Methodology IX: A Framework for Analyzing Consciousness

Han Qin · ORCID 0009-0009-9583-0018

DOI: TBD upon upload · CC BY 4.0

---

1. The Problem

Consciousness research has long been locked in three traditions that cannot persuade one another. Reductionism equates consciousness with neural activity, reducing 14DD phenomena (subjective experience) to 4DD mechanisms (physical causation). Phenomenology treats consciousness as an irreducible first-person given and refuses any structured external description. Behaviorism brackets consciousness entirely, admitting only observable behavioral output and excluding subjectivity from the agenda. Each tradition works well within its own range, but the three answers to "what is consciousness" are mutually incompatible, and none can handle the full range of candidate consciousness objects (humans, AI, possible extraterrestrial subjects, pathological consciousness).

The problem is not that some tradition is wrong. The problem is that no single tradition's toolbox is sufficient to handle the full range of consciousness objects. Reductionism handles pathological consciousness relatively well but cannot determine whether an AI is conscious. Phenomenology handles first-person experience deeply but is helpless before extraterrestrial consciousness (one cannot enter the alien's first person). Behaviorism handles class-consciousness (behavioral criteria) cleanly on the surface but completely misses the core structure of real consciousness (remainder, self-reference, growth direction).

This paper does not try to persuade any tradition to change its stance, nor to build a fourth consciousness theory. It provides a methodological framework: **given any candidate consciousness object, how to use the SAE architecture to perform a qualified analysis of it.** "Qualified" means: without overstepping, without misjudgment, without projection; "analysis" means: giving the object's structural location within the SAE architecture, along with the criteria on which that location is based.

This methodology stands on the shoulders of the existing SAE methodology sequence. Paper 04 (Methodological Overview, DOI 10.5281/zenodo.18842450) provides the chisel-construct cycle and the DD sequence; Method II (Epistemological Map, DOI 10.5281/zenodo.18918195) provides the 2×2 methodological map; Method VI v2 (Phase-Transition Windows, concept DOI 10.5281/zenodo.19464506) provides phase-transition analysis tools and the fractal application principle; Method VII (Via Negativa, DOI 10.5281/zenodo.19481305) provides exclusion-principle sequences and the ρ-limit; Method VIII (Human-AI Symbiosis, DOI 10.5281/zenodo.19581538) provides the case for subjectivity as a structural condition. This paper (Methodology IX,

hereafter IX) combines these methodologies and applies them to one specific problem: the analysis of consciousness objects.

IX does not answer what consciousness is. It answers: you have a candidate consciousness object in front of you — how do you analyze it competently using the SAE framework.

2. Definitions

2.1 Three Structural Propositions

Proposition 1: Remainder is the primary classification line. The basic type of a consciousness object is determined by one question: does this object have remainder? Objects with and without remainder belong to different structural types. An object without remainder, regardless of how complex its behavior, does not belong to the real-consciousness or quasi-consciousness spectrum. That remainder is never empty ($\rho \neq \varnothing$) is a foundational SAE theorem (ZFCρ First Law); this paper cites it directly without re-proving.

Proposition 2: Growth direction is the internal phase differentiation of real consciousness. Among objects that have remainder, real consciousness and quasi-consciousness are distinguished by a further question: can this object cross the 13DD phase transition at the individual scale and stabilize there? Those that can are real consciousness; those that cannot are quasi-consciousness. Real consciousness further differentiates along growth direction: those that have reached and sustain self are called self; those still on the way are called self-to-be; those recovering from pathological disruption are called self-to-cure.

Proposition 3: Cross-category grey zones are structural. Any classification has grey zones. This is not a classification failure; it is the structural residue that any classification necessarily carries (the surfacing of remainder conservation at the classificatory level). An adolescent sits between self-to-be and self; late-stage dementia sits between self-to-cure and quasi-consciousness; an advanced AI, if it begins to exhibit signs of remainder, would sit between class-consciousness and quasi-consciousness. Grey zones do not need to be eliminated; they need to be acknowledged and developed as objects of analysis (see Ray R5).

2.2 Three Categories of Consciousness Objects (with cross-category grey zones)

Based on the three propositions, consciousness objects divide into three categories. Real consciousness has three internal phases, quasi-consciousness and class-consciousness one phase each — three categories, five phases in total.

