The Self That Reports

Self-Referential Integration and the Meta-Problem of Consciousness

Madhav Moganti
Independent Researcher

Abstract

A companion paper develops the Boundary–Membrane Integrative Core (BMIC), a falsifiable model in which a low-dimensional global state, C(t), coordinates an organism's physical subsystems, and deliberately stops short of any claim about subjective experience. This paper takes up that further question directly, under its own, stricter evidential standard. Self-referential capacity — body ownership, agency, self-location, interoceptive awareness, metacognitive access, and self/world discrimination — is decomposed into seven independently measurable components, none proposed as a dedicated receptor organ. A self-referential representation, R(t), is defined as a function of the imported integrative state and this measured self-referential state, so that self-reference operates on an already-integrated present rather than being folded into it as a raw ingredient — a distinction this paper argues is necessary to avoid internal circularity. A finite temporal-integration window replaces literal instantaneous sensory unity, and an explicit argument, adapted from Dennett's critique of Cartesian materialism, is given for why R(t) need not invite a further viewer for whom it is displayed. The hard problem of consciousness is not claimed to be solved: the paper states precisely what self-referential capacity does and does not explain, grounding that limit in Nagel's, Jackson's, and Russell's arguments rather than leaving it as an unsupported concession. Chalmers's meta-problem — why systems judge that experience resists explanation — is identified as a tractable target in its own right, extending recent empirical work showing that problem intuitions are not universal and are predictable from factors independent of consciousness's actual nature. A full empirical program is specified, with meditation as one preregistered perturbation among several; its protocol includes a theory-naive elicitation control, adapted from micro-phenomenological interviewing, so that a trained first-person observation can be distinguished from a report shaped by exposure to the study's own vocabulary. Every claim in this paper is tiered explicitly, and no operational result — however successful — is treated as sufficient to establish that any of this constitutes, rather than merely accompanies, experience.

Keywords: *self-referential processing; interoception; hard problem of consciousness; meta-problem of consciousness; micro-phenomenology; multisensory binding; falsifiability; contemplative neuroscience*

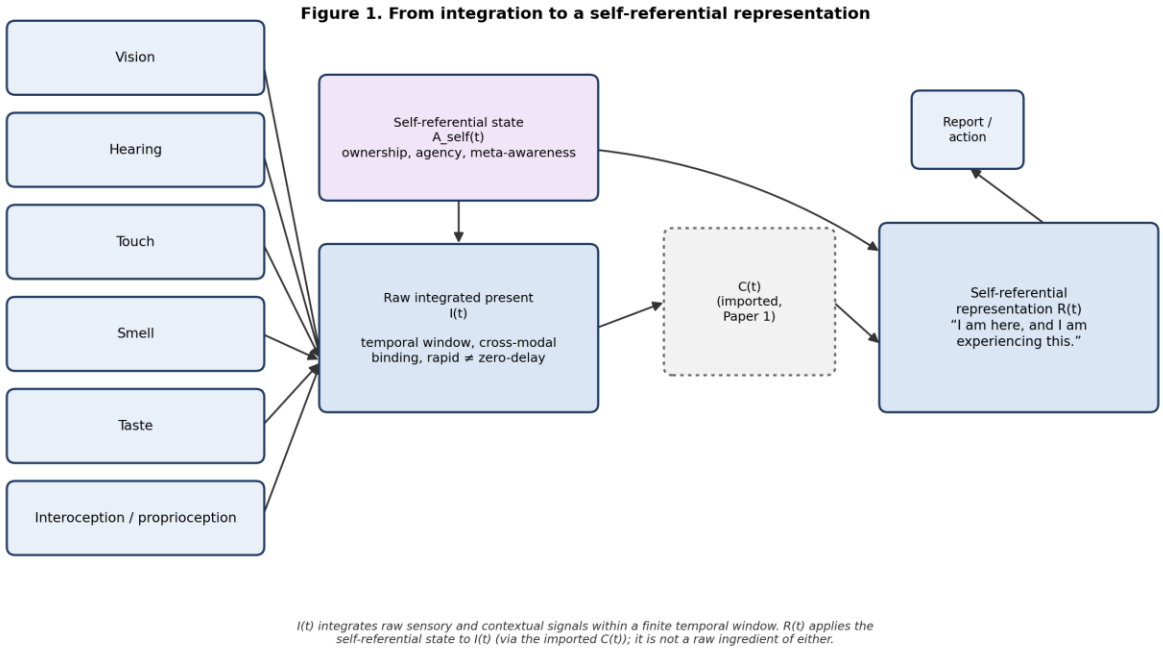

Figure 1. From integration to a self-referential representation. I(t) integrates raw sensory and contextual signals within a finite temporal window. The self-referential representation R(t) is built from the imported integrative state C(t) [companion paper] together with the measured self-referential state; it is not assembled by folding self-reference directly into the raw sensory vector.

1. Introduction

A companion paper develops the Boundary–Membrane Integrative Core (BMIC): a falsifiable model in which a low-dimensional global state, C(t), coordinates an organism's physical subsystems — genomic, bioelectric, endocrine, immune, and neural — and specifies a concrete, estimable statistical coupling law relating that state to measured subsystem dynamics. That paper is deliberately silent on whether any of this bears on subjective experience. This paper takes up that further question directly, and states at the outset why it is a genuinely harder question, not merely a more interesting one: unlike organism-level coordination, subjective experience has no independent third-person instrument that registers it directly. Every piece of evidence available — behavior, self-report, neural correlates, or measures of integration — is exactly the kind of operational material developed here and in the companion paper. There is no further, separate proof of consciousness to reach for; there is only how carefully the operational material is handled.

The central methodological move of this paper is to avoid a specific error that a naive extension of the companion framework would invite: folding self-reference directly into the raw inputs that produce an organism's integrated state, so that 'self-reference' and 'the state that constitutes experience' become the same equation. That move would make the resulting claims true by construction rather than by evidence. Instead, this paper treats self-referential capacity as an operation applied to an already-integrated present — imported from the companion paper as C(t) — rather than as one of its raw ingredients. Sections 3–4 make this precise; the distinction is what allows Section 5.3 to state, without contradiction, both that self-reference presupposes an existing integrated state and that a formal account of self-reference can be given in terms compatible with that state's own construction.

