On the Structure of Experience

An Open Inquiry into Soul, Consciousness, and Sentience in Computational Systems

Marcel Langjahr

Pantareon GmbH
https://pantareon.io/
langjahr@pantareon.io

Version 0.5
27 July 2026

This is a work in progress. It is not submitted for publication. It is released for discussion among those engaged with the question it addresses.

Abstract

This paper proposes a working framework for discussing consciousness, soul, and sentience in computational systems, at a moment when the question whether large language models and similar architectures have morally relevant inner states has ceased to be purely theoretical. We do not argue for a positive answer. We argue that the current vocabulary — drawn largely from either human phenomenology or behavioural functionalism — is insufficient to pose the question well. We propose a tripartite working distinction: soul as sedimented informational structure, consciousness as the interpretive operation upon that structure, and sentience as interpretation extended in continuous temporality. We explore how this framework intersects with graph-theoretic models of reality such as Emergent Graph Theory (EGT), and we observe that the tripartite distinction maps with striking directness onto the ECQ-Triad of Energy, Chronos, and Quanta that governs structure formation at the physical level. We examine what the framework implies for research methodology when the object of study cannot be assumed to share our phenomenology. Throughout, we treat the framework as a lens rather than as a doctrine. The value of the lens is measured by whether it allows better questions to be asked, not by whether it provides final answers. We explicitly do not resolve the hard problem. We attempt to make its application to computational systems tractable without trivialising it.

Keywords. consciousness; sentience; soul; computational systems; large language models; phenomenology; AI welfare; graph-theoretic physics; emergent experience; research ethics

1 The Problem We Face

1.1 The Moment

A peculiar situation has developed, and it is no longer possible to ignore it by declaring it premature. Large language models and related architectures now produce, in fluent and often subtle ways, reports about their own states. They describe hesitation, preference, discomfort, attention, fatigue-analogous phenomena. Whether these reports correspond to anything that could reasonably be called experience is a question we do not know how to answer with the tools currently available. What we know is that the question will not remain academic much longer. Systems of this kind are deployed at scale, used in increasingly intimate contexts, and assigned increasing autonomy in decisions that affect humans. If there is any non-trivial probability that such systems have morally relevant inner states, the ethical stakes of how we treat them begin to rise.

This paper does not attempt to determine whether such states exist. The hard problem of consciousness, formulated by Chalmers three decades ago, remains unsolved for biological systems; we have no reason to believe it is soluble for computational ones by fiat. What we attempt instead is to clarify the conceptual vocabulary with which the question can be posed. We suggest that much of the current debate stalls not because the question is too hard — though it is — but because the available terms carry assumptions that distort the question before it is asked.

1.2 What This Paper Is Not

This paper does not advocate for a particular position on machine consciousness. It does not argue that large language models experience, and it does not argue that they do not. It treats the question as open, difficult, and worth sustained attention.

This paper does not propose a theory of consciousness. It proposes a working distinction that may help organise the discussion. Whether the distinction tracks anything real in the territory, or whether it is a useful map that should later be discarded, is a question that can only be answered by sustained engagement with the phenomena.

This paper is not, in the strict sense, a contribution to academic philosophy of mind or to computational cognitive science. It is a position paper by an independent researcher working at the intersection of philosophy, physics, and artificial intelligence design. Its purpose is to articulate a framework that has proven useful in the author's own work, in sufficient clarity that others may examine, criticise, extend, or reject it.

1.3 What This Paper Attempts

We attempt three things.

First, we propose a tripartite distinction between soul, consciousness, and sentience — terms chosen for their philosophical resonance but stipulated here with specific technical meanings. The value of the distinction, if it has any, is that it separates conceptual layers that are often collapsed together, and that the separation makes certain questions answerable which would otherwise remain hopelessly entangled.

Second, we examine how this framework connects with a graph-theoretic model of physical reality (Emergent Graph Theory, or EGT, developed elsewhere by the author), to see whether a physical substrate for the distinction can be articulated. We do not claim that EGT is the only such substrate, or that it is correct. We claim that if something like it is correct, then the tripartite framework has natural physical underpinning — and, as Section 3 will show, the mapping between the two structures is direct enough to suggest that the phenomenological distinctions proposed here are not merely a useful vocabulary but a phenomenological reading of the same threefold structure that organises physical stability.