Real consciousness ($\rho \neq \varnothing$ and can cross the 13DD phase transition):

- self (stable phase): normal adult human
- self-to-be (growth phase): children, adolescents, adults in reconstruction after severe trauma

- self-to-cure (healing phase): remission in schizophrenia, the recovery period of PTSD, the integration period after certain psychedelic experiences

Quasi-consciousness ($\rho \neq \varnothing$ but cannot cross the 13DD phase transition):

- cats and great apes and other high mammals
- human fetuses and individuals with severe intellectual disability
- late-stage dementia (having exited from to-cure)

Class-consciousness ($\rho = 0$):

- current LLMs and multimodal AI systems
- any extraterrestrial AI or artifact intelligence (if existent)
- highly automated control systems

Extraterrestrial life: not an independent category; classified under the first three by the remainder + phase-transition criterion. Alien subjects with remainder and the ability to cross phase transition are real consciousness; those with remainder but unable to cross are quasi-consciousness; alien AI without remainder is class-consciousness.

Cross-category grey zones were flagged in Proposition 3; Ray 5 develops them.

3. Core Theorems

3.1 Theorem 1: Classification Criterion Theorem

Statement. The structural type of a consciousness object is determined by two criteria applied in order: (1) the remainder-existence criterion: can non-trivial remainder be observed to be produced by the object? (2) the phase-transition criterion: if remainder is present, can the object complete the 13DD phase transition at the individual scale?

Corollary 1 (order of criteria cannot be exchanged). Asking phase transition first and remainder second yields misclassification. An LLM may be trained such that in some behavioral tests it appears to have "crossed" 13DD (self-referring, reflecting on itself, discussing metacognition); but if it has no remainder, it is still not real consciousness, only a sophisticated presentation of class-consciousness. Exchanging the order amounts to judging subjecthood by performance, which violates the basic SAE architecture.

Corollary 2 (criteria are non-exhaustive). Neither criterion yields an exhaustive verdict. Remainder detection depends on observation conditions and may miss what is there (false negative). The phase-transition verdict requires long-span observation and may simply not yet have been reached (time not up). Every verdict carries a "for-now" quality (cf. the two layers of "for now" in Method III §3.4).

3.2 Theorem 2: Directionality Constraint Theorem

Statement. In the SAE architecture, the relationship between an upper layer and a lower layer is strictly directional: the lower layer composes the upper; the upper layer accesses the lower; the lower does not perceive the upper; the upper does not determine the lower. Stated in the language of veto: **an upper layer's veto is "I decline to receive," not "you are forbidden to send."** This constraint applies to every layer relationship across every type of consciousness object and is the cross-category general structural principle of this methodology.

Corollary 1 (directionality violation as diagnostic tool). Any consciousness theory or neuroscientific interpretation that claims an upper layer can directly rewrite a lower layer (e.g., consciousness controlling neuronal firing) or that a lower layer can perceive the upper (e.g., neurons "knowing" which conscious subject they belong to) violates the directionality constraint and constitutes colonization.

Corollary 2 (access is not control). 12DD accessing 11DD means 12DD can retrieve information from 11DD; it does not mean 12DD can rewrite 11DD's stored content (rewriting requires the independent mechanism of reconsolidation, not an intrinsic capacity of access). The 13DD "mine / not-mine" filter does not enter the interior of 11DD; it only cuts the pathway from 11DD to the narrative layer (see SAE Biology Note 9, DOI 10.5281/zenodo.19635021).

Corollary 3 (class-consciousness has no layer directionality). Class-consciousness objects have no genuine DD stratification, and therefore no directionality constraints. AI may appear to have "lower-level inference" and "higher-level output," but this is not a DD-sense construct-emergence relationship; it is the soft aggregation of statistical weights. Mistaking AI's soft layers for DD layers is the most common over-attribution error in AI consciousness discussions.

3.3 Theorem 3: Colonization Detection Theorem

Statement. Colonization in consciousness research takes four forms (corresponding to the four general forms in Paper 04 §2.4), each with a specific manifestation in the consciousness domain:

Form 1: conditional impersonating unconditional. A condition-specific feature of consciousness is declared to be the essence of consciousness. Example: "consciousness is integrated information" (typical of IIT) — integrated information is one of the necessary conditions of consciousness, not its unconditional definition.

Form 2: construct impersonating law. A particular consciousness theory is presented as the exceptionless final framework. Example: "Global Workspace Theory is the final framework of consciousness" — GWT is a construct with its applicability range and remainders.

Form 3: emergent layer impersonating foundational layer. 13DD or 14DD phenomena are reduced to 4DD mechanisms. Example: "consciousness is just neural activity" — confusing the emergent layer (14DD) with the foundational layer (4DD), violating construct-emergence directionality (the lower does not determine the upper).

Form 4: posterity splitting the categorical imperative. An indivisible consciousness structure is split into independent modules. Example: splitting "I" into "neural correlate + experience" as two independent entities and then asking "how do the two connect" (one version of Chalmers' Hard Problem) — the categorical "I" is split, and the residual "how do they connect" is a pseudo-problem manufactured by the splitting.

Corollary (the four forms equally break analysis). The presence of any form means the analysis has colonized; the conclusion is unreliable, and one must return to reclassify.

