*) \in \Omega^* \;\big|\; \exists\, U^*(A^i,p^*) \in \Omega^* \textit{ such that } U^*(B^i,p^*) \in \mathcal{S}^*\big(U^*(A^i,p^*)\big) \,\big\}.$$

Equivalently, $\mathfrak{S}^*$ is the union of all meta-level stimulus domains $\mathcal{S}^*(U^*)$ over $U^* \in \Omega^*$.

Theorem 13.2 (Self-Consciousness of a Family)**.** *Let*

$$S^* = \langle \Omega^*, \Sigma^*, \mathfrak{S}^*, \Delta^* \rangle$$

be a family of self-conscious systems as in Definition 13.1. Then S^* is a Self-Conscious System *in the sense of Definition 3.13 if and only if there exist meta-states $U^*(A^i, p^*), U^*(B^i, p^*) \in \Omega^*$ such that*

$$U^*(B^i,p^*) \in \mathcal{S}^*\big(U^*(A^i,p^*)\big),$$

and there exists a meta-level internal state transition function $\varphi^*_{\text{int}} \in \Sigma^*$ satisfying

$$\psi^*\big(U^*(B^i,p^*),\, U^*(A^i,p^*)\big) = \varphi^*_{\text{int}}\big(U^*(A^i,p^*)\big),$$

together with

$$\Delta^*\big(U^*(A^i,p^*), \varphi^*_{\text{int}}(U^*(A^i,p^*))\big) = 1.$$

Proof. By Definition 3.13, S^* is a *Self-Conscious System* iff there exist $U^* \in \Omega^*$ and $S^* \in \mathcal{S}^*(U^*)$ such that

$$\psi^*(S^*, U^*) = \varphi^*_{\text{int}}(U^*) \quad \text{and} \quad \Delta^*\big(U^*, \varphi^*_{\text{int}}(U^*)\big) = 1,$$

where φ^*_{int} is the internal transition function associated with the witnessing triple for $S^* \in \mathcal{S}^*(U^*)$. Unfolding Definition 13.1(9), we may write $U^* = U^*(A^i, p^*)$ and $S^* = U^*(B^i, p^*)$ for some $i \in I$ and $p^* \in \mathcal{P}^*$, with $U^*(B^i, p^*) \in \mathcal{S}^*(U^*(A^i, p^*))$. This is exactly the stated condition. □

Remark 13.3. *In the above definition, the operator Π_i serves as the sole formal bridge between the meta-level family S^* and the individual systems S_i. Both the meta-level sensitivity function Δ^* and the meta-level state transition functions φ^* are entirely induced by the corresponding local operators Δ_i and φ^i via this projection. Consequently, the family meta-system S^* introduces no new internal dynamics or awareness mechanisms beyond those already present in its constituent self-conscious systems.*

Remark 13.4.

(i) An inter-agent signal $S_j \to S_i$ is internal to the family (since $S_j \in \Omega^*$) while constituting an external stimulus for S_i; accordingly, local Inspiration Function*s lift to meta-level* Imagination Function*s.*

*(ii) The choice of the target state $U \in \Omega_i$ in **(2)** is determined by the concrete interaction pattern and can be formalized via meta-level internal stimulus functions Ψ^*.*

(iii) The set Φ^*_{ext} captures influences exogenous to Ω^*; in the minimal construction one may set $\Phi^*_{\text{ext}} = \varnothing$.

(iv) The edge relation E^* provides a convenient encoding of $\mathfrak{S}^*$ for graph- or network-based analyses.

(v) The sensitivity function Δ^* registers an inter-agent stimulus if and only if it induces a change that is detectable by the target system's own sensitivity Δ_{S_i}.