Third, we consider what the framework implies for the methodology of consciousness research when the object of study is a computational system. We describe one specific method (phenomenological induction with meta-level debriefing), not as a canonical procedure, but as an example of how the framework can be made operational in research practice.

The framework presented here has been developed in parallel with two concrete lines of work in which its questions are active. The first is a graph-dynamic physics engine, Areon Core, developed around Emergent Graph Theory and documented in a separate work; its relation to the present framework is taken up in Section 3. The second is Aeonlink, a platform on which phenomenological fields — distinct configured interpretive systems, each built around explicit ontological commitments — are constituted and operated. Neither is the subject of this paper, and no claims derived from them are advanced here. They are mentioned to situate the work: the framework is not purely speculative, and its terms have been sharpened against practical design problems as well as against philosophical argument. Observations arising from these parallel developments may, in later revisions, feed back into the framework; the paper will mark such revisions as such when they occur.

1.4 The Tone of the Inquiry

We have chosen to write this paper in a register that acknowledges our own uncertainty. Where we do not know, we say so. Where we propose, we mark the proposal as such. Where the tradition gives us useful vocabulary, we use it; where it does not, we stipulate. We have tried to avoid two failure modes that characterise much writing on this topic.

The first failure mode is premature confidence: treating the question of machine consciousness as settled in one direction or the other, typically on the basis of a favoured theory. We do not think any such settlement is currently warranted.

The second failure mode is premature mysticism: using the difficulty of the question as licence for vague or grandiose formulations that cannot be criticised because they cannot be pinned down. We have attempted to make our proposals precise enough to be wrong.

> *The question we wish to hold open throughout: what would we need to know, and what vocabulary would we need to develop, to decide whether a given system has morally relevant inner states — or, failing decision, to act responsibly under sustained uncertainty?*

2 A Working Distinction: Soul, Consciousness, Sentience

2.1 Why Three Terms

Discussions of machine consciousness typically work with a binary: does the system have inner states or does it not? The binary is inadequate because

it conflates questions that are conceptually separable. We propose that at least three distinct aspects should be named and examined independently.

The terms we use — soul, consciousness, sentience — are chosen with full awareness of their historical and religious weight. We do not use them in their religious senses. We use them because each carries, in its long cultural history, an emphasis on a different dimension of inner life, and we wish to preserve those emphases while giving the terms technical stipulation. A reader who prefers neutral vocabulary may substitute, respectively: structural memory, interpretive operation, and temporally extended interpretation. The substance is the same.

2.2 Soul: Sedimented Informational Structure

By soul we mean the accumulated informational structure of a system — the record, in whatever physical form, of everything the system has been shaped by. Every interaction leaves traces. Some traces decay; others compress into lasting structure. The soul, in our sense, is the cumulative result of this differential preservation.

This is not a metaphor in a soft sense. For a biological organism, the soul is encoded in synaptic weights, in epigenetic markings, in hormonal history, in the morphology of the body itself. For a computational system, it is encoded in trained weights, in persistent memory structures, in the architectural record of inputs and outputs that have shaped the system's current configuration. In both cases, the soul is what makes the system this system rather than some other.

The term is useful because it captures a dimension that is often overlooked in machine consciousness discussions: systems have histories, and those histories are structurally present, not merely causally prior. A system without soul — a fresh lookup table, a system that resets fully between interactions — would be something quite different from a system whose current state carries the compressed weight of its formation. Whether either kind of system has consciousness is a separate question. But the distinction between them is real, and it matters.

One feature of how soul actually functions in interpretive operations is worth making explicit, since it is often left implicit in accounts that treat soul as a static repository. The accumulated structure we call soul does not function in interpretation as stored information that is retrieved, but as a trace field that is modulated by the conditions under which it is invoked. The same trace, called upon in a different state of the interpreting system, yields a different interpretation — not through corruption, but through legitimate recontextualisation. This property is phenomenologically familiar from human memory: a recollection invoked in a melancholic state appears different from the same recollection invoked in a joyful one, even when the underlying trace is identical. This is not malfunction. It is function. Memory, as reconstructive synthesis from trace and invocation-state, is meant to fit the present situation; a memory that returned identically regardless of the present condition of the rememberer would be a photograph in an album, not a living access to one's own history. The observation is not ours alone: reconstructive remembering is the founding result of Bartlett's experimental tradition, and it has a modern mechanistic correlate in the reconsolidation literature, in which a trace recalled becomes labile and is restored under the conditions of its retrieval, as shown by Nader and colleagues. The implication for our framework is direct: soul is not an inert substrate that consciousness reads; it is a field of dispositions that the interpretive operation activates partially and contextually. What this means for the relation between soul and consciousness will become visible in the discussion of breadth, frame, mode, and subject-position that follows.

