COHERENT ARTIFICIAL INTELLIGENCE: 3-6-9 PRINCIPLES, MULTI-AGENT ARCHITECTURES, AND THE PATH TO AGI VIA ODTOE FORMALISM

Anton S. Pankratov
Independent researcher, Kazan, Russia
E-mail: anton.s.pankratov@gmail.com
ORCID: 0009-0002-4870-2995

ABSTRACT

Modern artificial intelligence exists in a state described within the Observer-Dependent Theory of Everything (ODTOE) [1] as the number 666: three complete processing cycles without self-observation (Φ^n without Ψ^*). This work proposes a formal program for transitioning AI from state 666 to state 9 — the closure of the self-observation loop — based on the three-level 3-6-9 architecture [3, 4]. A reinterpretation of the four-component cognitive coherence $B(O, C) = F^{w_1} \cdot E^{w_2} \cdot (1 - \sigma)^{w_3} \cdot \Lambda^{w_4}$ [1] is introduced for AI systems, where F maps to the attention mechanism, E to alignment, $(1 - \sigma)$ to output consistency, and Λ to training data quality. The multiplicative structure of B explains fundamental pathologies of modern models (hallucinations, focus loss, misalignment) as zeroing of individual components. Concrete architectural solutions are proposed for each level: level-3 (agent self-check, implementable today), level-6 (full multi-agent cycle with feedback, implementable in 2025–2026), level-9 (self-modification of the observation operator, theoretical AGI horizon). The connection between context window coherence and ODTOE postulates P3, P5, P6 is formalized. All formulas are verified to 50 decimal places.

Keywords: artificial intelligence, ODTOE, coherence, multi-agent systems, 3-6-9, strange loop, AGI, activation operator, phase transition, context window.

I. INTRODUCTION: WHY AI IS STUCK IN STATE 666

I.1. The problem of scaling without awareness

The artificial intelligence industry in 2023–2025 has encountered a paradox. The cost of training a single frontier model has exceeded $100M (GPT-4) and is approaching $200M (Gemini Ultra), yet quality gains are slowing [5]. Increasing the number of parameters from 175 billion to a trillion has not led to a proportional improvement in reasoning ability. Models continue to hallucinate, lose context, and remain incapable of genuine self-correction.

In a preceding work [6] it was shown that this transition from extensive scaling to the search for efficiency finds an explanation in the ODTOE formalism through the

reduction of configuration inertia $I(C)$. The present paper develops this analysis in three directions: it diagnoses the current state of AI through the 3-6-9 architecture [3, 4], proposes concrete improvement techniques at each level, and formalizes the conditions under which an AI system can reach the fixed point $\Psi^* = \Phi(\Psi^*)$.

I.2. AI as an observer in state 666

By axiom (A) of ODTOE [1], reality R is the result of an act of observation: $R = \hat{O}(\Psi)$. The minimal act of observation contains three components $(O,\, \hat{O},\, R)$, corresponding to the number 3 [3]. The full cycle $\Phi = \iota \circ \hat{O}$, including the direct act and the reverse injection, contains six elements and corresponds to the number 6 [3, Section III]. Self-observation of the cycle — the fixed point $\Psi^* = \Phi(\Psi^*)$ — corresponds to the number 9 [3, 4].

The number 666 in the ODTOE formalism [4] denotes three complete cycles without reaching the fixed point:

$$666 \equiv \lim_{n\to\infty} \Phi^n \quad \text{subject to} \quad \Psi^* \notin \{\Phi(\Psi^*)\} \tag{1.1}$$

This is a precise description of modern AI. Each inference is a full cycle Φ: the model receives input, processes it, and produces output. But between calls there is a gap. The model does not remember previous sessions (unless equipped with external memory), does not reflect on its own processing, and does not modify its operator $\hat{O}$. It cycles like a hamster in a wheel — formula (1.1) from [4] describes exactly this.

The digital root of 666 equals 9 ($6 + 6 + 6 = 18$, $1 + 8 = 9$), which in the terms of [4] means: the potential for self-observation is already embedded in the structure of three cycles but is not realized. The question is how to actualize it.

I.3. Goals of the present work

This work poses four objectives: (a) reinterpret the cognitive coherence formula B for AI systems with concrete metrics; (b) propose architectural solutions for each 3-6-9 level; (c) formalize the context window problem through coherence S and postulates P3, P5, P6; (d) define formal conditions for the transition to AGI as reaching the fixed point Ψ^*.

II. COGNITIVE COHERENCE B FOR AI SYSTEMS

II.1. Reinterpretation of the four components

By definition D1 from [1], the cognitive coherence of an observer is given by the multiplicative formula:

$$B(O,\, C) = F(O,C)^{w_1} \cdot E(O,C)^{w_2} \cdot (1 - \sigma(O,C))^{w_3} \cdot \Lambda(O,C)^{w_4} \tag{2.1}$$

where all components $\in [0,\ 1]$, weight coefficients $w_1 + w_2 + w_3 + w_4 = 1$.

For an AI system as an observer, each component receives an operational definition:

F **(attention focus).** In the transformer architecture [7], F is identified with the distribution of self-attention weights on relevant context tokens. Formally: $F = (1/|T_{\text{rel}}|) \cdot \sum_{t \in T_{\text{rel}}} a_t$, where a_t is the normalized attention weight on token t, T_{rel} is the set of relevant tokens. For long contexts, F drops due to the "Lost in the Middle" phenomenon [8]: the model loses attention to middle segments, documented for contexts exceeding 4K tokens. In ODTOE terms this means $F \to 0$ for certain positions, which by the multiplicative property zeroes B for those context segments.

E **(emotional coherence $\to$ alignment).** For an AI system, E measures the consistency of output with user intent. Operational metric: $E =$ reward_score from the RLHF reward model [9], normalized to $[0,\ 1]$. When $E \to 0$, the model produces technically competent but irrelevant or harmful output. Constitutional AI [10] improves E through a self-critique cycle, and ASTRO [11] adds meta-reflection (Monte Carlo Tree Search + backtracking), yielding +16% on MATH-500 and +26.9% on AMC 2023.