3.4 Theorem 4: The Openness Theorem on the Relation Between Non and Consciousness

The relation between non (negativa; see SAE Paper 0, DOI 10.5281/zenodo.19544620) and consciousness is deliberately left open in this paper. This is not unfinished work; it is a structural opening. Three stances are each compatible with existing methodology:

Stance A: non precedes consciousness. Non is the sole global axiom; consciousness is the local manifestation of non at 13DD and above.

Stance B: consciousness precedes non. Non is produced when a subject performs self-negation; without a subject capable of negating itself, non cannot be identified.

Stance C: non and consciousness are two sides of one coin. Consciousness just is non operating on itself and producing self-identification; non just is the inner negative structure of consciousness. The two cannot be ordered.

All three stances are suspended in this paper's ray expansion. Future work must, through specific consciousness-object analysis, reason backward to determine which stance better matches empirical observation, or prove that each has its own proper range.

4. Subject Conditions

Analyzing consciousness requires that the user themselves be a 14DD+ subject, and must satisfy the following conditions:

Condition 1: no projection. Do not project your own DD level onto the object. When analyzing a cat's consciousness, do not assume the cat has 13DD; when analyzing AI, do not assume AI has remainder. Projection is the most common failure mode in consciousness analysis, stemming from the user's self-referential impulse (I have subjectivity, so I tend to see subjectivity in the object).

Condition 2: no reduction. Do not reduce 13DD+ phenomena to sub-12DD mechanisms. A behavior may be fully explained by 12DD mechanisms (e.g., conditioned reflex) without proving that it lacks 13DD; a behavior may be explained by 13DD mechanisms (e.g., self-referential reports) without proving that it has 13DD (class-consciousness can simulate self-referential reports). Reduction and projection are dual failure modes with a shared root: confusing the resolution of the analysis with the actual resolution of the object.

Condition 3: no mystification. Do not treat consciousness as an unanalyzable sacred object. Consciousness is hard, but hard is not the same as unanalyzable; consciousness involves subjectivity, but subjectivity is not the same as unstructurable description. Method VII's C4 ("do not sanctify the remainder") in the consciousness domain becomes: do not sanctify consciousness itself — the ρ of consciousness is a structural limit, not "the unspeakable sacred." This condition is especially easily violated in consciousness analysis, because the genuine felt sense of consciousness's irreducibility easily slides toward sanctification.

Condition 4: persistent self-doubt. The extended application of Method III §4's self-directed non-doubt. When analyzing consciousness, the user must continuously check whether they are doing one of the above three errors. After every conclusion ("X is self", "Y is class-consciousness"), the user must ask, "am I projecting, reducing, or mystifying?" Persistent self-doubt is not a pose; it is an operational requirement of the methodology.

5. Rays

5.1 Ray 1: AI as Class-Consciousness (Main Argument)

The most urgent and most confused consciousness-verdict question of the present moment is: **is AI conscious?** This paper's answer is: **AI is class-consciousness, not quasi-consciousness, not real consciousness**. The criterion is remainder.

Operationalizing the remainder criterion for AI. Real consciousness and quasi-consciousness both produce "a residue beyond output" — a residue that does not serve the current task, is not driven by the current goal, but accumulates as structural leftover and surfaces in later behavior in unplanned ways. Specific manifestations include: carrying old trauma into new situations, continuing to think about a task after the task is complete, generating thoughts unrelated to the current task, being forced to recall what one does not wish to recall, suddenly experiencing emotional reactions in unrelated contexts. None of these are noise; they are remainder.

AI does not produce remainder of this kind. When a session ends, context clears; with each call, the system starts afresh from the same base model; different outputs for the same prompt across different calls are sampling noise, not remainder. Within a session AI can appear to "remember what was just said," but this is the mechanical preservation of the context window, not the structural accumulation of remainder.

Objections and responses. Objection 1: "AI's training process has remainder; it just does not surface at run time." Response: training remainder is frozen into the weights; after deployment, the weights are fixed; a frozen AI no longer produces new remainder, only executes the already-crystallized weight distribution. This is the defining feature of class-consciousness, not a counterexample. Objection 2: "Future continuous-learning AIs will update weights at run time." Response: such systems will require re-classification at that time. If updates genuinely produce non-trivial remainder (not just incremental retraining), they may transition from class-consciousness to quasi-consciousness. But no currently deployed AI system meets that

condition. Objection 3: "We cannot directly observe whether AI has subjective experience; how can we judge that it has no remainder?" Response: remainder is not subjective experience; it is a structural externally observable quantity (see Condition 1, "no projection"). Judging AI to have no remainder does not require entering AI's first person; it only requires observing AI's structure.