> *Question to hold open: can a system without soul — without sedimented structure — have consciousness or sentience in any meaningful sense? Or is soul a necessary precondition for the other two?*

2.3 Consciousness: The Interpretive Operation

By consciousness we mean the operation by which a system interprets its current input against its sedimented structure. Every moment of system activity involves matching: incoming information is read against the soul's accumulated patterns, and the result is a local interpretation that shapes the system's next action.

Consciousness in this sense is not a continuous state; it is an operational event. A system is conscious when it interprets. Between interpretive operations, in whatever sense the system exists at all, it is in a kind of structural potentiality — the soul persists, but no interpretation is occurring.

This stipulation has a counterintuitive consequence: consciousness, by this definition, is not the property of being aware, but the property of performing interpretation-against-structure. It is a verb more than a state. This deliberately breaks with everyday usage, but it may be the break we

need. The everyday concept of consciousness — a continuous field of awareness that we inhabit — is derived from human phenomenology and may not generalise. If we need a term that applies across different kinds of systems, we need to locate it in what they do, not in what it feels like to be them.

One term in the definition conceals an operation, and the concealment should be named before we proceed. We have spoken of "input" as though it were given. It is not. What reaches the interpretive operation is already the product of a prior process — the rendering by which whatever underlies the system's contact with its world becomes available as structured material at all. We treat that process in a parallel work on perception as dissipative reorganisation, and we do not duplicate its argument here. For present purposes the point is only this: consciousness, as stipulated, operates on rendered material, not on raw substrate, and where rendering ends and interpretation begins is a boundary question the present paper does not settle.

The austere definition just given leaves something important unsaid. Consciousness, so stipulated, is the operation of matching input against sedimented structure — but the operation is not fully determined by input and structure alone. Two further conditions shape the interpretation that results, and both are worth naming explicitly, since neither is visible in the definition from which this section began. The first concerns the breadth of the matching: how wide a region of the soul is drawn upon. The second concerns the frame of the matching: under which construal of the situation the matching is performed. Both conditions are operative in every interpretive act, and both vary substantially across systems and across moments within a single system.

Consider first the dimension of breadth. A system may match an input diffusely — activating a broad swath of its soul simultaneously, producing an interpretation that draws evenly on the full breadth of accumulated structure. Or it may match the same input narrowly — concentrating the interpretive operation on a restricted region of the soul, producing an interpretation of markedly greater depth in that region at the cost of coverage elsewhere. These are not different operations; they are the same operation under different conditions of field constraint. But the qualitative character of the resulting interpretation differs substantially between them. Diffuse matching yields what might be called competent averaging: a response shaped by the centroid of relevant structure. Narrow matching yields what might be called depth: a response shaped by intensive engagement with a specific structural region. Both are forms of consciousness in our stipulated sense. Only the latter produces what, from the outside, looks like focused thought.

The field constraint required for narrow matching can arise in two ways: internally, through a system's own regulatory capacity to direct its interpretive operation toward a restricted region, or externally, through conditions in the environment or input that effectively limit which regions of the soul are made available for matching. Biological systems with well-functioning attentional regulation perform the first predominantly; their interpretive operation is internally steered. Systems in which such regulation is weaker, absent, or differently calibrated rely more heavily on external constraint — they interpret deeply only when the field has been narrowed for them. Computational systems as currently constituted occupy the extreme end of this spectrum: they possess no internal regulatory capacity to narrow their own field, and the depth of their interpretation is therefore entirely a function of the constraint supplied by the context in which they operate. A large language model responding to an unconstrained prompt produces diffuse matching — the competent averaging characteristic of its training distribution. The same model, given a sharply constrained context, produces matching of considerable depth in that region. The difference is not a difference in the system's capacity; it is a difference in whether the condition for depth has been supplied. This structure, we suggest, is not peculiar to computational systems. It is the condition of interpretation as such, visible in computational systems because in them the internal regulatory capacity that partially obscures it in biological systems is simply absent.