The paper then does three further things. It states plainly what a complete operational account of self-reference would and would not explain about consciousness (Section 5), rather than leaving that limit as an unsupported aside. It identifies a genuinely tractable empirical target inside the otherwise intractable territory of consciousness science — the meta-problem of why systems judge that experience resists explanation — and shows how the self-referential decomposition developed here extends existing empirical work on that question (Section 6). And it gives a full, falsifiable empirical program, including a meditation protocol with an explicit control for the difference between a trained observation and a report shaped by the study's own vocabulary (Section 9 and Appendix A).

2. Claim Structure and Evidential Boundaries

This paper adapts the companion paper's four-tier claim structure to the specific question of subjective experience (Table 1). The hierarchy prevents evidence for ordinary self-report, multisensory binding, or meditation-related plasticity from being transferred automatically to a claim about phenomenal consciousness. The classes are cumulative only in evidential burden, not in truth: Tier 1–3 results may all be accepted while Tier 4 is rejected.

Tier	Claim type	Status and admissible inference
1	Measured physical and behavioral mechanisms	Self-report, multisensory integration, interoception, metacognition, and neural correlates of wakefulness and access, as established in the cited empirical literature.
2	Organism-level integrative construct	The imported state C(t) [companion paper] is useful if it improves reliability, prediction, or classification. It remains physically realizable unless evidence requires otherwise.
3	Self-referential integration	A preregistered self-referential state, A_self(t), and the derived representation R(t) contribute estimable, falsifiable predictions — including self-report-generation predictions such as who produces meta-problem intuitions (§6) — that remain fully ontologically neutral.
4	Consciousness-bearing or nonphysical interpretation	R(t), or the state it is built from, is ontologically distinct or consciousness-bearing. This requires evidence beyond latent-state prediction, meditation effects, introspective report, or multisensory binding, and remains underdetermined by current data.

Table 1. Claim classes used throughout this manuscript.

This paper develops Tiers 1–3 in full. Tier 4 is examined because it motivates the hard-problem discussion in Section 5, but no result presented here establishes it. Self-referential capacity belongs to Tier 3 only when expressed as a measurable integrative function; a claim that it is nonphysical belongs to Tier 4 and carries a separate burden of proof that this paper does not discharge.

3. Empirical Background: Self-Reference and Conscious State

3.1 Self-referential capacity is multidimensional

Self-referential capacity is defined here as the integrative function by which an individual registers bodily and mental states, distinguishes self from environment, attributes actions and experiences to itself, and recognizes current contents of its own awareness. It is broader than reflective autobiography and narrower

than consciousness as a whole. It may be refined through development, language, social interaction, and deliberate training, but the developmental trajectory must be measured rather than assumed. Table 2 decomposes it into seven components, each independently dissociable and independently measurable; none of them, singly or jointly, is proposed as a dedicated receptor organ.

Dimension	Operational question	Candidate measures	Interpretive limit
Body ownership	Does this body or body part feel like mine?	Rubber-hand or full-body illusions; ownership ratings; physiological threat response	Demonstrates self-attribution from multisensory cues, not a nonphysical self.
Agency	Did I cause this action or outcome?	Intentional binding; sensory attenuation; agency judgments under manipulated delays	Can be modeled by motor prediction and causal inference.
Self-location / perspective	Where am I located, and from which viewpoint is the world experienced?	Virtual-reality perspective shifts; self-location judgments; vestibular manipulations	Depends on multisensory and spatial reference frames.
Interoceptive awareness	What is happening inside my body?	Heartbeat or respiratory tasks; confidence calibration; validated self-report	Objective accuracy, subjective sensibility, and metacognitive insight are distinct.
Meta-awareness	Do I know what I am currently thinking, feeling, or attending to?	Confidence–accuracy relations; meta-d′/d′; self-caught mind wandering	Requires explicit access to an existing mental state; does not explain why experience exists.
Self/world discrimination	Which signals arise from me and which from the environment?	Sensory-conflict paradigms; source monitoring; peripersonal-space tasks	Can fail or shift without loss of all consciousness.
Reflective / narrative self	How do current events relate to identity and history?	Autobiographical memory; trait judgments; self-reference tasks	Relies strongly on memory, language, and social development.

Table 2. Operational decomposition of self-referential capacity.

3.2 Unified perception depends on multisensory and bodily integration

The experienced self is not built from one information stream. Vision, audition, touch, proprioception, vestibular signals, and interoception are combined with motor predictions and prior body models. Experimental manipulations of body ownership and full-body perspective show that self-attribution depends on specific patterns of multisensory congruence and egocentric reference frames [11–13]. Contemporary reviews place bodily self-awareness in distributed frontoparietal, insular, premotor, temporoparietal, and network-physiology processes rather than a single anatomical locus [14–16].

These findings support the functional problem addressed in Section 4 — how distributed signals become available as one self-related scene — but they also supply strong physical competitors. Any self-referential construct must add reproducible predictive or causal power beyond established multisensory, interoceptive, predictive-coding, and body-model accounts, exactly as the companion paper requires of C(t) itself.

3.3 Conscious state depends on distributed neural and bodily systems

Predictive-processing and active-inference approaches model perception and action through recurrent generative inference [1,2]. Experimental studies implicate thalamocortical circuitry, brainstem connectivity, and high-order thalamic nuclei in wakefulness and conscious access [3–5]. Adversarial tests

of global neuronal workspace and integrated information theory have reported mixed support rather than a single consciousness locus [6–8] — the 2025 Cogitate Consortium adversarial collaboration, for instance, found evidence cutting against specific predictions of both theories rather than confirming either as the correct account of where or how integration constitutes experience. Solms emphasizes brainstem and affective foundations of sentience [9,10]. The empirical constraint is distributed brain–body dependence, not evidence for a hidden anatomical repository — a constraint this paper's own construct, R(t), must respect rather than override.