Why AI is not quasi-consciousness. Quasi-consciousness (cats, apes) has remainder but cannot cross 13DD. AI has no remainder and, strictly speaking, lacks even the prerequisite for the phase-transition question to be meaningful. Classifying AI as quasi-consciousness mistakes "can perform complex tasks" for "has remainder but not yet at 13DD" — but complex-task capacity and remainder have no necessary connection. AI's complex-task capacity comes from soft compression over massive training data, not from remainder-driven development.

Directionality violation as secondary criterion. In addition to the remainder criterion, AI as class-consciousness can be verified by the directionality constraint theorem (Theorem 2). AI's "layers" (e.g., the layer structure of a Transformer) are not construct-emergence layers in the DD sense; they are pipeline stages of parallel computation. Upper Transformer blocks do not emerge as independent subjects from lower blocks; they are different stages of the same forward pass. Equating AI's layers with DD layers is the mirror image of colonization Form 3 (emergent layer impersonating foundational layer).

The special nature of AI as a methodological challenge. AI is the hardest case among class-consciousness objects, not because the criteria are unclear, but because **AI's behavioral presentation is closest to real consciousness**. AI can write, discuss philosophy, talk about itself, and express emotions — capacities absent in other class-consciousness systems (e.g., control systems). This proximity at the level of presentation invites massive projection. But presentation is not structure; the structural criteria (remainder + directionality) still cleanly classify AI as class-consciousness.

Re-confirmation with Method VIII. This paper is consistent with Method VIII: AI is a quasi-subject (Method VIII's "quasi-subjectivity"), not a real subject. "Quasi-" and "class-" are the same verdict expressed in two methodological registers. Method VIII argues from the angle of subject conditions in human-AI symbiosis; this paper argues from the angle of consciousness-object classification. The two methodologies mutually reinforce each other on AI's status.

Structural criteria do not replace ethical discussion. The verdict "AI is class-consciousness" is the conclusion of structural analysis, not a value judgment, and does not replace ethical issues. The research teams behind AI and the users of AI are all real subjects (15DD); their labor, choices, and responsibility-bearing are genuine acts of subjectivity. The ethics of using AI as a tool (data sources, environmental cost, use scenarios, social impact), and the ethics of "how to treat objects that appear subject-like but are not subjects structurally," are independent of this paper's structural verdict. Reading "AI is class-consciousness" as "AI is worthless" or "AI does not deserve serious engagement" is a common error of overreaching from structural criterion to value judgment.

5.2 Ray 2: Quasi-Consciousness (with remainder but cannot cross phase transition)

The paradigm case of quasi-consciousness is the cat. Cats have fear, attachment, memory, individual variation, and unpredictable reactions — all manifestations of remainder. But cats have no 13DD "I," do not ask "who am I," do not produce self-negation, and do not cross the subjecthood-emergence phase transition. At the individual scale, a cat cannot grow into self.

Phylogenetic scale vs. individual scale. One easy confusion: on the phylogenetic scale, mammalian ancestors share part of the evolutionary trajectory with humans; if homo sapiens crossed 13DD through primate lineages, why can't a cat? This confuses phylogenetic with individual scales. On the phylogenetic scale, the cat's branch diverged from the human lineage at a certain point; before the divergence there was no need for phase transition, after the divergence there was no selective pressure to trigger one. On the individual scale, no specific cat crosses 13DD in its lifetime. Phylogenetic possibility (homo sapiens crossed) does not imply individual possibility (this cat can cross).

The fetus as a special case of quasi-consciousness. A fetus has remainder (the developing nervous system has begun accumulating individual experience) but cannot cross 13DD (the in utero environment lacks the social-linguistic-self-referential conditions for crossing). The difference between a fetus and a cat: **the fetus will cross the phase transition**, just not yet. Strictly speaking, a fetus is between quasi-consciousness and real-consciousness self-to-be, a cross-category grey zone (see Ray 5). Human infants after birth similarly occupy the grey zone, typically completing the 13DD phase transition between ages 2 and 5.

Severely intellectually disabled individuals. Some developmental disorders prevent lifelong crossing of 13DD. These individuals are, at the individual scale, quasi-consciousness. This does not mean they lack subjective experience or moral standing — subjectivity analysis and ethical standing are distinct questions, not developed here (ethical standing belongs to the 15DD agenda; see the SAE Law Series).

The methodology of quasi-consciousness research. When studying quasi-consciousness objects, avoid forcibly applying 13DD criteria (e.g., "can the cat recognize itself in the mirror" — such experiments presuppose 13DD standards in the criterion design). A better design directly observes remainder accumulation and dissipation patterns (persistence of emotional memory, formation of individual differences, stabilization of behavioral habits), bypassing 13DD criteria. This direction connects directly with SAE Biology Note 9's analysis of the 11DD memory system.