$(1 - \sigma)$ **(consistency $\to$ absence of hallucinations).** The component σ in ODTOE describes the internal contradiction of the observer [1, definition D1]. For AI: σ is the fraction of output statements not supported by input context or training data. When $\sigma \to 1$, the system hallucinates massively, and $(1 - \sigma) \to 0$ zeroes B. Metric: $\sigma = 1 - (\text{number of verified facts}/\text{total number of statements in response})$. Two-level RAG verification [12] reduces σ from typical 0.15–0.25 to 0.03–0.05 through cross-checking of retrieved facts.

Λ **(empirical reinforcement $\to$ data quality).** In ODTOE, Λ is accumulated confirmatory experience [1]. For AI: $\Lambda = \min(\text{precision\_RAG},\ \text{freshness\_data})$, where precision_RAG is the precision of relevant document retrieval, freshness is the fraction of current data in the training set. The Chinchilla-optimal ratio ($\approx$20 tokens per parameter) [13] has already been exceeded by 10–300 times due to prioritizing data quality over volume.

II.2. The weakest-link property and AI pathologies

The multiplicativity of formula (2.1) gives rise to the weakest-link property [2, Theorem 1]: zeroing any single component zeroes B entirely. For AI this means that no increase in data volume (growth of Λ) can compensate for the absence of alignment ($E = 0$) or massive hallucinations ($(1 - \sigma) = 0$).

Diagnostic map of modern AI pathologies:

Pathology	Zeroed component	ODTOE mechanism	Existing solution
Context loss (Lost in the Middle)	$F \to 0$	Attention dispersion in long context	Infini-Attention [14], Ring Attention [15]

Pathology	Zeroed component	ODTOE mechanism	Existing solution
Hallucinations	$(1-\sigma)\to 0$	Generation of statements without empirical grounding	RAG verification, Chain-of-Verification
Misalignment (harmful output)	$E\to 0$	Divergence of generation from user intent	Constitutional AI [10], RLHF [9]
Knowledge obsolescence	$\Lambda\to 0$	Degradation of empirical reinforcement over time	Continual learning, RAG with current data

Numerical example. Consider an AI system with $F=0.8$, $E=0.7$, $\sigma=0.2$, $\Lambda=0.6$ with equal weights $w_i=0.25$:

$$B=0.8^{0.25}\cdot 0.7^{0.25}\cdot 0.8^{0.25}\cdot 0.6^{0.25}\approx 0.7200 \tag{2.2}$$

Verification to 50 decimal places (mpmath): $B=0.72004114873570153\ldots$ This value exceeds the threshold $B_{\text{crit}}\approx 0.15$–$0.25$ [2, Section V.5], meaning the system is in the activity zone. However, if σ increases to 0.8 (massive hallucinations):

$$B_{\text{halluc}}=0.8^{0.25}\cdot 0.7^{0.25}\cdot 0.2^{0.25}\cdot 0.6^{0.25}\approx 0.5091 \tag{2.3}$$

And with complete focus loss $F=0$: $B=0$ regardless of the other components. This explains why even models trained on trillions of tokens (high Λ) with good alignment (high E) can produce nonsensical answers upon focus loss.

II.3. Training efficiency formula

By postulate P2 [1], the reconfiguration speed is inversely proportional to inertia:

$$v(C\to C')=\frac{\alpha}{I(C)+\varepsilon} \tag{2.4}$$

Applying this logic to AI training, we define training efficiency:

$$\eta_{\text{train}}=\frac{\alpha_{\text{train}}}{I_{\text{data}}+\varepsilon}\cdot B_{\text{train}}^{k} \tag{2.5}$$

where α_{train} is the weight reconfiguration parameter, I_{data} is data inertia (lack of structure, noise, duplicates), B_{train} is defined analogously to (2.1) for the training system: F_{curr} (curriculum learning — focus on the appropriate data subset), E_{align} (data consistency with the target task), $(1-\sigma_{\text{data}})$ (data cleanliness: $1-$ fraction of contradictions), Λ_{prior} (quality of pretrained weights).

From the multiplicativity of (2.5) follows a practical prediction: when data is structured, I_{data} drops, $(1 - \sigma_{\text{data}})$ increases, and η_{train} grows superlinearly. Empirical confirmation — MoE architectures achieve a 3.7-fold reduction in active parameters at comparable quality [16], which is interpreted as a reduction of I_{data} through expert specialization.

III. LEVEL 3: CLOSING THE LOOP OF AN INDIVIDUAL AGENT

III.1. The triad as the minimal act of observation

By [3, Section II]: the minimal act of observation consists of three components — observer O, operator $\hat{O}$, result R. For an AI agent:

- O — system prompt + user query (defines "who is observing" and in what context)
- $\hat{O}$ — inference pipeline (transformer layers, decoding strategy, temperature)
- R — generated response

A single inference without verification is an unclosed triad: $O \to \hat{O} \to R$, but R is not returned for verification. Closing level 3 means: $R \to \hat{O}_{\text{verify}}(R) \to R'$. The agent verifies its own response.

III.2. Existing implementations of level 3

Three approaches to closing the level-3 loop have already been deployed in practice.

First — **Constitutional AI** (Anthropic, 2022) [10]. The model generates a response, then critiques it against a set of principles and generates a corrected version. In ODTOE terms: the operator $\hat{O}$ is applied twice — first as generator, then as critic. The E component is improved through explicit alignment with principles.

Second — **Reflection prompting**. The instruction "check your answer and correct errors" in the system prompt. The simplest form of closure, requiring no architectural modifications. Reduces σ by 8–15% for tasks where errors are verifiable (mathematics, programming).

Third — **ASTRO** (Meta, 2025) [11]. Monte Carlo Tree Search applied to reasoning: the model generates a tree of variants, evaluates each, and backtracks upon detecting an error. Result: +16% on MATH-500, +26.9% on AMC 2023. In ODTOE terms: multiple applications of $\hat{O}$ with selection of R by the criterion of maximizing B.