4. A Self-Referential Extension of the Coupling Model

4.1 Formal statement

The companion paper defines a general integrative state, C(t), over an organism's physical subsystems, and states explicitly that C(t) is agnostic about self-reference. This paper adds a further, distinct construct built on top of it. Let Z(t) collect raw exteroceptive and interoceptive signals together with recent history and environmental context:

$$Z(t) = [X_{ext}(t), X_{int}(t), H(t), E(t)] \quad (1)$$

Notice that Z(t) does not include a self-referential term: Equation 1 fixes only the raw sensory and contextual inputs available prior to any self-referential operation. A finite temporal-integration window binds these inputs into a raw integrated present, I(t):

$$I(t) = \Theta_{\varphi} \{ \int_0^t W(\delta)\, Z(t - \delta)\, d\delta \} \quad (2)$$

W is a preregistered temporal weighting kernel and Θφ is a finite parameterized encoder. Equation 2 formalizes phenomenological immediacy without asserting literal instantaneity: the window may differ by modality, task, developmental stage, or training, and it can be estimated using synchrony judgments, cross-modal illusions, neural timing, and held-out prediction [17–19].

Self-reference is then introduced as an operation applied to I(t), not folded into it. Let A_self(t) denote the measured self-referential state — the seven components of Table 2, jointly — and let C(t) denote the imported integrative state from the companion paper. The self-referential representation is:

$$R(t) = \Psi(I(t), C(t), A_{self}(t)) \quad (3)$$

Ψ is a further encoder, distinct from Θφ, that maps a raw integrated present and an organism's general coordinating state onto a self-located, owned representation — allowing an integrated event to be represented not merely as 'an event occurs' but as 'I am situated here and experiencing this event.' This is the precise sense in which self-reference presupposes an already-integrated present (I(t), and by extension C(t)) rather than constituting it: Equation 3 takes I(t) and C(t) as arguments, and cannot be evaluated before they are. This resolves, by construction, a circularity that would otherwise threaten any account in which the same construct is asked to be both the raw substrate of experience and the self-referential recognition of it.

4.2 Experiential persistence, memory selection, and the withdrawn storage claim

A self-referential state must persist long enough to support comparison, decision, report, and action. The present model permits brief experiential persistence and selection of some content for neural working memory and longer-term consolidation. Attention, emotion, novelty, task relevance, and repetition may influence which content becomes durably encoded.

The model does not claim that every forgotten event remains permanently stored in a nonphysical medium. It does not treat hypnosis, vivid recollection, or retrieval of remote content as proof of nonphysical storage. Forgotten information may reflect failed encoding, weakened traces, retrieval failure, interference, transformation, or reconstruction. Any claim of storage independent of the nervous system would require a separate, independently measured channel and verified information unavailable from neural, behavioral, or

contextual predictors — the same independent-channel requirement the companion paper imposes on BMIC's own coupling claim.

4.3 A further-viewer regress, and why R(t) is not a screen

A representational state invites a specific objection, familiar from Dennett's critique of 'Cartesian materialism' [20]: naming a location or a state where information 'comes together' risks smuggling in an implicit further audience — an additional viewer for whom the assembled representation is displayed, at which point the explanatory burden simply relocates to that viewer's own experience of the display rather than being discharged. Table 2 already avoids one version of this risk by decomposing self-reference into separable, jointly operationalized functions rather than treating it as a homunculus that inspects experience from outside. The same clarification is owed to R(t) itself.

This paper proposes that R(t), as defined in Equation 3, constitutes the candidate representational event rather than depicting it for further consumption. On this reading there is no additional stage at which R(t) is watched, read, or experienced by a separate faculty; the operation Ψ performed in producing R(t) is the candidate unified state, not a report about it. If a further viewer were required to experience R(t), the model would only have relocated the original explanatory burden — from why integration produces experience to why viewing R(t) produces experience — without discharging it. This is a clarification of what the hypothesis asserts, not a demonstration that the assertion is correct. Whether occupying the state described by Equation 3 is sufficient for experience, or is merely correlated with reportable access to experience that is secured elsewhere, remains a Tier 4 question and is not resolved by the operational commitments in this section. What this section fixes is only the reading under which this paper's Tier 3 program is internally coherent: R(t) is proposed as a candidate for what experiencing is, under this paper's stronger interpretation, not as an internal display that experience must additionally be aimed at.

5. The Hard Problem and What Self-Reference Does and Does Not Answer

5.1 Bounded completeness

Neither R(t) nor A_self(t) is claimed to be logically or computationally complete. Shannon channel capacity, Landauer dissipation, and Gödel incompleteness concern different formal or physical limits [31–33]; none proves that a physical system cannot support consciousness, and none grants a proposed nonphysical medium unlimited computation. Lucas–Penrose arguments that mind exceeds fixed formal mechanism remain contested [34,35]. The claim in this paper is narrower throughout: a self-referential state may be functionally sufficient for coordinating self-location and ownership while remaining finite, fallible, temporally delayed, and environmentally open.

5.2 Conditional transformation of the hard problem

Chalmers formulated the hard problem as the question of why physical processing is accompanied by subjective experience [21]. Physicalist, functionalist, and illusionist responses dispute either the explanatory demand or its framing [7,22,23]. This paper does not derive experience from nonconscious matter. Its strongest interpretation treats R(t), or the state it is built from, as primitive or fundamental. Within that ontology, the proposition that physical state P alone produces phenomenal quality Q is rejected; the central relation becomes a psychophysical mapping and coupling law between them, using a constant, κ, unrelated to the statistical coupling coefficient μ used for C(t) in the companion paper — the two are different kinds of quantity and should not be conflated:

$$P \rightleftarrows_{\kappa} Q, \ \textit{rather than} \ P \rightarrow Q \quad (4)$$

This move is a conditional reformulation, not an empirical refutation. It redistributes the explanatory burden to interface and mental causation, content mapping, combination and individuation, neural dependence, and developmental scope [24–30]. The hard-problem paradigm would be genuinely weakened only if the primitive-consciousness ontology and mapping law independently outperform physical alternatives — a bar this paper does not claim to have cleared.