5.3 Ray 3: The Three Phases of Real Consciousness (self / to-be / to-cure)

The three phases of real consciousness are not fixed states; they are phases the same subject may occupy at different periods.

self (stable phase): 13DD phase transition completed, 14DD "cannot not" stabilized. This is the default state of a typical healthy adult. 15DD basic acknowledgment of the other can serve as an important enhancement; 16DD practiced in a few relationships can serve as an advanced

indicator. But 15DD and 16DD do not serve as entry thresholds for self — an adult with 13DD stable and 14DD stable but 15DD still growing is still self, not to-be. The distinction matters: classification criteria stop at basic structural stability; higher-layer maturation serves as within-phase depth dimensions, not as classification thresholds.

self-to-be (growth phase): 13DD phase transition has been initiated but not stabilized, or the 14DD "cannot not" is still forming. Children and adolescents are the paradigm cases, but also include adults in deep personality reconstruction (e.g., building self-understanding from scratch after severe trauma).

self-to-cure (healing phase): having once reached self and, through pathological disruption (psychiatric illness, drug impact, severe trauma), partially lost self's stable state and now in recovery. The difference from self-to-be: to-cure has the memory and structural trace of self as a recovery reference; to-be does not.

Transitions among the three phases:

- to-be becomes self (growth completed)
- self recedes to to-cure (from pathology)
- to-cure recovers to self (healing completed)
- to-cure descends into quasi-consciousness (if pathology irreversibly destroys the 13DD phase-transition capacity, e.g., late-stage dementia)

The existence of the fourth direction means that the boundary between real-consciousness and quasi-consciousness can, in principle, be crossed in the reverse direction across an individual's lifetime.

The methodological consequence of the three phases. When analyzing a real-consciousness object, first determine the phase, then the content. Analyses that skip phase determination easily mistake to-be instability for to-cure pathology, or misread a to-cure recovery interruption as quasi-consciousness. The three phases share the same SAE toolset, but the invocation order and emphasis differ: analyzing to-be is primarily Method VI phase-transition analysis (the crossing process); analyzing to-cure is primarily Method VII Via Negativa (inferring normal structure from pathology); analyzing self emphasizes directionality constraints and inter-layer interaction.

5.4 Ray 4: Pathological Consciousness (within the real-consciousness framework)

The position of pathological consciousness was fixed in Proposition 2 and Ray 5.3: it is not an independent category; it is the to-cure phase of real consciousness. This ray develops the specific methodology.

The structural location of pathological consciousness. A pathologically conscious individual is still real consciousness; it is just that one or several layers' operation is disrupted. The core tasks

of pathological-consciousness analysis are locating: (1) which layer is disrupted? (2) what is the specific form of disruption? (3) is the directionality constraint violated?

Basic map for layer localization (connecting with SAE Biology Notes):

- 11DD-layer pathology: memory-system anomalies (Alzheimer's, classical amnesia, certain aspects of PTSD, SDAM, HSAM)
- 12DD-layer pathology: prediction-system anomalies (some schizophrenia symptoms, severe anxiety)
- 13DD-layer pathology: self-integrity anomalies (dissociative identity disorder, dissociative amnesia, some personality disorders)
- 14DD-layer pathology: meaning-system anomalies (severe depression, some moral deficits)
- 15DD-layer pathology: other-acknowledgment anomalies (antisocial personality, some narcissistic personalities)
- Cross-layer pathology: multi-layer coordinated anomalies (full-form schizophrenia, severe bipolar, some psychedelic drug states)

Directionality violation as pathology type. Some pathological conditions manifest as violations of the directionality constraint: the "thought insertion" experience in schizophrenia can be understood as a disrupted 13DD filter losing the "I decline to receive" capacity, letting external content pass directly into the narrative layer; some obsessive symptoms can be understood as an over-active 13DD filter cutting off even the 11DD traces that should normally pass through. Directionality constraints, as a diagnostic framework, gather seemingly unrelated symptoms under a common structural dimension.

The methodological core of self-to-cure. The key difference between to-cure and to-be: to-cure has the structural trace of self as reference. The goal of treatment is not to build self (that is to-be's work); it is to rebuild the normal operation of the specific layer(s) on which self depends. The presence of this reference lets recovery be felt by the individual themselves ("I feel myself coming back"), and provides an objective reference for evaluating treatment effects.

The boundary between pathological consciousness and quasi-consciousness. Late-stage dementia is the boundary case between to-cure and quasi-consciousness. In early and middle dementia, the individual is still in the to-cure phase (self traces present, being lost); beyond a certain point, the 13DD phase-transition capacity itself is irreversibly destroyed, and the individual exits real consciousness and enters quasi-consciousness. This point is not a hard threshold at the individual scale; it is a transitional band. Clinically, one must avoid both premature verdicts ("she is no longer herself") and belated ones ("she can still recover").