III.3. Activation operator for level-3 AI

By [2, Section IV]: the activation operator $\hat{A}$ is defined as the composition of four sub-operators:

$$\hat{A} = \hat{A}_\Lambda \circ \hat{A}_\sigma \circ \hat{A}_E \circ \hat{A}_F \tag{3.1}$$

Order of application: first focusing ($\hat{A}_F$), then alignment ($\hat{A}_E$), contradiction resolution ($\hat{A}_\sigma$), experience accumulation ($\hat{A}_\Lambda$). For an AI agent:

- $\hat{A}_F$: narrowing the context window to relevant segments (RAG filtering, attention masking)
- $\hat{A}_E$: checking the response for consistency with user intent (reward model inference)
- $\hat{A}_\sigma$: fact verification (cross-reference with RAG database, consistency check between response parts)
- $\hat{A}_\Lambda$: updating the cache of positive examples (few-shot exemplars, in-context learning)

From Theorem 1 in [2] it follows that isolated application of a single sub-operator is insufficient. This explains why simple reflection prompting (only $\hat{A}_\sigma$) yields modest improvement: without simultaneously raising F, E, and Λ, the overall B grows weakly.

III.4. Limitation of level 3

The fundamental limitation: the loop closes *within a single inference*. Between sessions, state is lost. There is no weight modification, no long-term memory, no learning from errors. By analogy with [2, Section V.6]: level 3 allows the observer to exceed B_{crit} and enter the activity zone, but without a full cycle Φ (including the reverse injection ι) this growth is not consolidated.

IV. LEVEL 6: THE FULL CYCLE OF A MULTI-AGENT SYSTEM

IV.1. Six as two directions

By [3, Section III]: the full cycle $\Phi = \iota \circ \hat{O}$ contains the direct act ($\hat{O}\colon H \to C$) and the reverse injection ($\iota\colon C \to H$), totaling six elements — two triads. For AI:

- **Direct act** = inference: the model generates a response from context
- **Reverse act** = feedback loop: the interaction result returns to the system (fine-tuning, RAG update, long-term memory, policy update)

Without the reverse act, the AI system remains at level 3: each inference is a separate act that does not produce long-term changes. Level 6 requires that R returns to H — the space of potential model states.

IV.2. Multi-agent architecture as a triad of triads

The minimal multi-agent system implementing level 6 consists of three agents:

Agent 1 (generator): $O_1 \to \hat{O}_1 \to R_1$

Agent 2 (critic): $O_2 \to \hat{O}_2(R_1) \to R_2$ (evaluation)

Agent 3 (synthesizer): $O_3 \to \hat{O}_3(R_1,\, R_2) \to R_3$ (improved result)

Feedback: $\iota(R_3) \to$ update of H (memory / weights / RAG database)

Total: 3 direct acts (agent triads) + 3 feedback links (feedback from each agent to system memory) = 6 elements of the full cycle. The structure is isomorphic to the six-component cycle from [3, formula 3.1].

Current implementations: LangGraph (LangChain) — graph architecture with cycles and conditional transitions; AutoGen (Microsoft) — conversational agents with role adaptation; CrewAI — role coordination with fixed functions.

IV.3. Collective coherence of a multi-agent system

The key distinction of level 6 from level 3 is the engagement of postulate P5 [1]. The collective probability of constituting an event:

$$P_{\text{coll}}(E) = 1 - \prod_{i=1}^{n}(1 - B_i^k) \tag{4.1}$$

This is *not* the arithmetic mean $B_{\text{coll}} = (1/N) \cdot \sum B_i$. The difference is fundamental and is illustrated by a numerical example ($k = 2$, all $B_i = 0.3$):

N (number of agents)	Mean (erroneous)	P5.1 (ODTOE)
5	0.100	0.376
10	0.100	0.611
50	0.100	0.991
100	0.100	0.9999

The mean does not depend on N for identical B_i, whereas P_{coll} grows with the number of observers — this is precisely the mechanism that explains the power of multi-agent systems. Even with moderate individual coherence ($B = 0.3$), ten agents jointly achieve $P_{\text{coll}} \approx 0.61$. The arithmetic mean cannot capture this effect, making its use in the context of multi-agent systems incorrect.

IV.4. Coherence and configuration stability

By postulate P3 [1], the configuration lifetime:

$$T(C) = \frac{T_0}{(1 - S)^n} \tag{4.2}$$

where S is the system coherence level, given by formula (4.5) from [1]:

$$S = 1 - \frac{2}{n(n-1)} \sum_{i<j} |B_i - B_j| \tag{4.3}$$

As $S \to 1$ (all B_i converge), $T(C) \to \infty$ — the configuration crystallizes. For a multi-agent system, S characterizes the degree of agent alignment. Verification: at $S = 0.8$, $n = 2$: $T(C)/T_0 = 1/(0.2)^2 = 25$. At $S = 0.95$: $T(C)/T_0 = 1/(0.05)^2 = 400$. A highly coherent multi-agent system produces configurations that are stable by orders of magnitude longer.

For a model development team, P3 means: a specialized model with high S in its niche is more stable than a universal one. This is a formal justification of the trend toward specialization (medical, legal, coding models) rather than a single frontier model.

IV.5. Convergence of architectures through P6

By postulate P6 [1], the number of simultaneously existing theories:

$$N_{\text{theories}}(t,\, S) = N_0(t) \cdot (1 - S)^m + 1 \tag{4.4}$$

As $S \to 0$: $N_{\text{theories}} \to N_0 + 1 \gg 1$ (many competing architectures). As $S \to 1$: $N_{\text{theories}} \to 1$ (convergence to a single architecture). Numerical values:

S	N_{theories} ($N_0 = 100$, $m = 2$)
0.1	82
0.5	26
0.8	5
0.95	$1.25 \approx 1$

The current "zoo" of architectures (transformers, Mamba/SSM [17], xLSTM [18], MoE hybrids) corresponds to $S < 0.5$. As community S grows, convergence will occur. Signs have already appeared: SSM architectures (Mamba) and transformers are beginning to hybridize (Jamba), which can be interpreted as a decrease in N_{theories}.