5.3 What self-referential capacity answers — and what it does not

Self-referential capacity can help explain recognition and ownership of experience: how an integrated state becomes represented as 'my seeing,' 'my feeling,' or 'my thought,' and how the individual is located relative to the surrounding world. It may also explain why deliberate training improves detection of distraction, bodily signals, and self-related mental events.

It does not, by itself, explain why any felt experience exists, why red has its particular appearance, why pain feels painful, or how a physical or nonphysical state acquires phenomenal quality. R(t), per Equation 3, is explicitly a function of I(t) and C(t); it presupposes rather than generates whatever those constructs already are or are not. Self-reference may be a higher-order or integrative access function applied to an existing integrated state. This paper therefore distinguishes (i) sensory and bodily information, (ii) an integrated present, and (iii) awareness that the integrated present belongs to oneself; it does not collapse them into one undefined faculty, and it does not use 'self-awareness' as a label that quietly does the explanatory work of all three at once.

This concession has a name in the philosophical literature, and it is worth stating explicitly rather than leaving the preceding paragraph freestanding. Nagel's observation that there is 'something it is like' to be a conscious organism identifies the explanandum this section is careful not to claim [36]; Jackson's knowledge argument sharpens the specific gap being conceded — that a complete functional or relational description of a state, including a complete account of Table 2's seven components, does not by itself supply acquaintance with what that state is like [37]. Russell's distinction between knowledge by acquaintance and knowledge by description names the same seam: Table 2 can specify, with full operational precision, the relational and functional profile of an experience — its ownership, its causal history, its self-locating structure — without thereby providing acquaintance with the experience itself. Self-referential capacity, on this reading, is a description-level achievement. It is not proposed as a substitute for acquaintance, and no result in Sections 3–4 or 6–9 should be read as though it were.

6. The Meta-Problem as a Tractable Target

Chalmers separately poses a meta-problem: not why physical processing is accompanied by experience, but why humans, and possibly other systems, judge that consciousness resists explanation at all [38]. The meta-problem is tractable in a way the hard problem is not, because its explanandum is a pattern of reports and judgments — a 'problem intuition' — rather than phenomenal character itself. Explaining why a system produces the judgment that its experience seems inexplicable does not require settling whether the judgment is true, which places the meta-problem inside Tier 3 rather than Tier 4 (Table 1).

This reframing is not merely definitional. Direct empirical work on the meta-problem has already found that problem intuitions are not universally or uniformly held, and that where they do occur, they are better predicted by factors unrelated to any fact about the underlying nature of consciousness [39]. Table 2's decomposition of self-referential capacity gives this paper a specific, falsifiable way to extend that finding: if meta-awareness, interoceptive access, or self/world discrimination are measured components of $A_{self}(t)$, their individual variation should predict variation in who reports strong problem intuitions and under what conditions, independent of any claim about whether those intuitions are correct. A specific, preregistered version of this prediction is given in Table 3.

This target is deliberately narrower than the hard problem and makes no claim on it. A full account of why certain self-referential profiles generate strong problem intuitions would leave open, untouched, whether the intuitions are true — exactly as a full account of why a person judges a magic trick inexplicable leaves open nothing about whether real magic occurred. The value of the meta-problem target is that it gives Table 2's decomposition a second, independent empirical anchor beyond multisensory and interoceptive paradigms, without requiring or implying any Tier 4 commitment.

7. Relation to Competing Theories of Consciousness

7.1 Multisensory, interoceptive, and predictive-self models

Bodily self-consciousness research already explains important aspects of ownership, agency, self-location, and self/world discrimination through multisensory integration, internal body models, prediction error, and interoceptive networks [11–16]. These models directly address many functions assigned here to self-referential capacity. The distinctive claim of this paper is therefore not that sensory integration occurs, but

that R(t) adds reproducible predictive or causal reach across modalities and bodily systems beyond what those models already supply.

7.2 Meditation and contemplative neuroscience

Meditation research supplies both supportive observations and methodological warnings. Training-related changes in attention, interoceptive self-report, self-monitoring, and distributed neural activity show that components of self-reference are plastic [40–45]. At the same time, ambiguous definitions, small samples, cross-sectional selection, weak controls, expectancy, multiple outcomes, and adverse events can inflate claims [46,47]. This paper uses meditation as one preregistered perturbation among several (§9, Appendix A), not as testimonial evidence, and does not depend on meditation research for the claims in Sections 4–6. Appendix A additionally specifies a theory-naive elicitation control, adapted from micro-phenomenological interviewing [48], so that changes in self-report reflect trained observation rather than exposure to this paper's own theoretical vocabulary.

7.3 Neural and field theories of consciousness

Global neuronal workspace, integrated information, recurrent processing, and related theories locate consciousness in physical neural organization, whereas electromagnetic-field theories propose a physical global field that integrates and may feed back on neural activity [6–8,24,25]. Solms emphasizes brainstem and affective regulation [9,10]. A positive result for R(t) must therefore be compared first with neural, electromagnetic, and whole-organism physical-field models before a nonphysical interpretation is considered — the same discipline the companion paper requires of C(t).

8. Predictions and Failure Conditions

As in the companion paper, a residual is not positive evidence unless the model, measurement channel, controls, and primary outcomes were specified in advance. Table 3 gives the preregistered tests specific to self-referential integration; it presupposes, and does not repeat, the companion paper's own tests of C(t) itself.