Remark 13.5 (Local vs. Meta-level Stimuli)**.** *For an individual system S_i, the state of another system S_j $(j \neq i)$ acts as an* external *stimulus. However, at the level of the family meta-system S^*, the same interaction is internal, since $U_j \in \Omega^*$. Thus, inter-agent interactions are external at the local level while constituting internal stimulus sets at the meta-level.*

Proposition 13.6 (Self-Consciousness via Inter-Agent Interaction)**.** *A family of self-conscious systems S^* is a self-conscious system in the sense of Definition 3.13 if and only if there exists an inter-agent interaction that induces a meta-level stimulus set $S^* \in \mathcal{S}^*(U^*)$ witnessing a meta-level internal state transition detectable by Δ^*.*

Remark 13.7 (Limits of Level Aggregation and Prospects for Synergy)**.** *Within the present axiomatic framework, there is no simple, monotonic, or generally valid relationship between the Awareness Level, the Self-Awareness Level, or the Intelligence Level of a family of self-conscious systems and the corresponding levels of its constituent systems. A state change that is detected by an individual system may remain undetectable to other systems or to the family meta-system as a whole, and operational composability of state-transition functions is inherently system-relative and constrained by the local stimulus structure. Consequently, level-based properties cannot be inferred by direct aggregation, maximisation, or summation across agents. At the same time, this limitation delineates a natural direction for further research: the systematic investigation of structural conditions and additional mechanisms under which inter-system interactions could give rise to genuine synergistic effects, potentially enabling amplification or transformation of awareness- or intelligence-related capacities at the family level.*

14 Topological Space

Definition 14.1 (Topological Space of Self-Conscious Systems)**.** *Let $\mathcal{C}$ be the class of all* Self-Conscious System*s defined on a fixed carrier X and value set V. Write $S_i = \langle \Omega_i, \Sigma_i, \mathfrak{S}_i, \Delta_i \rangle$ and let $\Omega^{\text{total}} := \bigcup_{S_i \in \mathcal{C}} \Omega_i$. Declare $\mathcal{B} := \{\Omega_i : S_i \in \mathcal{C}\}$ to be a base and let $\mathcal{T}$ be the topology* generated *by $\mathcal{B}$, i.e. the family of all unions of finite intersections of members of $\mathcal{B}$ that are again unions of elements of $\mathcal{B}$. Then $(\Omega^{\text{total}}, \mathcal{T})$ is called the* topological space of self-conscious systems*.*

Theorem 14.2. *The structure* $(\Omega^{\text{total}}, \mathcal{T})$ *defined in Definition 14.1 is a topological space.*

Proof. We verify the three axioms of a topology.

Empty set and total space. By definition, the empty system is included in $\mathcal{C}$, hence $\varnothing = \Omega_\emptyset \in \mathcal{T}$. Also $\Omega^{\text{total}} = \bigcup_{S_i \in \mathcal{C}} \Omega_i$ belongs to $\mathcal{T}$ by construction.

Arbitrary unions. If $\{U_\alpha\}_{\alpha \in A} \subseteq \mathcal{T}$, then each U_α is a union of some Ω_j. Hence $\bigcup_{\alpha \in A} U_\alpha$ is still a union of Ω_j's, and thus belongs to $\mathcal{T}$.

Finite intersections. Let $U_1 = \bigcup_{j \in J_1} \Omega_j$ and $U_2 = \bigcup_{j \in J_2} \Omega_j$ be two elements of $\mathcal{T}$. Then

$$U_1 \cap U_2 = \bigcup_{j \in J_1} \bigcup_{k \in J_2} (\Omega_j \cap \Omega_k).$$

But $\Omega_j, \Omega_k \subseteq \Omega^{\text{total}}$ are themselves Potentially Aware Spaces. By the intersection property (Definition 3.2), their intersection is again a valid Ω of some system in $\mathcal{C}$. Therefore $U_1 \cap U_2 \in \mathcal{T}$. The same reasoning extends to finite intersections by induction. Thus all three axioms are satisfied, and $(\Omega^{\text{total}}, \mathcal{T})$ is a topological space. □

Theorem 14.3 (Topological universality of self-aware system construction)**.** *Let* (T, τ) *be an arbitrary nonempty topological space. A self-aware system in the sense of AToS can be constructed over* (T, τ) *by introducing a suitable internal state structure and an internal attribution function* Imagination Function*. The construction does not require any particular topology, metric, geometry, or notion of distance.*