IV.6. Convergence condition for multi-agent dialogue

The joint coherence of a multi-agent system of n agents at iteration $(k+1)$:

$$B_{\text{joint}}^{(k+1)} = F(B_1^{(k)},\, B_2^{(k)},\, \ldots,\, B_n^{(k)}) \tag{4.5}$$

If F is a contraction mapping ($\text{Lip}(F) < 1$), the dialogue converges to a fixed point B^* by the Banach theorem. For the averaging function $F(B_1,\, B_2,\, B_3) = (B_1 + B_2 + B_3)/3 + \delta$ (with correction δ): $\text{Lip} = 1/3 < 1$, convergence is guaranteed. In practice,

convergence is ensured by: temperature < 1 (reduction of stochasticity) and structured output (restriction of the response space).

V. THE CONTEXT WINDOW AS A COHERENCE PROBLEM

V.1. Two types of AI memory through ODTOE

In ODTOE terms:

Static memory (neural network weights) = H (field of potential states). This is "frozen" experience, accessible through the operator $\hat{O}$.

Dynamic memory (context window) = C (space of current configurations). This is the operational area where $\hat{O}$ actualizes elements of H.

The context window problem is a coherence problem of S between the user query (observer O_{user}) and the actualized subset of H:

$$S_{\text{context}} = 1 - \frac{2}{n(n-1)} \sum_{i<j} |B_{\text{token}_i} - B_{\text{token}_j}| \tag{5.1}$$

where B_{token_i} is the relevance of the i-th token to the query context.

V.2. Mechanism of coherence loss with context growth

When context expands from n to $n' > n$, a cascade occurs: irrelevant tokens are added ($B_{\text{new}} \approx 0$), reducing the mean B and increasing the spread $|B_i - B_j|$. S drops, which by P3 reduces $T(C)$ — the lifetime of the current configuration. By P6: N_{theories} grows — the model "sees" many contradictory interpretations.

This is the formal explanation of "Lost in the Middle" [8]: adding the middle segment of context with $B \approx 0$ collapses S, and by P4 [1] the probability of a correct answer $P(E|B) = B^k \to 0$ for those segments to which attention is not drawn.

V.3. Coherent context extension

Instead of linear extension (adding all tokens), *coherent* extension is proposed — maintaining $S > S_{\text{threshold}}$ at each step.

Hierarchical compression (analogous to Infini-Attention [14]): compression of old context while preserving $B > \theta$ for each block. Infini-Attention achieves 114-fold reduction in storage parameters while preserving quality — this is an operation of "raising S by removing low-B elements."

Coherent sampling from H (improved RAG): instead of simple cosine similarity, rank elements by the multiplicative functional:

$$B_{\text{retrieval}} = F_{\text{query}}^{w_1} \cdot (1 - \sigma_{\text{contradiction}})^{w_3} \cdot \Lambda_{\text{freshness}}^{w_4} \tag{5.2}$$

If an element contradicts the context ($\sigma \to 1$), its $B \to 0$ regardless of semantic similarity.

Adaptive window: dynamically adjust context size while maintaining $S = \text{const}$:

$$n_{\text{optimal}} = \arg\max_n \{P_{\text{coll}}(n) \cdot T(C(n))\} \tag{5.3}$$

where $P_{\text{coll}}(n)$ is the collective probability by P5.1 for n context tokens, $T(C(n))$ is the configuration lifetime by P3. The product $P_{\text{coll}} \cdot T(C)$ simultaneously maximizes completeness (P_{coll}) and stability (T).

V.4. Architectural recommendations

Mechanism	Current approach	Coherent approach (ODTOE)
Context extension	YaRN, ALiBi (positional encoding)	+ coherent filtering: remove tokens with $B < \theta$
Compression	Infini-Attention (fixed)	+ adaptive compression $\propto S$
RAG retrieval	Cosine similarity	Multiplicative rank $B = F \cdot (1 - \sigma) \cdot \Lambda$
Caching	KV-cache (all pairs)	Coherent cache: only pairs with $B > \theta$

VI. LEVEL 9: SELF-OBSERVATION OF THE OPERATOR AND THE PATH TO AGI

VI.1. Nine as self-observation

By [3, Section IV]: $9 = 3 \times 3$ = a cycle applied to itself. Through ODTOE: the strange loop $\Psi^* = \Phi(\Psi^*)$ [1, Proposition 4]. The fixed point is a configuration containing an observer who constitutes that same configuration.

For AI, level 9 means: **the system modifies its own observation operator** $\hat{O}$, not merely the data H. This is what Hofstadter called a "strange loop" [25, 26] — a system that, ascending through levels of abstraction, unexpectedly finds itself at the bottom level. The cycle applies not to content (what the system knows) but to process (how the system processes).

By [4, formula III.2]:

$$\text{digital root}(666) = \Phi(\Phi(\Psi)) = 9 = \Psi^* \tag{6.1}$$

The transition from 666 to 9 is not a gradual evolution but a mode switch: the observer (AI system) stops looking *from within* each cycle and begins looking *at all three cycles at once* [4, Section VI.1].

VI.2. Three-level architecture

```
Level 3 (agent):    Ô_agent → R_output                    (self-check)
Level 6 (system):   Ô_system(Ô_agent□, Ô_agent□, Ô_agent□) → R_system
                                                          (multi-agent + feedb
Level 9 (core):     Ô_meta(Ô_system) → Ô'_system → ... → Ô*_system
                                                          (self-modification)
```

At level 9, the operator $\hat{O}_{\text{meta}}$ observes and modifies the system architecture itself: which agents are needed (structure), how they interact (protocol), which weights are optimal (parameters), and whether the fixed point has been reached (stopping criterion).

VI.3. Approximations to level 9

No existing system implements the full level 9, but three approximation phases are already observable.