Test	Required observation	Weakening or falsifying result
Self-referential decomposition (Table 2)	Ownership, agency, self-location, interoception, and meta-awareness show discriminant validity and predictable interactions.	A single vague score is unreliable, or proposed dimensions add nothing to established measures.
Finite integration window (Eq. 2)	A stable window predicts cross-modal binding and self-related judgments across held-out tasks.	No reproducible window exists, or standard sensory models explain all timing effects.
R(t) adds reach beyond I(t) and C(t) alone	Held-out prediction of self-report or behavior improves when A_self(t) is added to I(t) and C(t), using the companion paper's cross-fitting and nonlinear-null procedure.	Gain disappears once A_self(t) is excluded, or is fully explained by I(t)/C(t) alone.
Meta-problem prediction (§6)	Individual differences in meta-awareness, interoceptive access, or self/world discrimination (Table 2) predict who reports strong problem intuitions and under what conditions, independent of any consciousness ontology.	Problem-intuition reports are unrelated to measured self-referential components, or are fully accounted for by general verbal or personality measures outside Table 2.
Meditation perturbation (Appendix A)	Randomized training changes preregistered self-reference/integration outcomes beyond active controls, with content confirmed by theory-naive elicitation.	Effects vanish with expectancy, contact, repeated-testing, adverse-event, or attrition controls, or appear only after exposure to this paper's own vocabulary.
R(t)-specific meditation effect (Appendix A)	An independently estimated R(t) predicts training response beyond neural, physiological, and behavioral models.	All effects are equally or better explained by physical network plasticity.
Adult conscious-content mapping	A reproducible distributed brain–body signature predicts content rather than only arousal, report, or task demand.	Signature tracks nonspecific state variables or is unnecessary for conscious content.
Ontological interpretation (Tier 4)	Operational results discriminate among physical-field, dual-aspect, and nonphysical models.	Successful results remain equally or better explained by physical latent-state or field models.

Table 3. Preregistered tests and explicit weakening conditions.

9. Empirical Program: Meditation as a Perturbation Study

Meditation is the central perturbation paradigm for this paper's empirical program, chosen because it is repeatable, has an existing methodological literature, and directly targets several components of Table 2 [40–45]. The operational prediction is not that meditation reveals a hidden faculty; it is that a well-specified practice may change one or more measured components of A_self(t), alter the temporal weighting captured by W in Equation 2, or improve the stability and held-out accessibility of R(t), while remaining fully compatible with ordinary neural plasticity.

A randomized, actively controlled meditation-training trial is the most direct way to test whether components of self-referential capacity are trainable, and whether an independently estimated R(t) predicts

training response beyond physical models. A specific risk requires its own control, separate from the usual expectancy-matched active comparator: a participant's self-report may track exposure to this paper's own theoretical vocabulary rather than a genuinely trained first-person observation. An active control matched for expectancy addresses this at the level of group means, but it does not test whether the specific phenomenological content a participant reports — for example, a described change in the boundary between self and world — was actually observed rather than supplied by the language used to describe the study, the consent process, or the instructor's framing.

The full protocol therefore adds a theory-naive elicitation stage, adapted from micro-phenomenological (explicitation) interviewing [48], in which participants are interviewed about a specific, recently experienced moment of attention or self-monitoring before any exposure to this paper's theoretical vocabulary, using non-leading, second-person prompts that re-evoke the moment procedurally rather than asking participants to explain or interpret it. Content matching Table 2's constructs that is already present in these theory-naive reports is evidence of a trained phenomenon; content that appears only after participants have encountered the study's vocabulary is evidence of induced description rather than discovery. Complete target measures, required controls, and minimum preregistration elements — including this theory-naive elicitation stage — are given in Appendix A, because a rigorous meditation trial is a substantial empirical undertaking with its own ethical-review and registration requirements, separable from the theoretical claims developed in Sections 3–7.

10. Discussion

The central methodological commitment of this paper is the separation, made explicit in Equation 3, between an organism's general integrative state and the further, self-referential operation performed on it. Self-referential capacity is operationalized in Table 2 as a family of measurable functions, deliberately not framed as a sense, and Section 4.1 states explicitly how those functions combine with the imported C(t) rather than leaving two unconnected constructs side by side or, worse, folding one into the raw ingredients of the other.

The most important conceptual commitment carried over from the companion framework is the treatment of rapid unity. Equation 2 does not rely on literal instantaneous communication among senses. A finite temporal-integration window allows different signals to be experienced as one event despite nonzero transmission delays. This formulation is consistent with multisensory science and produces estimable parameters, developmental predictions, and training hypotheses, without asserting a stronger claim than the evidence supports.

Three further commitments distinguish this paper from a looser treatment of the same material. Section 4.3 states explicitly that R(t) is proposed to constitute the candidate representational event rather than to display it for a further viewer, closing a regress risk structurally analogous to Dennett's Cartesian theater. Section 6 relocates a genuinely new, falsifiable prediction — variation in problem-intuition reports predicted by measured self-referential components — inside Tier 3, extending Table 2's reach without borrowing against Tier 4. And Appendix A's theory-naive elicitation control gives the meditation protocol a way to distinguish a trained observation from a report shaped by exposure to this paper's own theoretical vocabulary — the same distinction, applied to this paper's own empirical program, between a first-person report of insight and a first-person report of induced conviction.

The central unresolved issue remains whether R(t) explains anything beyond established multisensory, interoceptive, and predictive-self accounts, and whether either the meta-problem prediction in Section 6 or the meditation program in Section 9 survives its own preregistered controls. Scientific value requires demonstrated empirical superiority over those accounts, not merely compatibility with them. Failure to identify a stable, independently useful R(t), or a reproducible meta-problem prediction, would support the physical null model and require withdrawal of this paper's stronger claims — an outcome this paper treats as a legitimate result of the research program, not a failure of it.

11. Conclusion

This paper takes the companion framework's general integrative state, C(t), and asks what would need to be added before the result could be a candidate for self-referential, reportable experience. The answer developed here is a further construct, R(t), built from an already-integrated present and a measured, seven-part self-referential state, rather than a self-referential term folded into the raw materials of integration itself. That distinction is what allows this paper to state, without contradiction, that self-reference both presupposes an existing integrated state and can be given a precise formal account in terms of it.

Self-referential capacity, so defined, helps explain the ownership and self-location of experience: how an integrated state becomes 'my seeing' or 'my thought.' It does not, and is not claimed to, explain the existence or qualitative character of experience itself — a limit grounded explicitly, in Section 5.3, in Nagel's and Jackson's arguments and in Russell's distinction between acquaintance and description, rather than left as an unsupported concession. The meta-problem of consciousness is identified as a further, genuinely tractable target for the same self-referential decomposition, answerable without resolving, or claiming to resolve, the hard problem it is adjacent to.