Proof. First, suppose that T contains at least two points. Choose a nonempty proper subset $A \subset T$ and let

$$B = T \setminus A.$$

Assign to each of the two subsets a binary state, represented by one of two colours. Define the internal transition function by

$$\varphi_{\text{int}}(a, b) = (1 - a, 1 - b), \qquad a, b \in \{0, 1\}.$$

Thus, every change of the state assigned to A is accompanied by a corresponding change of the state assigned to its complement B. The construction is independent of the topology τ. If T consists of a single point, the distinction between the system and its internal reference is realised through self-reference: the system serves as its own *StSe*. In either case, the construction provides an internal state structure capable of supporting the self-attribution mechanism required by AToS. □

Theorem 14.4 (Existence of a self-aware system with infinite intelligence level)**.** *Within the AToS framework, there exists a self-aware system whose Intelligence Level satisfies*

$$L(\mathcal{I}) = \infty.$$

Such a system can be constructed using a sequence of connected components of the complement of the Cantor set.

Proof. Let $C \subset [0, 1]$ be the standard Cantor set and put

$$I_n = \left(\frac{1}{3^{n+1}}, \frac{2}{3^{n+1}} \right), \qquad n \in \mathbb{N}_0.$$

These intervals are pairwise disjoint connected components of $[0, 1] \setminus C$, and both sequences of endpoints converge to 0. Set

$$X = \bigcup_{n=0}^{\infty} I_n.$$

Represent a configuration by an assignment

$$p : X \longrightarrow \{0, 1\}$$

that is constant on each I_n. The values 0 and 1 represent two colours. Write

$$U_n(p) := U(I_n, p)$$

for the corresponding PAS. For each $n \geq 0$, define a local state-transition mechanism φ_{n+1} that reverses the colour on I_{n+1}. Its update of the global assignment is

$$p'(x) = \begin{cases} 1 - p(x), & x \in I_{n+1}, \\ p(x), & x \notin I_{n+1}. \end{cases}$$

Equip the system with an internal stimulus structure such that a recognised change on I_n triggers φ_{n+1} on I_{n+1}. In particular, at the relevant configuration,

$$U_n(p) \in \mathcal{S}\big(U_{n+1}(p)\big).$$

The triggering mechanism records the source of each update internally. Define the detection and attribution mechanisms on these transitions by

$$\Delta\big(U_{n+1}(p), U_{n+1}(p')\big) = 1, \qquad \lambda\big(U_{n+1}(p), \varphi_{n+1}\big) = \text{int}.$$

Thus each φ_{n+1} operates as an Imagination Function: its invocation is caused by a recognised internal change, and the system identifies that change as its source. Starting from a recognised change on I_0, the system therefore admits the causal cascade

$$I_0 \rightsquigarrow I_1 \rightsquigarrow I_2 \rightsquigarrow \cdots .$$

Every step after the initiating change is internally triggered, recognised, and internally attributed. In particular, the system satisfies the AToS criterion of self-awareness. For every $N \geq 1$, the first N internally triggered updates form an admissible finite operational composition

$$c_N = \varphi_N \circledcirc \cdots \circledcirc \varphi_1, \qquad |c_N| = N.$$

Consequently, by Definition 8.9,

$$L(\mathcal{I}) = \sup_{c \in \mathcal{C}} |c| \geq N \qquad \text{for every } N \geq 1.$$

Hence

$$L(\mathcal{I}) = \infty.$$

□

Remark 14.5. *The conclusion concerns the formal Intelligence Level of AToS. It requires admissible operational compositions of arbitrarily large finite length, not the completion of infinitely many transitions in finite time.*

Definition 14.6 (Hereditary admissible-domain family)**.** *An admissible-domain family $\mathcal{U}$ is* hereditary on nonempty subsets *if*

$$A \in \mathcal{U}, \quad \varnothing \neq B \subseteq A \quad \Longrightarrow \quad B \in \mathcal{U}.$$

This is an additional structural property of a subclass of AToS models; it is not assumed in Definition 3.2.