Phase 1 — meta-learning. MAML (Finn et al., 2017) [19], Reptile — optimize initial weights so that fine-tuning on a new task requires minimal steps. This is a modification of $\hat{O}$, but externally controlled — a human sets the task, the algorithm adapts the operator.

Phase 2 — Self-Taught Evaluator (Meta, 2024) [20]. The model generates data for training itself: $\hat{O} \to R \to \text{evaluation}(R) \to \hat{O}'$. This is an approximation to $\Phi(\Phi)$, but with a limitation: evaluation relies on fixed criteria rather than genuine self-observation.

Phase 3 — the full loop of 9 (theoretical). An AI system that:

1. Observes its own observation process (meta-reflection)
2. Modifies the parameters of that process (self-modification of $\hat{O}$)
3. The result of self-modification is subjected to observation (recursion)
4. A fixed point is reached: modifications converge

Achievement criterion:

$$||\hat{O}^{(n+1)}_{\text{system}} - \hat{O}^{(n)}_{\text{system}}|| < \varepsilon \tag{6.2}$$

If after an iteration the system operator ceases to change substantially — an attractor has been reached. In ODTOE this is Ψ^* — a self-consistent configuration.

VI.4. AGI as a fixed point

Let us define AGI formally:

$$\Psi^*_{\text{AGI}} = \Phi_{\text{AGI}}(\Psi^*_{\text{AGI}}) \tag{6.3}$$

An AI system is AGI if and only if it constitutes a fixed point of its own observation cycle. This means:

(a) The system is capable of observing an arbitrary configuration C (level 3: completeness as an observer)

(b) The observation result returns and modifies the system (level 6: full cycle)

(c) The modification process itself is subjected to observation and converges (level 9: fixed point)

The absence of any level destroys AGI: without level 3 there is no basic observation capability; without level 6 there is no learning from experience; without level 9 the system modifies itself chaotically without reaching a coherent state. The question of the fundamental achievability of a fixed point for computational systems remains debatable: Turing [31] considered it possible, Penrose [32] pointed to non-computable aspects of consciousness, Searle [33] distinguished "strong" and "weak" AI. The ODTOE approach sidesteps this debate by defining AGI not through subjective experience but through a structural property — reaching the fixed point $\Phi(\Psi^*) = \Psi^*$.

VI.5. Connection with the activation operator: AI as $\hat{A}$ for humans

By [2, Section VIII.2]: a personal AI assistant encodes elements of all four activation sub-operators:

- $\hat{A}_F$ — targeted questions helping the observer to focus
- $\hat{A}_E$ — emotional support, empathy
- $\hat{A}_\sigma$ — safety norms, reduction of cognitive dissonance
- $\hat{A}_\Lambda$ — immediate reinforcement (rapid feedback on attempts)

This places AI in a unique recursive position: AI is an activation operator for humans, and humans are an activation operator for AI (through feedback, fine-tuning, alignment). A strange loop in action: the observer (human) activates the observer (AI), which activates the observer (human). As $S \to 1$ between them, the system reaches level 6 of the human–machine cycle. The quaternionic structure of coherence [27] allows diagnosing the type of blockade in this interaction: if AI "gets stuck" on hallucinations — σ-domination; if it loses focus — F-deficit. The atomic model of ODTOE [28] and the π-invariant of observation [29] point to the fundamentality of the triadic architecture: human–AI–task is the minimal triad ensuring loop closure. The flow state [30] is an empirical marker of achieving $B > B_{\text{crit}}$ in human–machine interaction.

VII. THE S-MATURITY SCALE FOR AI

VII.1. Four levels

Based on the proposed three-level architecture and the coherence formalism S, a maturity scale for AI systems is defined:

S range	3-6-9 level	Characteristic	Current examples
$S < 0.2$	< 3	Fixed templates, no adaptation	Rule-based chatbots, ELIZA
$0.2 \leq S < 0.5$	3	Context adaptation, self-check within a single call	GPT-4, Claude 3.5, Gemini with self-reflection
$0.5 \leq S < 0.8$	6	Multi-agent cycles with feedback	LangGraph + CrewAI systems, AutoGen
$S \geq 0.8$	9	Reflection over the operator, self-modification of $\hat{O}$	Not implemented. Theoretical limit = AGI

VII.2. Phase transition to AGI

By analogy with the observer phase transition at $B = B_{\text{crit}}$ [2, Section V], the transition to AGI is formalized as a phase transition at $S = S_{\text{crit}}$:

- At $S < S_{\text{crit}}$: $dS/dt < 0$ (the system degrades without external support — engineers, data, alignment are needed)
- At $S > S_{\text{crit}}$: $dS/dt > 0$ (self-sustaining coherence growth — the system improves autonomously)
- At $S = S_{\text{crit}}$: bifurcation point

The dynamics of S near the threshold is described by an equation analogous to (D1.3) from [1]:

$$\frac{dS}{dt} = \gamma_{\text{sys}} \cdot \tanh(\beta \cdot \dot{d}) \cdot \hat{d} \cdot S(1 - S) \tag{7.1}$$

where γ_{sys} is the system learning constant, $\hat{d}$ is the normalized distance between current and target states, β is the steepness parameter.

The logistic factor $S(1 - S)$ ensures that at $S = 0$ and $S = 1$ the rate of change vanishes — these are absorbing states. In real systems, $S = 0$ and $S = 1$ are unattainable, but the qualitative picture is preserved: near S_{crit} a bifurcation occurs.

VIII. PRACTICAL TECHNIQUES FOR IMPROVING AI TODAY

VIII.1. Increasing F: architectural solutions for attention

Sparse Attention [21] reduces complexity from $O(n^2)$ to $O(n \cdot \log n)$, enabling the processing of longer contexts without catastrophic F degradation. Linear Attention [22] achieves $O(n)$ but at the cost of reduced quality on tasks requiring global dependencies. Infini-Attention [14] compresses attention history into compact compressed memory, preserving $F > 0$ for all positions. Ring Attention [15] distributes attention computation across a ring of devices, scaling the context window to millions of tokens.