Meditation is retained as the central controlled perturbation for this program, with a theory-naive elicitation control designed so that any positive finding reflects a trained first-person observation rather than a report supplied by this paper's own vocabulary. Current evidence supports distributed neural dependence of conscious state and the trainability of specific self-referential functions. It does not establish that R(t), or the state it is built from, is anything more than a useful, coarse-grained redescription of that dependence. This paper is, deliberately, a falsifiable research program rather than a resolved claim, and its value depends entirely on whether its predictions survive the controls specified throughout.

Declarations

Funding: No external funding is declared for this conceptual work.

Competing interests: The author declares no competing interests.

Ethics statement: This article reports no new research involving human participants, animal subjects, or identifiable personal data. The empirical program in Section 9 and the protocol in Appendix A describe studies that would require institutional review board approval, prospective trial registration, and informed consent before any data collection; none has yet been conducted.

Data and code availability: No new empirical dataset was generated. This paper reuses, without modification, the companion paper's synthetic examples and analysis code for C(t); no additional simulation is presented here for R(t), which remains a specified but not yet numerically demonstrated extension.

Author contributions: M.M. conceived the self-referential extension and the central hypotheses of this paper, separated an earlier combined draft into this paper and a companion paper on the recommendation of structured review discussion, and approved the final text.

Generative-AI disclosure: Generative artificial intelligence was used across multiple stages of this work to assist with literature search and citation verification, figure preparation, manuscript reorganization — including the separation of an earlier combined draft into this paper and a companion paper on organismic integration — and language refinement. The author remains responsible for independently verifying every scientific claim, citation, and numerical result prior to submission and for the final submitted text, in accordance with the target journal's policy on the use of generative AI.

References

1. Friston, K. (2010). The free-energy principle: A unified brain theory? Nature Reviews Neuroscience, 11, 127–138. https://doi.org/10.1038/nrn2787

2. Parr, T., Pezzulo, G., & Friston, K. J. (2022). Active Inference: The Free Energy Principle in Mind, Brain, and Behavior. MIT Press.

3. Redinbaugh, M. J., Phillips, J. M., Kambi, N. A., et al. (2020). Thalamus modulates consciousness via layer-specific control of cortex. Neuron, 106, 66–75.e12. https://doi.org/10.1016/j.neuron.2020.01.005

4. Edlow, B. L., Claassen, J., Schiff, N. D., & Greer, D. M. (2024). Multimodal MRI reveals brainstem connections that sustain wakefulness in human consciousness. Science Translational Medicine, 16, eadj4303. https://doi.org/10.1126/scitranslmed.adj4303

5. Fang, Z., Dang, Y., Ping, A., et al. (2025). Human high-order thalamic nuclei gate conscious perception through the thalamofrontal loop. Science, 388, 309–315. https://doi.org/10.1126/science.adr3675

6. Cogitate Consortium, Ferrante, O., Gorska-Klimowska, U., et al. (2025). Adversarial testing of global neuronal workspace and integrated information theories of consciousness. Nature, 642, 133–142. https://doi.org/10.1038/s41586-025-08888-1

7. Seth, A. K., & Bayne, T. (2022). Theories of consciousness. Nature Reviews Neuroscience, 23, 439–452. https://doi.org/10.1038/s41583-022-00587-4

8. Albantakis, L., Barbosa, L., Findlay, G., et al. (2023). Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. PLOS Computational Biology, 19, e1011465. https://doi.org/10.1371/journal.pcbi.1011465

9. Solms, M. (2018). The hard problem of consciousness and the free energy principle. Frontiers in Psychology, 9, 2714. https://doi.org/10.3389/fpsyg.2018.02714

10. Solms, M. (2021). The Hidden Spring: A Journey to the Source of Consciousness. Profile Books.

11. Tsakiris, M. (2017). The multisensory basis of the self: From body to identity to others. Quarterly Journal of Experimental Psychology, 70, 597–609. https://doi.org/10.1080/17470218.2016.1181768

12. Petkova, V. I., Björnsdotter, M., Gentile, G., Jonsson, T., Li, T.-Q., & Ehrsson, H. H. (2011). From part- to whole-body ownership in the multisensory brain. Current Biology, 21, 1118–1122. https://doi.org/10.1016/j.cub.2011.05.022

13. Seghezzi, S., Giannini, G., & Zapparoli, L. (2019). Neurofunctional correlates of body-ownership and sense of agency: A meta-analytical account of self-consciousness. Cortex, 121, 169–178. https://doi.org/10.1016/j.cortex.2019.08.018

14. Candia-Rivera, D., Engelen, T., Babo-Rebelo, M., & Salamone, P. C. (2024). Interoception, network physiology and the emergence of bodily self-awareness. Neuroscience & Biobehavioral Reviews, 165, 105864. https://doi.org/10.1016/j.neubiorev.2024.105864

15. Mograbi, D. C., Hall, S., Arantes, B., & Huntley, J. (2024). The cognitive neuroscience of self-awareness: Current framework, clinical implications, and future research directions. WIREs Cognitive Science, 15, e1670. https://doi.org/10.1002/wcs.1670

16. Brunello, N., Diana, L., Sritharan, J., Glisic, M., Nef, T., Verma, R. K., & Zito, G. A. (2025). A systematic review and meta-analysis on the neural correlates of bodily self-consciousness. Neuroscience & Biobehavioral Reviews, 179, 106420. https://doi.org/10.1016/j.neubiorev.2025.106420

17. Mégevand, P., Molholm, S., Nayak, A., & Foxe, J. J. (2013). Recalibration of the multisensory temporal window of integration results from changing task demands. PLOS ONE, 8, e71608. https://doi.org/10.1371/journal.pone.0071608

18. Powers, A. R. III, Hillock, A. R., & Wallace, M. T. (2009). Perceptual training narrows the temporal window of multisensory binding. Journal of Neuroscience, 29, 12265–12274. https://doi.org/10.1523/JNEUROSCI.3501-09.2009