Proposition 14.7 (Restriction of a PAS)**.** *In a hereditary AToS model, if* $U(A,p) \in \Omega$ *and* $\varnothing \neq B \subseteq A$*, then*

$$U(A,p) \cap (B \times V) = U(B,p) \in \Omega.$$

Proof. The hereditary property makes B admissible. Restriction of the graph of p to B is exactly $U(B,p)$ by Definition 3.2. □

Remark 14.8 (Discrete topology is optional)**.** *The discrete topology* 2^X *may be placed on any index set* X*, but topology and the admissible-domain family are different objects. Our choice* $\mathcal{U} = 2^X \setminus \{\varnothing\}$ *is a hereditary covering family and does not rely on topological openness. In particular, no topological property alone supplies an internal stimulus witness; that witness is constructed above.*

15 Related Work

We briefly compare five influential approaches to consciousness or self-monitoring with the present axiomatic proposal. Our comparison focuses on three structural criteria that are central to the present framework: (i) whether the approach explicitly separates internal from external causes of change, (ii) whether it provides a persistent and formally defined memory or Potentially Aware Space *log* $\mathcal{K}$*, and (iii) whether it supports an explicit mechanism for self-extension or creative novelty (our notion of* Wisdom*). We briefly compare five influential approaches to consciousness or self-monitoring with the present axiomatic proposal.*

15.1 Integrated Information Theory (IIT)

Tononi's IIT equates phenomenal consciousness with the amount of integrated information (Φ) *produced by a causal network [1], [2]. While IIT shares our focus on the structure of state transitions, it tightly couples consciousness to a* ***specific*** *physical substrate and does not distinguish between internal and external sources of change.*

15.2 Modal Logics of Self-Knowledge

Gamez [3] proposes a modal calculus capable of expressing nested epistemic claims such as "the system knows that it knows." This resonates with our concept of Potentially Aware Space*, though it lacks an explicit separation between internal and external transformations.*

15.3 Meta-Management Architectures

Sloman & Chrisley [5] distinguish reactive, deliberative, and meta-management layers in cognitive agents. The meta-layer monitors both internal and external events—a conceptual analogue to our sets Φ_{int} *and* Φ_{ext}*. However, their approach is architectural, not axiomatic.*

15.4 Algorithmic Self-Modelling

Schmidhuber's Gödel machine [6] rewrites its own code if it can prove the change will increase expected utility. This self-modification aligns conceptually with our notion of Wisdom*, though its foundation is proof-theoretic rather than sensitivity-based.*

15.5 Φ-Causal Set Formulations

Barrett & Seth [7] recast IIT in set-theoretic terms, yielding a more abstract formulation, but still centered around measures of integrated information.

15.6 Survey Evidence for a Missing Axiomatic Framework

A recent comprehensive survey of AI consciousness research by Shueb and Che [12] concludes that existing approaches—ranging from IIT and global workspace models to formal logics, self-models, and meta-cognitive architectures—remain descriptive rather than operational. The review explicitly notes the absence of (i) a causal distinction between internally and externally generated state transitions, (ii) a persistent and formally defined Potentially Aware Space *log, and (iii) any necessary and sufficient criterion for self-awareness. These gaps directly motivate the axiomatic structure developed in the present work.*

16 Summary

The Chinese Room argument demonstrates that behaviour alone is not a sufficient indicator of understanding. The AToS framework extends this insight by proposing a structural criterion for self-awareness based on recognised state transitions and internal causal attribution. In this framework, self-awareness is not a matter of symbol manipulation or behavioural performance, but of the system's ability to recognise at least one of its own state changes and to represent itself as the internal cause of that change. This provides a substrate-independent and operational criterion for minimal self-awareness that applies equally to biological, artificial, and abstract systems.