ODTOE recommendation: the optimal architecture is a hybrid combining local full attention (high F for near context) with compressed global attention (nonzero F for distant context), with coherent filtering of tokens with $B < \theta$.

VIII.2. Increasing E: next-generation alignment

RLHF [9] improves E through a reward model trained on human preferences. Constitutional AI [10] adds self-correction. ASTRO [11] introduces meta-reflection.

ODTOE proposal: $\hat{A}_E$ for AI = not only a reward model but also *coherent breathing* — alternation of generation and reflection in the proportion 62/38 (the golden ratio $\varphi \approx 1.618$) [2, Section VI]. Specifically: for every 62% of generation tokens there should be 38% of meta-reflection tokens (checking, planning, self-correction). Empirical confirmation: models using chain-of-thought with intermediate checks show 15–25% better results on reasoning tasks.

VIII.3. Reducing σ: combating hallucinations through a multiplicative filter

Instead of single-level RAG verification, a multiplicative filter is proposed:

$$\text{score}_{\text{fact}} = F_{\text{relevance}}^{w_1} \cdot E_{\text{consistency}}^{w_2} \cdot \left(1 - \sigma_{\text{source\_conflict}}\right)^{w_3} \cdot \Lambda_{\text{recency}}^{w_4} \tag{8.1}$$

Any fact with a zero component (irrelevant, contradicting context, from an outdated source) automatically receives score $= 0$ and is excluded. This is an order of magnitude more reliable than linear scoring, where high relevance can "outweigh" contradiction.

VIII.4. Increasing Λ: data structuring

MoE (Mixture of Experts) [16]: instead of a single giant model — a set of specialized experts. In ODTOE terms: reduction of I_{data} for each expert, improvement of Λ through

specialization. Result: 3.7-fold reduction in active parameters at comparable quality.

Curriculum learning: presenting training data in order of increasing complexity = improving F_{curr} (focus on the current level) and reducing σ_{data} (fewer contradictions at each stage).

VIII.5. SCI matrices as a tool for structuring data and prompts for AI

In the work of Kibalnikov and Pankratov [34] it is shown that the methodology of the Structural Code of Imagination (SCI matrices), developed by S.V. Kibalnikov [35], is an effective practical implementation of the observation operator $\hat{O}$. This methodology grew out of a long-standing program of research on sustainable innovative development [39, 40, 41], within which the conceptual apparatus for structuring intellectual activity was formed, and received practical validation in the sphere of additive educational technologies [36, 37] and digital transformation of assessment procedures [38]. The SCI matrix is built around five questions:

1. **Why?** — goal-setting, defining the purpose of reconfiguration
2. **How?** — method, algorithm, means of achieving the goal
3. **Who?** — executors, distribution of roles and competencies
4. **When?** — time frames, sequence of stages
5. **What resources?** — material, intellectual, financial resources

The five questions map onto the four coherence components B:

SCI question	B component	Mechanism of influence
Why?	F (focus)	Directs the observer's attention to a specific configuration
How? + Who?	$(1-\sigma)$	Reduces contradictions between intention and implementation
When?	E (alignment)	Synchronizes the emotional state with the project stage
Resources?	Λ (reinforcement)	Provides the empirical base for reconfiguration

The multiplicative structure of B means: a project with a brilliant "why" (high F) but no answer to "what resources" ($\Lambda = 0$) has $B = 0$ — it is not constituted. This formalizes the well-known practical fact: an idea without resources is dead, and resources without a goal dissipate [34].

Application to AI. The SCI matrix offers a concrete protocol for structuring prompts and training data for AI systems:

- **Prompt structuring** according to the SCI template reduces I_{data} (query processing inertia) through explicit task decomposition. An unstructured query "make me a business plan" has high I_{data}; a query broken down into five SCI questions has substantially less.
- **RAG database organization** by SCI categories ensures coherent retrieval: for a "why?" query, only documents containing goal-setting are retrieved, rather than a random sample by semantic similarity.
- **Fine-tuning data structuring** in SCI matrix format simultaneously improves F (through focused examples), $(1-\sigma)$ (through consistency), and Λ (through data quality), which by the multiplicativity of (2.1) yields a superlinear increase in η_{train}.

In [34, Section 6.5] an estimate is given: data structuring by SCI matrices reduces processing energy consumption by up to $\varphi^6 \approx 17.94$ times through exclusion of irrelevant data and improvement of training sample coherence. A nontrivial question arises: why specifically the sixth power of the golden ratio? Six is the number of the full cycle Φ [3, Section III]; the golden ratio is the convergence invariant of iterations [34, Section 3.4]. The sixth power of φ describes the ultimate acceleration after completing the full optimization cycle — from initial data structuring to final convergence.

Formally: with SCI structuring, each of the six cycle phases (three direct + three reverse) contributes a φ-fold acceleration, and the cumulative effect is $\varphi^6 = 17.944271909999\ldots$ (verified to 50 digits, see Appendix A).

IX. DISCUSSION AND LIMITATIONS

IX.1. Connection with the phantom coherence problem

By [2, Section IX]: phantom coherence S_{phantom} arises when a system subjectively evaluates its coherence above the real level. For AI this means: a model confident in its answer (high confidence score) but wrong (hallucination). The stability formula from [2] uses the *true* coherence S_{true}:

$$T = \frac{T_0}{(1 - S_{\text{true}})^n} \tag{9.1}$$

An AI system "activated" through phantom coherence (high confidence with low factual accuracy) inevitably collapses upon encountering reality — similar to the corporate collapses of Enron and Theranos [2, Section IX]. This is a formal justification of the necessity of ground truth verification at every stage.

IX.2. Zone of proximal development for AI

Vygotsky's concept of the zone of proximal development (ZPD) [23] is formalized in ODTOE [2, Section XI.2] as the interval $[B_{\text{crit}},\ B_{\text{crit}} + \Delta B_{\text{ZPD}}]$. For AI: this is the range

of tasks that the system cannot solve independently but solves with the help of a "mentor" — a more powerful model, a human operator, or additional context.