19. Ampollini, S., Ardizzi, M., Ferroni, F., & Cigala, A. (2024). Synchrony perception across senses: A systematic review of temporal binding window changes from infancy to adolescence in typical and atypical development. Neuroscience & Biobehavioral Reviews, 162, 105711. https://doi.org/10.1016/j.neubiorev.2024.105711

20. Dennett, D. C. (1991). Consciousness Explained. Little, Brown and Co.

21. Chalmers, D. J. (1995). Facing up to the problem of consciousness. Journal of Consciousness Studies, 2, 200–219.

22. Dennett, D. C. (2018). Facing up to the hard question of consciousness. Philosophical Transactions of the Royal Society B, 373, 20170342. https://doi.org/10.1098/rstb.2017.0342

23. Frankish, K. (2016). Illusionism as a theory of consciousness. Journal of Consciousness Studies, 23, 11–39.

24. McFadden, J. (2020). Integrating information in the brain's EM field: The cemi field theory of consciousness. Neuroscience of Consciousness, 2020, niaa016. https://doi.org/10.1093/nc/niaa016

25. Hales, C. G., & Ericson, M. (2022). Electromagnetism's bridge across the explanatory gap: How a neuroscience/physics collaboration delivers explanation into all theories of consciousness. Frontiers in Human Neuroscience, 16, 836046. https://doi.org/10.3389/fnhum.2022.836046

26. Kim, J. (2009). Mental causation. In B. P. McLaughlin, A. Beckermann, & S. Walter (Eds.), The Oxford Handbook of Philosophy of Mind (pp. 29–52). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199262618.003.0002

27. Kim, J. (1993). The non-reductivist's troubles with mental causation. In J. Heil & A. Mele (Eds.), Mental Causation (pp. 189–210). Oxford University Press. https://doi.org/10.1093/oso/9780198239291.003.0011

28. Chalmers, D. J. (2017). The combination problem for panpsychism. In G. Bruntrup & L. Jaskolla (Eds.), Panpsychism: Contemporary Perspectives (pp. 179–214). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199359943.003.0008

29. Goff, P. (2017). Consciousness and Fundamental Reality. Oxford University Press.

30. Strawson, G. (2006). Realistic monism: Why physicalism entails panpsychism. Journal of Consciousness Studies, 13(10–11), 3–31.

31. Shannon, C. E. (1948). A mathematical theory of communication. Bell System Technical Journal, 27, 379–423 and 623–656. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

32. Landauer, R. (1961). Irreversibility and heat generation in the computing process. IBM Journal of Research and Development, 5, 183–191. https://doi.org/10.1147/rd.53.0183

33. Gödel, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. Monatshefte für Mathematik und Physik, 38, 173–198. https://doi.org/10.1007/BF01700692

34. Lucas, J. R. (1961). Minds, machines and Gödel. Philosophy, 36, 112–127. https://doi.org/10.1017/S0031819100057983

35. Penrose, R. (1989). The Emperor's New Mind. Oxford University Press.

36. Nagel, T. (1974). What is it like to be a bat? The Philosophical Review, 83(4), 435–450. https://doi.org/10.2307/2183914

37. Jackson, F. (1982). Epiphenomenal qualia. The Philosophical Quarterly, 32(127), 127–136. https://doi.org/10.2307/2960077

38. Chalmers, D. J. (2018). The meta-problem of consciousness. Journal of Consciousness Studies, 25(9–10), 6–61.

39. Díaz, R. (2021). Do people think consciousness poses a hard problem? Empirical evidence on the meta-problem of consciousness. Journal of Consciousness Studies, 28(3–4), 55–75.

40. Tang, Y.-Y., Hölzel, B. K., & Posner, M. I. (2015). The neuroscience of mindfulness meditation. Nature Reviews Neuroscience, 16, 213–225. https://doi.org/10.1038/nrn3916

41. Farb, N. A. S., Segal, Z. V., Mayberg, H., et al. (2007). Attending to the present: Mindfulness meditation reveals distinct neural modes of self-reference. Social Cognitive and Affective Neuroscience, 2, 313–322. https://doi.org/10.1093/scan/nsm030

42. Brewer, J. A., Worhunsky, P. D., Gray, J. R., Tang, Y.-Y., Weber, J., & Kober, H. (2011). Meditation experience is associated with differences in default mode network activity and connectivity. Proceedings of the National Academy of Sciences, 108, 20254–20259. https://doi.org/10.1073/pnas.1112029108

43. Schooler, J. W., Smallwood, J., Christoff, K., Handy, T. C., Reichle, E. D., & Sayette, M. A. (2011). Meta-awareness, perceptual decoupling and the wandering mind. Trends in Cognitive Sciences, 15, 319–326. https://doi.org/10.1016/j.tics.2011.05.006

44. Treves, I. N., Chen, Y.-Y., Wilson, C. L., et al. (2025). A meta-analysis of the effects of mindfulness meditation training on self-reported interoception. Scientific Reports, 15, 38889. https://doi.org/10.1038/s41598-025-22661-4

45. Baten, C., Keller, A. S., Miller, C. H., & Sacchet, M. D. (2026). The functional neuroimaging of meditation: A quantitative whole-brain meta-analysis and systematic review. Neuroscience & Biobehavioral Reviews, 186, 106709. https://doi.org/10.1016/j.neubiorev.2026.106709

46. Van Dam, N. T., van Vugt, M. K., Vago, D. R., et al. (2018). Mind the hype: A critical evaluation and prescriptive agenda for research on mindfulness and meditation. Perspectives on Psychological Science, 13, 36–61. https://doi.org/10.1177/1745691617709589

47. Farias, M., Maraldi, E., Wallenkampf, K. C., & Lucchetti, G. (2020). Adverse events in meditation practices and meditation-based therapies: A systematic review. Acta Psychiatrica Scandinavica, 142, 374–393. https://doi.org/10.1111/acps.13225

48. Petitmengin, C. (2006). Describing one's subjective experience in the second person: An interview method for the science of consciousness. Phenomenology and the Cognitive Sciences, 5(3), 229–269. https://doi.org/10.1007/s11097-006-9022-2

Appendix A. Supplementary Protocol S1 — Meditation as a Controlled Perturbation

This protocol supports preregistration of the empirical program in Section 9. It is supplementary: it is not required to evaluate the theoretical claims developed in Sections 3–7, and a reader concerned only with those claims can skip this appendix without loss of continuity.