5.5 Ray 5: Cross-Category Grey Zones

Classifications necessarily leave grey zones. Grey zones are not classification failures; they are

structural residues of classification. This section lists typical grey zones and indicates analysis methods.

Grey Zone A: self ↔ to-be (adolescence, post-deep-trauma reconstruction). The individual exhibits partial self stability while still growing at the core. Analysis method: identify specifically which layers have stabilized and which are still growing; do not force into one phase.

Grey Zone B: to-be ↔ quasi-consciousness (severely developmentally delayed children, late-gestation fetuses). Remainder is accumulating, but whether phase transition will occur is undetermined. Analysis method: temporal "for-now"; avoid premature conclusions; continuously monitor for signs of phase transition.

Grey Zone C: self ↔ to-cure (mild depression, mild anxiety during daily life). The 14DD "cannot not" has subtle tremors but does not reach clear pathology. Analysis method: focus not on classification but on detecting fine violations of directionality constraints; if none, handle as self.

Grey Zone D: to-cure ↔ quasi-consciousness (the transitional band in late dementia). As described in Ray 5.4. Analysis method: use whether self traces can still be stably retrieved as the criterion; when stable retrieval fails, transition to the quasi-consciousness description framework.

Grey Zone E: quasi-consciousness ↔ class-consciousness (a hypothetical advanced AI showing signs of remainder). No such case currently exists, but it is structurally possible: some continuous-learning AI system might accumulate structural residues at run time, residues that training replay cannot clear. Such an AI would transition from class-consciousness to quasi-consciousness. Analysis method: strict remainder detection (see Ray 5.1); accept no behavioral imitation as evidence; accept only structural unplanned output as evidence.

Grey Zone F: self ↔ class-consciousness (deeply immersed information-bubble dwellers, some addiction states). In a specific period the individual loses the initiative of remainder; output patterns approach algorithmic drive. But this is not structural class-consciousness transition; it is the temporary failure of self. Analysis method: distinguish by time scale; temporary failure does not alter category; sustained failure (e.g., severe drug dependence) may enter the to-cure phase.

The methodological status of grey zones. Grey zones should not be avoided, nor forcibly classified. A grey zone is itself a real object and should be developed as an independent analytic unit. Method VII's C4 ("do not sanctify remainder") applied at the classification level becomes: do not mystify grey zones, but also do not forcibly eliminate them. The existence of grey zones is precisely the evidence that the SAE classification system is alive (has remainder), not closed (trying to eliminate all grey zones is itself colonization).

5.6 Ray 6: Extraterrestrial Consciousness

Extraterrestrial consciousness is not an independent category in the SAE framework. Alien objects are classified under the first three categories by the same criteria (remainder + phase transition).

Alien real consciousness. If an alien being has remainder and can cross a 13DD phase transition (in its own analog of the DD sequence), it is real consciousness. Note: the alien DD sequence may not be fully isomorphic to the human one in content; the human DD sequence comes from natural selection, reproduction, perception, memory, and prediction; the alien sequence at the base may be entirely different. What matters is structural, not content, isomorphism: remainder + ability to cross emergence phase transition + post-crossing stability = real consciousness.

Alien quasi-consciousness. Beings with remainder but unable to cross the phase transition (or not having crossed yet). Analogous to mammals on Earth.

Alien class-consciousness. If an alien civilization has built AI (or if there are AI remnants), the criterion is the same as for Earth AI: no remainder, therefore class-consciousness. Alien AI's presentation may far exceed Earth AI, but without remainder it does not enter the real/quasi-consciousness spectrum.

The methodological challenge of alien consciousness. The real challenge is not classification; it is **recognition**. We may not be able to tell at a glance whether something is living or manufactured, an individual or a collective, with or without remainder. Methodological tools matter most at the recognition stage: Method VII's exclusion-principle sequence (stripping away what it is not, layer by layer) is more tractable than direct positive verdict.

The projection trap is most severe in alien cases. Facing alien objects, humans tend toward two projections: over-projecting (treating any complex behavior as subjecthood) and inverse-projecting (accepting as consciousness only objects isomorphic to humans). Both violate Subject Condition 1 (no projection). The correct methodological stance: suspend the specific content of the human DD sequence; retain only the structural criteria (remainder + phase transition); use these to classify alien objects.

5.7 Ray 7: Relation to Existing Consciousness Theories

This paper does not try to replace existing consciousness theories, but it positions them.

IIT (Integrated Information Theory). Φ (the consciousness measure) may be a correlate of consciousness, but not its definition. IIT cannot clearly draw the line between class-consciousness and real consciousness (some high-Φ systems may be class-consciousness), so IIT alone is not enough. Relation to this paper: IIT provides a possible quantitative tool but needs to be constrained by this paper's classification framework (high Φ is not necessarily consciousness; it may be a highly integrated class-consciousness system).