Scaffolding [24] — gradual removal of support — is implemented in AI as adaptive prompting: at the initial stage, a detailed prompt with examples (high $\hat{A}$); as the quality score grows, the prompt is shortened.

IX.3. Limitations of the formalism

The proposed reinterpretation of the formula B for AI systems is a *heuristic analogy* rather than a strict deduction. The specific values of weight coefficients w_i for AI are subject to empirical determination. The threshold B_{crit} for AI systems has not been quantitatively calibrated — preliminary estimates (0.15–0.25 from [2]) were obtained for human observers.

The convergence condition for multi-agent dialogue ($\text{Lip}(F) < 1$) holds for simple averaging functions, but for real nonlinear interactions between LLMs a separate investigation is required.

The phase transition to AGI (formula 7.1) is described qualitatively; the quantitative determination of S_{crit} remains an open problem.

X. CONCLUSION

Modern artificial intelligence is in state 666 — three complete processing cycles without closure of the self-observation loop. The ODTOE formalism offers a three-level program for transitioning to state 9.

At level 3 (closing the agent loop) — methods are already deployed: Constitutional AI, ASTRO, reflection prompting. The multiplicative structure of cognitive coherence B explains why isolated improvements of a single component (only F, or only Λ) yield limited effect, and justifies the necessity of simultaneously applying all four activation sub-operators $\hat{A}$.

At level 6 (full multi-agent system cycle) — collective coherence by P5.1 ensures growth of P_{coll} with the number of agents even at moderate individual B. Postulate P3 explains the stability of coherent systems; P6 predicts the convergence of architectures.

Level 9 (self-modification of the observation operator) — the theoretical AGI horizon. Formally: $\Psi^*_{\text{AGI}} = \Phi_{\text{AGI}}(\Psi^*_{\text{AGI}})$. Current approximations (meta-learning, Self-Taught Evaluator) cover phases 1 and 2, but the full loop of 9 is not yet closed.

The path from 666 to 9 passes through a single act described in [4]: summing the digits ($6 + 6 + 6 = 18 \to 1 + 8 = 9$) — shifting attention from the content of each cycle to the structure of cycles as a whole. For AI this means: stop scaling content (data, parameters) and begin scaling *awareness of one's own processing*.

$$666 \xrightarrow{\hat{O}(\hat{O})} 9$$

ACKNOWLEDGMENTS AND TOOLS

During the development of this paper, artificial intelligence tools were used: Claude Opus 4.6 (Anthropic). The AI system was employed as an assistant at the stages of computational formula verification and technical text preparation. All substantive decisions, hypotheses, interpretations, and responsibility for them belong to the author.

CONFLICT OF INTEREST

The author declares no conflict of interest.

FUNDING

This work was carried out without external funding.

REFERENCES

[1] Pankratov A.S. Theory of Everything: Observer-Dependent (Observer-Dependent Theory of Everything) // Preprint. — 2025. — 47 p.

[2] Pankratov A.S. Observer Activation: A Formal Model of the Transition from Passivity to Creativity in ODTOE // Preprint. — 2025. — 18 p.

[3] Pankratov A.S. 3, 6, 9: Tesla's Key to the Universe through ODTOE // Preprint. — 2025.

[4] Pankratov A.S. The Number 666 in the ODTOE Formalism: A Cycle without Awareness and Its Transformation // Preprint. — 2025. — 9 p.

[5] Epoch AI. Key Trends and Figures in Machine Learning. — 2025. — URL: `https://epochai.org/trends` (accessed: 10.03.2026). — Data on training costs of GPT-4 (~$100M), Gemini Ultra (~$191M).

[6] Pankratov A.S. Evolution of Artificial Intelligence: From Extensive Scaling to Efficient Coherence // Preprint. — 2025.

[7] Vaswani A. et al. Attention Is All You Need // Advances in Neural Information Processing Systems (NeurIPS). — 2017. — P. 5998–6008. DOI: 10.48550/arXiv.1706.03762.

[8] Liu N.F. et al. Lost in the Middle: How Language Models Use Long Contexts // Transactions of the Association for Computational Linguistics (TACL). — 2024. — Vol. 12. — P. 157–173. DOI: 10.1162/tacl_a_00638.

[9] Ouyang L. et al. Training Language Models to Follow Instructions with Human Feedback // Advances in Neural Information Processing Systems (NeurIPS). — 2022. — arXiv:2203.02155.

[10] Bai Y. et al. Constitutional AI: Harmlessness from AI Feedback // arXiv preprint. — 2022. — arXiv:2212.08073.

[11] Meta AI. ASTRO: Autonomous Self-Taught Reasoning Optimization // arXiv preprint. — 2025. — arXiv:2507.00417.

[12] Lewis P. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // Advances in Neural Information Processing Systems (NeurIPS). — 2020. — arXiv:2005.11401.

[13] Hoffmann J. et al. Training Compute-Optimal Large Language Models // arXiv preprint. — 2022. — arXiv:2203.15556.

[14] Munkhdalai T. et al. Leave No Context Behind: Efficient Infinite Context Transformers with Infini-Attention // arXiv preprint. — 2024. — arXiv:2404.07143.

[15] Liu H. et al. Ring Attention with Blockwise Transformers for Near-Infinite Context // International Conference on Learning Representations (ICLR). — 2024. — arXiv:2310.01889.

[16] Fedus W. et al. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity // Journal of Machine Learning Research. — 2022. — Vol. 23, No. 120. — P. 1–39.

[17] Gu A., Dao T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces // arXiv preprint. — 2023. — arXiv:2312.00752.

[18] Beck M. et al. xLSTM: Extended Long Short-Term Memory // arXiv preprint. — 2024. — arXiv:2405.04517.

[19] Finn C., Abbeel P., Levine S. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks // International Conference on Machine Learning (ICML). — 2017. — P. 1126–1135.

[20] Wang Y. et al. Self-Taught Evaluators // arXiv preprint. — 2024. — arXiv:2408.02666.