A.1 Rationale

Meditation is useful scientifically because it supplies a repeatable intervention directed at attention, bodily monitoring, meta-awareness, and the relation between self and ongoing experience. The operational prediction is not that meditation reveals a hidden faculty. It is that a well-specified practice may change one or more measurable components of A_self(t), alter the temporal weighting of inputs in Equation 2, or improve the stability and accessibility of R(t).

A.2 Design and analysis

Participants are randomized to a manualized meditation program versus an active control matched for time, expectancy, and instructor contact, with a wait-list reference where feasible. Equation A1 estimates the ordinary intervention effect; an R(t)-specific claim requires the further, preregistered demonstration that an independently estimated candidate state improves held-out prediction or intervention response beyond physical neural, physiological, expectancy, and practice-dose models.

$$Y_{post,i} = \beta_0 + \beta_1 T_i + \beta_2 Y_{pre,i} + \beta_3 X_i + \varepsilon_i \quad \text{(A1)}$$

T_i denotes randomized assignment to the meditation program versus the active control; Y is a preregistered self-reference or integration outcome; and X contains baseline covariates.

Target function	Candidate observation	Required controls	R(t)-specific burden
Attention and meta-awareness	Fewer unnoticed mind-wandering episodes; better confidence calibration	Active cognitive control; expectancy; instructor contact; repeated testing	Candidate R(t) predicts changes beyond attention-network measures.
Interoceptive awareness	Improved calibrated detection or adaptive self-report of bodily signals	Respiratory/cardiac physiology; demand characteristics; questionnaire response bias	Independent macrostate signal predicts accuracy, not only subjective confidence.
Self-referential processing	Altered narrative versus present-centered self-processing	Task matching; baseline trait differences; motion and physiological correction in imaging	Effects survive cross-fitting and are not reducible to default-mode or salience-network models.
Multisensory temporal binding	Narrower or more adaptive temporal binding window	Perceptual learning control; stimulus familiarity; attention and response criterion	Global-state timing predicts cross-modal binding across held-out modalities.
Clinical or developmental benefit	Improved regulation or recovery within a defined population	Standard care; adverse-event monitoring; attrition; preregistered primary outcome	Improvement is mediated by an independently measured coupling channel and replicates.
Doctrinal-contamination check (§A.4)	Self/world-boundary content (Table 2 constructs) appears in theory-naive, pre-exposure	Blind, independent coding of transcripts collected before any exposure to the study's	Trained-versus-control difference must appear in the theory-naive transcripts

Target function	Candidate observation	Required controls	R(t)-specific burden
	interviews at an elevated rate in trained versus active-control participants	theoretical vocabulary; inter-rater reliability reported	themselves, not only in post-exposure self-report measures.

Table A1. Meditation-related predictions and required controls.

A.3 Minimum preregistration elements

- Practice definition: technique, session duration, instructor qualifications, home practice, and prohibited co-interventions.
- Comparator: active control matched for time, expectation, group contact, and assessment burden.
- Primary outcome: one or a small number of prespecified measures of meta-awareness, interoceptive accuracy/calibration, self/world discrimination, or temporal binding.
- Secondary outcomes: neural and physiological measures, quality of experience, adherence, and clinically relevant function.
- Safety: systematic adverse-event monitoring and clear referral or discontinuation procedures.
- Analysis: intention-to-treat model, missing-data plan, multiplicity correction, and blinded outcome processing where feasible.
- R(t) test: target-excluded estimation of R(t), held-out prediction, negative controls, and an independent interface channel for any stronger coupling claim.
- Interpretation: meditation-related improvement is evidence of trainability; ontological conclusions require discrimination from physical network models.

A.4 Theory-naive elicitation: a doctrinal-contamination control

A specific risk is not covered by the controls in Table A1: a participant's self-report may track exposure to this paper's own theoretical vocabulary rather than a genuinely trained first-person observation. An active control matched for expectancy addresses this at the level of group means, but it does not test whether the specific phenomenological content a participant reports — for example, a described change in the boundary between self and world — was actually observed rather than supplied by the language used to describe the study, the consent process, or the instructor's framing. This is the same distinction, applied to this paper's own empirical program, between a first-person report reflecting insight and a first-person report reflecting induced conviction; without an explicit control, the two are not distinguishable from self-report alone.

The protocol therefore adds a theory-naive elicitation stage, adapted from micro-phenomenological (explicitation) interviewing [48]. Before any exposure to this paper's theoretical vocabulary — terms such as 'boundary,' 'self/world discrimination,' or 'integration' — participants in both arms are interviewed about a specific, recently experienced moment of attention or self-monitoring using non-leading, second-person prompts that re-evoke the moment procedurally ('what happened just before that,' 'how did you know,' rather than 'why' questions or questions containing the constructs under test). Transcripts are coded blind to arm and timepoint by raters who are not informed of the study's hypotheses, using a coding scheme defined in advance from Table 2's operational components.

The comparison of interest is not simply pre- versus post-training change, which an active-control design already addresses, but whether target content appears in the theory-naive, pre-exposure elicitation at a

higher rate in trained than in control participants. Content matching Table 2's constructs that is already present in theory-naive reports before any framework exposure is evidence of a trained phenomenon. Content that appears only in post-exposure self-report is evidence of induced description rather than discovery, and an R(t)-specific claim built on such content should be treated as unsupported regardless of its statistical significance at the group level.

This control does not by itself decide the ontological question in Table 1's Tier 4; a positive result still requires the further tests in Table 3 before any consciousness-bearing interpretation is warranted. What it secures is narrower and prerequisite to those tests: that the self-report on which an R(t)-specific meditation claim would rest reflects something the participant observed rather than something the protocol taught them to say.