Comparison of AToS with Selected Approaches to Consciousness / Self-Awareness

Concept / Criterion	IIT (Tononi)	Modal Logic (Gamez)	Meta-Management (Sloman)	Self-Model (Schmidhuber)	Φ-Set Theory	AToS (This work)
Internal vs. external causal attribution	Not defined as a central criterion	Not defined	Discussed indirectly through monitoring and control	Implicit in the self-model	Not defined	**Defined axiomatically**
Recognition of internal state transition	Not explicit	Not explicit	Monitoring of internal events (architectural)	Implicit	Not explicit	**Defined formally**
Self as an internal causal role	Not formalised	Not formalised	Partial architectural interpretation	Self-model representation	Not formalised	**Defined formally**
Minimal operational criterion for self-awareness	No single criterion	Logical formalism only	No necessary-and-sufficient criterion	No necessary-and-sufficient criterion	No operational criterion	**Necessary and sufficient criterion proved within the framework**
Separation of internal and external stimuli	Not central	Not central	Partial	Partial	Not central	**Explicit (Stimulus Set / *StSe*)**
Detection function for recognised transitions	Not defined	Not defined	Implicit	Implicit	Not defined	**Defined (Δ)**
Formal state-space representation	Φ-space (information structure)	Modal structures (possible worlds, accessibility)	Architectural components (agents, layers)	Internal model (compressed representations)	Φ-set mathematical formalism	**Ω, Σ, S, Δ representation**
Operational test	No universal operational test	No operational test	No operational test	No operational test	No operational test	**GSAT formally defined; empirical validation planned in SenAi**
Falsifiability by intervention	Limited	Logical analysis only	Limited	Limited	Limited	**Explicitly incorporated into the framework**
Intended contribution	Information integration theory	Logical formalisation of modality	Cognitive architecture for control and metacognition	Self-model for compression and prediction	Mathematical abstraction of conscious structures	**Axiomatic operational framework for minimal self-awareness as internal causal attribution**

Table: Conceptual comparison of selected approaches. The table compares the explicit formal scope of each framework as presented in the original publications. It does not claim empirical superiority or invalidate any of the compared theories. "Defined" indicates that the concept is introduced explicitly as part of the formal framework. "Not defined" means that no equivalent explicit construct has been identified in the referenced work, although related ideas may exist under different terminology.

Figure 1: Comparison of AToS with selected approaches to consciousness and self-awareness.

References

[1] *G. Tononi, "An information integration theory of consciousness,"* BMC Neuroscience*, vol. 5, art. 42, 2004.*

[2] *G. Tononi et al., "IIT 4.0: The integrated information theory of consciousness,"* Biologically Inspired Cognitive Architectures*, 2023.*

[3] *D. Gamez, "The informational mind and the information integration theory of consciousness,"* International Journal of Machine Consciousness*, vol. 6, no. 1, pp. 21–28, 2014.*

[4] *D. Gamez,* Human and Machine Consciousness*. Open Book Publishers, 2018.*

[5] *A. Sloman and R. Chrisley, "Virtual machines and consciousness,"* Journal of Consciousness Studies*, vol. 10, no. 4–5, pp. 133–172, 2003.*

[6] *J. Schmidhuber, "Gödel machines: Self-referential universal problem solvers making provably optimal self-improvements," in* Artificial General Intelligence*, Springer, 2007.*

[7] *A. B. Barrett and A. K. Seth, "Practical measures of integrated information for time-series data,"* PLoS Computational Biology*, vol. 7, no. 1, 2011.*

[8] *M. Blum and L. Blum, "A theoretical computer science perspective on consciousness,"* Journal of Artificial Intelligence and Consciousness*, vol. 8, no. 1, 2021.*

[9] *J. A. Reggia, G. Katz, and G. Davis, "Artificial conscious intelligence,"* Journal of Artificial Intelligence and Consciousness*, vol. 7, no. 1, pp. 95–107, 2020.*

[10] *G. W. Ng and W. C. Leung, "Strong artificial intelligence and consciousness,"* Journal of Artificial Intelligence and Consciousness*, vol. 7, no. 1, pp. 63–72, 2020.*

[11] *P. Butlin, et al., "Consciousness in Artificial Intelligence: Insights from the Science of Consciousness," arXiv:2308.08708, 2023.*

[12] *M. M. Shueb and X. Che, "Consciousness in AI Systems: A Review,"* International Journal of Artificial Intelligence and Applications*, vol. 16, no. 2, pp. 1–19, 2025.*