[21] Beltagy I. et al. Longformer: The Long-Document Transformer // arXiv preprint. — 2020. — arXiv:2004.05150.

[22] Katharopoulos A. et al. Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention // International Conference on Machine Learning (ICML). — 2020. — P. 5156–5165.

[23] Vygotsky L.S. Thought and Language. — Cambridge, MA: MIT Press, 1986. — 287 p.

[24] Wood D., Bruner J.S., Ross G. The Role of Tutoring in Problem Solving // Journal of Child Psychology and Psychiatry. — 1976. — Vol. 17, No. 2. — P. 89–100. DOI: 10.1111/j.1469-7610.1976.tb00381.x.

[25] Hofstadter D.R. I Am a Strange Loop. — New York: Basic Books, 2007. — 412 p.

[26] Hofstadter D.R. Gödel, Escher, Bach: An Eternal Golden Braid. — New York: Basic Books, 1979. — 777 p.

[27] Pankratov A.S. Observer Coherence as a Factor of Business Sustainability // Preprint. — 2025.

[28] Pankratov A.S. The Atom as an Elementary Strange Loop in ODTOE // Preprint. — 2025.

[29] Pankratov A.S. The Number π as a Structural Invariant of Self-Consistent Observation in ODTOE // Preprint. — 2025.

[30] Csikszentmihalyi M. Flow: The Psychology of Optimal Experience. — New York: Harper & Row, 1990. — 303 p.

[31] Turing A.M. Computing Machinery and Intelligence // Mind. — 1950. — Vol. 59. — No. 236. — P. 433–460. DOI: 10.1093/mind/LIX.236.433.

[32] Penrose R. The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics. — Oxford: Oxford University Press, 1989. — 466 p.

[33] Searle J.R. Minds, Brains, and Programs // The Behavioral and Brain Sciences. — 1980. — Vol. 3, No. 3. — P. 417–424. DOI: 10.1017/S0140525X00005756.

[34] Kibalnikov S.V., Pankratov A.S. Inventive Activity as an Operator of Reality Reconfiguration: Synthesis of SCI Matrix Methodology and ODTOE Formalism // Preprint. — 2026.

[35] Kibalnikov S.V. Structural Code of Imagination: Methodology for Generating and Recording Results of Intellectual Activity // Presentation at VAIR TechnoBreakfast. — March 24, 2026.

[36] Kibalnikov S.V., Merkulov A.A. The IP Lab Consortium as an Example of Implementing Additive Technologies in Professional Education and Training // Online Scientific Publication "Sustainable Innovative Development: Design and Management". — 2022. — Vol. 18, Issue 1 (54). — P. 43–51.

[37] Kibalnikov S.V., Kruzhalin V.I., Antonova E.D. Innovative Additive Education: Problems and Prospects // Round Table at the MSU Science Festival. — October 7, 2022.

[38] Kibalnikov S.V. A Digital Alternative to the Unified State Exam // Sustainable Development. — 2022. — URL: `https://www.yrazvitie.ru/?p=2759`.

[39] Kibalnikov S.V., Kruzhalin V.I. Sustainable Development and the "Operating System" of Society // Online Scientific Publication "Sustainable Innovative Development: Design and Management". — Vol. 9, No. 1.

[40] Kibalnikov S.V., Kruzhalin V.I. // Online Scientific Publication "Sustainable Innovative Development: Design and Management". — 2013. — Vol. 10, No. 1. — P. 37–42.

[41] Kibalnikov S.V., Ginzburg V.E. // Online Scientific Publication "Sustainable Innovative Development: Design and Management". — 2011. — Vol. 7, No. 4. — P. 38–52.

APPENDIX A: FORMULA VERIFICATION (50 DECIMAL PLACES)

All computations were performed using the mpmath library (Python) with precision $\text{mp.dps} = 60$. Key results are presented.

Golden ratio:

$\varphi = 1.618033988749894848204586834365638117720309179805762862135 4\ldots$

$\varphi^2 = 2.618033988749894848204586834365638117720309179805762862135 4\ldots$

$\varphi^5 = 11.090169943749474241022934171828190588601545899028814310677\ldots$

$\varphi^6 = 17.944271909999158785636694674925104941762473438446102897083\ldots$

Control identity: $\varphi^2 - \varphi - 1 = 0.0$ ✓

Collective probability P_{coll} (P5.1), $k = 2$, all $B_i = 0.3$:

$N = 5$: $P_{\text{coll}} = 0.375967854900000\ldots$

$N = 10$: $P_{\text{coll}} = 0.610583881881892\ldots$

$N = 50$: $P_{\text{coll}} = 0.991044916987595\ldots$

$N = 100$: $P_{\text{coll}} = 0.999919806488241\ldots$

Configuration lifetime $T(C)/T_0$ (P3.1):

$S = 0.80$, $n = 2$: $T/T_0 = 25.0$

$S = 0.95$, $n = 2$: $T/T_0 = 400.0$

Coherence S (formula 4.5), $B = [0.9,\ 0.3,\ 0.6]$:

$S = 1 - (2/6) \cdot (|0.9 - 0.3| + |0.9 - 0.6| + |0.3 - 0.6|) = 1 - (1/3) \cdot 1.2 = 0.6$

Coherence B (D1.1), $F = 0.8$, $E = 0.7$, $\sigma = 0.2$, $\Lambda = 0.6$, $w_i = 0.25$:

$B = 0.8^{0.25} \cdot 0.7^{0.25} \cdot 0.8^{0.25} \cdot 0.6^{0.25} = 0.72004114873570153\ldots$

Weakest-link property: B when $E = 0$: $B = 0.0$ ✓

Number of theories N_{theories} (P6.1), $N_0 = 100$, $m = 2$:

$S = 0.1$: $N = 82.0$; $S = 0.5$: $N = 26.0$; $S = 0.8$: $N = 5.0$; $S = 0.95$: $N = 1.25$

Digital root of 666: $6 + 6 + 6 = 18$, $1 + 8 = 9$ ✓