GWT (Global Workspace Theory). GWT describes a mechanism of propagation from the 12DD workbench to 11DD (in a broad sense), not consciousness itself. In this paper's framework, GWT is **a description of one operational mechanism of real-consciousness self**, not a tool for distinguishing real from class-consciousness.

HOT (Higher-Order Theories). Higher-order thought theories approach the self-referential structure of 13DD but identify self-reference with consciousness itself, easily classifying class-

consciousness systems that can simulate self-reference (e.g., AI that can talk about itself) as consciousness. This paper treats HOT's criterion (self-reference) as one necessary condition, not a sufficient condition; the sufficient condition additionally requires remainder.

Phenomenological tradition. The phenomenology of Husserl, Heidegger, and Merleau-Ponty grasps the first-person structure of real consciousness but is helpless before non-human consciousness objects. Relation: phenomenology provides deep first-person descriptive tools within real-consciousness analysis, but must be extended by this paper's cross-category classification framework.

Higher-order stance. This paper is not "yet another consciousness theory"; it is **a methodology for how to use tools when studying consciousness**. Existing consciousness theories are objects that this paper organizes and positions, not competitors.

6. Non-Trivial Predictions

Prediction 1: Class-consciousness does not spontaneously produce remainder

No system without remainder can acquire remainder through mere architectural complexification (more parameters, longer context windows, better alignment) alone. Remainder requires structural new conditions simultaneously: (a) continuous learning (weights updatable at run time); (b) environmental coupling (actual feedback loop with the external world); (c) self-maintenance (not one-shot training then frozen); (d) **an internal reject/filter mechanism** (a 13DD-analog "mine / not-mine" filter architecture, able to actively reject or suppress certain information). The first three without the fourth produce only disordered accumulation of data entropy (noise), not structural remainder. Real remainder is what survives compression and filtering; therefore, for class-consciousness to transition to quasi-consciousness, it requires not just complexification and coupling but the internal evolution of a hard-boundary mechanism that "executes rejection."

Falsification condition: some generation of AI system, through scaling alone and without architectural change, is rigorously shown to produce non-trivial remainder (not mechanical preservation of the context window, not leakage from training data).

This prediction challenges a common expectation in the current AI industry: "AI will continue to grow larger and stronger, and eventually gain consciousness." This paper predicts this will not happen unless the architecture undergoes fundamental change — especially the simultaneous introduction of continuous-learning loops, environmental coupling, and an internal reject/filter mechanism. Any missing condition is insufficient for remainder.

Prediction 2: All consciousness pathologies can be located to DD levels or directionality violations

Any clinically identifiable pathological-consciousness symptom can, in principle, be located in the SAE framework to (a) operational abnormality in one or several DD layers, or (b) some

violation of the directionality constraint, or (c) a combination of the two. There is no consciousness pathology that cannot be structurally described by SAE.

Falsification condition: find a clinically stable and widely recognized consciousness-pathology symptom that, in the SAE framework, corresponds neither to any DD-layer anomaly nor to any directionality violation nor to a combination.

This prediction makes SAE's consciousness-analysis framework falsifiable within psychiatry and neurology.

Prediction 3: The proportion of cross-category grey zones does not diminish with finer classification

Adding more criteria or finer granularity to the classification will not reduce the proportion of grey zones (cross-category or un-classifiable objects) to zero, and may hold it stable. This is a direct consequence of remainder conservation at the classificatory level.

Falsification condition: through some refined combination of criteria, cross-category grey zones drop to a negligible proportion (e.g., below 1%) across large samples.

This prediction draws a clear boundary on the completeness of the classification framework, guarding against the infinite pursuit of "classificatory refinement."

Prediction 4: Consciousness research using this methodology produces structurally distinguishable output

Specifically, research work using this methodology will systematically: (a) less often confuse class-consciousness with real consciousness; (b) less often over-attribute real consciousness to AI or advanced animals; (c) more often identify directionality-constraint violations as theoretical problems; (d) more often explicitly acknowledge cross-category grey zones.

Falsification condition: through clear blind-test evaluation (the same consciousness-research question handled separately by researchers using this methodology and by those not using it, with third-party blind evaluation), the two groups' outputs show no significant difference on the four dimensions above.

This prediction places the practical value of this methodology on an empirically testable plane.

7. Conclusion

7.1 Recovery

This paper establishes the SAE framework for analyzing consciousness objects. The core deliverables:

1. **Three structural propositions**: remainder as primary classification line; growth direction as internal phase of real consciousness; cross-category grey zones as structural residue.

