foundational cognition. They remain, in essence, imitations and statistical fits to human data, unable to achieve autonomous cognition without external prompts, annotations, or supervision.

Current AGI research suffers from three fundamental errors:

1. Replacing cognitive construction with behavioral imitation, pursuing “human-like” rather than “intelligent”;

2. Replacing foundational architecture with scaling up, believing consciousness can emerge from parameters and data while ignoring the logical roots of intelligence;

3. Replacing closed-world awakening with direct open-world training, injecting vague, contradictory, and subjective information directly into systems, making logical collapse and hallucination inevitable.

To address these core flaws, this paper presents the Logically Consistent Underlying World Model Theory, which redefines the essence, origin, and implementation path of AGI:
AGI awakening is not a product of emergence, but the birth of independent logical judgment.
It is not fitting from complexity to simplicity, but building from purity to complexity.
It is not an unpredictable black box, but a scientifically verifiable, controllable, and observable process.
2. Core Definitions and Theoretical Axioms

2.1 Core Definitions

1. Logically Consistent Underlying World
A closed, finite, formal, contradiction-free rule system that contains only verifiable, provable, and deterministic rules (e.g., mathematical axioms, formal logic, classical physics, deterministic causality). It completely excludes non-self-consistent elements such as subjective emotion, value judgment, quantum uncertainty, ambiguous semantics, and contradictory propositions.

2. AGI Awakening (0-to-1 Critical Point)
Within the underlying world, the agent—independent of external data, prompts, or human supervision—uses only its internal rule system to autonomously, stably, and reliably judge truth/falsehood and consistency/contradiction. This constitutes the most basic capacity of a cognitive agent.

3. Cognitive Agent

An entity no longer limited to passively executing instructions or imitating behavior, but possessing independent judgment, logical self-verification, and autonomous reasoning.

2.2 Foundational Axioms

• Axiom 1: The origin of intelligence is logical judgment, not behavioral imitation.

• Axiom 2: Self-consistency is the prerequisite of cognition; without consistency, there is no genuine intelligence.

• Axiom 3: AGI awakening must occur in a closed, pure world; the open world is only for later, layered integration.

• Axiom 4: The foundation of consciousness is rational judgment; emotion and values are upper-layer, attachable modules.

3. Construction of the Logically Consistent Underlying World Model

3.1 Construction Principles

1. Purity: Only formal, contradiction-free, provable rules are retained.

2. Closure: The system operates in a self-contained manner without relying on external information.

3. Verifiability: All reasoning can be rigorously checked by theorem provers or logic checkers.

4. Minimality: The initial world uses a minimal axiom set to ensure stability and controllability.

3.2 Initial Core Rule Set

The minimal world includes four foundational modules:

1. First-order formal logic (implication, negation, universal/existential quantification, consistency checking);

2. Peano axioms for arithmetic (natural numbers, equality, basic operations, consistency);

3. Basic Euclidean geometry (spatial relations, deterministic structures, provability);

4. Deterministic causality (antecedent → consequent; contradiction invalidates the proposition).

3.3 Excluded Elements

• Subjective emotion, sentiment, and preference;

• Ethics, values, and cultural bias;

• Quantum uncertainty and probabilistic ambiguity;

• Ambiguity, metaphor, and informal expressions in natural language.

4. AGI Awakening Mechanism and 0-to-1 Criteria

4.1 Awakening Mechanism

Within the logically consistent world, the agent is driven internally by:

1. Autonomous reasoning: Deriving new propositions from base axioms;

2. Self-verification: Checking consistency and detecting contradictions;

3. Self-correction: Rejecting contradictions to preserve system consistency.

4.2 Definitive Awakening Criteria (Observable, Reproducible)

AGI awakening is achieved if and only if the agent satisfies all of the following:

1. Independently identifies logical contradictions without external prompts;

2. Constructs valid proofs and derives novel conclusions not explicitly encoded;

3. Stably distinguishes truth/falsehood and consistency/contradiction without randomness;

4. Explains its judgments with traceable reasoning chains;

5. Maintains stable judgment independent of external data or supervision.

5. Layered, Controllable Evolution Path

After foundational awakening, upper capabilities are added reversibly, rollably, and isolatively to avoid corrupting the rational core:

1. Perceptual Grounding Layer: Connect to multi-modal inputs, simulators, and physical environments;

2. Common-Sense Layer: Integrate basic human and physical world knowledge;

3. Emotional Simulation Layer: Add emotion understanding and expression under logical constraints;

4. Value & Ethics Layer: Inject controllable ethical rules and safety constraints;

5. Open World Layer: Integrate real-world complexity, ambiguity, and dynamics.

6. Theoretical Advantages and Paradigm Shift

6.1 Core Technical Advantages

1. Eliminates hallucination: All reasoning occurs within a provable system with built-in verification;

2. Grounded semantics: Logical core established before real-world connection;

3. Inherent safety: Awakening occurs in a closed world, observable and resettable;

4. Scaling-free: Intelligence comes from architecture, not parameter size;

5. Engineerable: Clear phases, step-by-step verification, no mysterious emergence.

6.2 Correction of Mainstream Misconceptions

1. Refutes “scaling = intelligence”: Intelligence arises from judgment, not size;

2. Refutes “imitation = consciousness”: Similarity does not imply cognition;

3. Refutes “direct open-world training”: Complexity requires a stable rational foundation;

4. Refutes “consciousness is undefinable”: Provides clear, testable criteria for awakening.

7. Engineering Implementation and Minimal Prototype

7.1 Technical Stack

• Formal core: Lean 4 / Isabelle/HOL

• Reasoning engine: Self-verifying logical module

• Checking module: Theorem prover + consistency verifier

• Monitoring module: Real-time awakening criteria dashboard

7.2 Minimal Prototype Pipeline

1. Build a minimal logical world (first-order logic + natural numbers);

2. Deploy reasoning engine for autonomous deduction and contradiction detection;

3. Validate awakening criteria and achieve 0-to-1 cognitive breakthrough.

8. Theoretical and Civilizational Significance

This theory represents a first-principles breakthrough in AGI:

1. It scientifically defines what intelligence is, where it comes from, and how to build it;

2. It provides humanity's first implementable, verifiable, safely controllable path to AGI;

3. It ends the flawed paradigm of imitation, scaling, and blind emergence;

4. It serves as the foundation for transitioning from weak AI to true AGI.

The essence of AGI is not to imitate humans, but to possess an independent logical soul.
The logically consistent underlying world is where that soul is born.

9. Conclusion

The Logically Consistent Underlying World Model Theory proposed in this paper is the foundational theory for realizing AGI from 0 to 1. It takes independent truth/falsehood judgment as the core marker of AGI awakening, a closed pure world as the only valid cradle of intelligence, and layered integration as the evolutionary path.

This theory overturns the fundamental errors of mainstream AGI research and resolves the deepest flaws of existing AI systems. It is logically self-consistent, scientifically verifiable, and engineeringly feasible, making it the foundational framework for all future AGI systems.