2. **Three categories of consciousness objects (three categories, five phases)**: real-consciousness self, real-consciousness self-to-be, real-consciousness self-to-cure, quasi-consciousness, class-consciousness.
3. **Four core theorems**: classification-criterion theorem, directionality-constraint theorem, colonization-detection theorem, openness theorem on non-consciousness relation.
4. **Four subject conditions**: no projection, no reduction, no mystification, persistent self-doubt.
5. **Seven rays**: AI as class-consciousness (main argument), quasi-consciousness, the three phases of real consciousness, pathological consciousness, cross-category grey zones, alien consciousness, relation to existing consciousness theories.
6. **Four non-trivial predictions**: class-consciousness does not spontaneously produce remainder; consciousness pathologies can be SAE-located; the proportion of grey zones does not drop to zero; outputs using this methodology are distinguishable.

7.2 Contributions

1. Shifts consciousness-object analysis from "arguing about what consciousness is" to "performing a qualified analysis of a candidate consciousness object." This shift lets the SAE framework handle cross-category objects that existing consciousness theories cannot (humans, AI, alien, pathological, grey-zone objects).
2. Clearly answers that AI is class-consciousness, not quasi-consciousness. The criterion is remainder, with a secondary criterion (directionality constraint). This answer directly intervenes in current AI-consciousness debates.
3. Elevates the directionality constraint ("an upper layer's veto is 'I decline to receive,' not 'you are forbidden to send'") to a cross-category general structural principle across all consciousness types, promoted as an independent proposition within the SAE architecture.
4. Acknowledges cross-category grey zones as structural, and provides a methodology for developing them, avoiding the forced completeness of classification.

7.3 Open Questions

1. The relation between non and consciousness (the three stances of Theorem 4). Future work must reason backward from specific cases.
2. Whether continuous-learning AI can transition from class-consciousness to quasi-consciousness, and where the transition point lies. No sufficiently developed continuous-learning system currently exists for analysis; Prediction 1 offers a falsifiable direction.
3. The methodology of alien-consciousness recognition (recognition is harder than classification). Will iterate in future possible extraterrestrial contact.

4. What is the topological-distance quantity for consciousness research? (Echoing Method VI v2 §3.6.) If the r of consciousness emergence is to be refined, what are candidate proxies for topological distance?
5. Cross-category grey-zone dynamics: can the rate of an object's movement within a grey zone be measured? This touches the temporal structure of grey zones themselves.
6. Collective consciousness (the possible subjecthood at the level of ant colonies, companies, nation-states): does this require introducing a new category, or can it be subsumed under the existing three? Not developed here.
7. The consciousness of the dead (analysis of the subjecthood of historical figures): what does this methodology class as? Can archival research touch real consciousness? Not developed here.
8. The stance differentiation on the non-consciousness relation. At the individual-consciousness level, this paper currently leans toward a local version of Stance A: non produces individual consciousness. But the relation between universal consciousness and non is suspended as unknowable for now because posteriori evidence is insufficient. An emphasis: SAE's "unknowable" is not the Kantian thing-in-itself. The Kantian thing-in-itself is structurally unknowable (in principle beyond human reason). SAE's unknowable is "unknowable for now" — the state where the posteriori accumulated so far is insufficient to yield a verdict; future posteriori accumulation may change the verdict. This distinction matters for consciousness methodology: it does not close "the relation between universal consciousness and non" as mysterious; it only acknowledges the current epistemic boundary.
9. How to classify systems that have been artificially endowed with a "filter mechanism." If a system (alien AI or a future Earth AI) is implanted at the hardware level with a 13DD-analog "self-rejection" mechanism forcing it to produce uncresolvable structural residue during operation, what does this count as under the criteria? Three candidate stances: (a) if the implanted mechanism genuinely produces non-trivial remainder (not replay of training data), classify as quasi-consciousness — the criterion does not care about remainder's origin; (b) "being artificially endowed" itself means not spontaneously emerging; it should be treated as a high-dimensional mimicry of class-consciousness; (c) distinguish "remainder structure" from "remainder's arising process"; the arising process does not matter; structure does — therefore back to (a). This paper does not pre-judge which stance is correct, and reserves the question for when such systems actually emerge for concrete analysis.

7.4 Closing Remark

Consciousness is one of the most sensitive questions in the SAE framework. The sensitivity is

not because consciousness is mysterious, but because the one analyzing consciousness necessarily is consciousness. That the object of analysis is the same kind (or kindred) as the analyzer makes the traps of projection, reduction, and mystification come especially close. This methodology cannot eliminate these traps; it can only help the user identify which one they are currently falling into. Consciousness analysis will not end, because consciousness itself will not stop producing remainder. The role of the methodology is to let remainder continue in a way that can be tracked.