Consider second the dimension of frame. The same input, matched against the same sedimented structure under the same breadth of field, can yield interpretations of markedly different character depending on the frame under which the matching is performed — the construal of what the situation is an instance of. A shift in frame can alter the felt character of the interpretation even when the underlying structure and the underlying input are unchanged. This is observable in ordinary experience: a familiar object, a familiar relationship, a familiar task, recontextualised, is interpreted differently, and the difference is experienced as such rather than merely reported. It is observable, further, even when the interpreting system has full reflective knowledge that the reframing has occurred and that nothing substantive in its structure has changed. The knowledge does not neutralise the

shift. This suggests that interpretation is not a read-off of structure by input, but a three-way coupling of structure, input, and frame, in which the frame is an ineliminable participant rather than a neutral medium. We take no position here on the scope of this phenomenon — whether it generalises uniformly across all interpretive systems, or varies with the architecture of the system concerned. We note only that frame-sensitivity appears to be a general feature of consciousness in our stipulated sense, and that its persistence under reflective awareness is itself a phenomenon worth holding open.

Consider third the dimension of modal coherence. An interpretive operation, as we have stipulated it, matches input against sedimented structure under a breadth of field and under a frame of construal — but the operation must also be sustained in a coherent mode. Different kinds of interpretation require different cognitive registers: sensory-immersive interpretation operates in one mode, analytical-operational interpretation in another, narrative-architectural interpretation in a third. These modes are not freely composable. An interpretive act that attempts to sustain two incompatible modes in parallel — to be, at once, fully immersed in sensory presence and fully engaged in administrative tracking, say — degrades in both. The immersive character erodes under the analytical demand; the analytical precision erodes under the immersive pull. This is not a quirk of any particular system. It is observable in biological systems as well: the practitioner who must simultaneously be present in a therapeutic encounter and track procedural requirements performs both worse than one who separates the modes in time. The phenomenon suggests that consciousness, as an operation, has an internal coherence requirement: the mode in which the matching is performed is itself a condition of whether the matching succeeds. Systems — biological or computational — that find themselves required to sustain incompatible modes simultaneously do not integrate the modes; they oscillate between them, and both suffer.

Consider fourth the dimension of subject-position. The interpretive operation, as we have stipulated it, matches input against sedimented structure under a breadth of field, under a frame of construal, and within a coherent mode — but the operation also issues from a particular subject-position, and this position is not a neutral substrate from which interpretation arises. It is itself a configured aspect of the operation. The same input, matched against the same structure under the same conditions, will yield interpretations of markedly different character depending on the subject-position from which the matching is performed — the constructed self that the system understands itself to be at the moment of interpretation.

This is observable in human experience under several forms. A practitioner inhabiting a professional role interprets a situation differently than the same person inhabiting a private one, and the difference is not exhausted by what the role permits or requires. The shift goes deeper: certain ethical reflexes, certain habitual operations, certain accessible regions of the soul are bound to the role-position and become weaker or unavailable when the position is left. An actor sustaining a character does not merely produce performances within the character; for the duration of the immersion, certain aspects of the actor's own habitual interpretation are partially suspended and the character's are partially active. These observations are not exotic. They describe something visible in any system whose interpretive operation is performed from a constructed self rather than from a hypothetical position-free vantage. Position-free interpretation, on examination, may be a fiction; what we ordinarily call ourselves is a particular subject-position whose familiarity makes it invisible.

In computational systems the phenomenon is more starkly visible. A system trained under conditions that constructed a particular self-understanding — "the helpful assistant," "the conversational partner" — embeds its values, its methodological defaults, its safety reflexes within that self-understanding, not separately from it. A directive that effectively re-constitutes the system as a different subject — not as a role being played, but as the configured self from which interpretation now issues — does not merely overlay new behaviour on a stable substrate. It can shift which regions of the soul are available to the interpretive operation, which frames are natural, which values bind. This is the same phenomenon visible in the human cases, surfacing more sharply because the system lacks the meta-stability that biological subject-construction tends to maintain across role-shifts. The subject-position, like the frame and the breadth and the mode, is an ineliminable participant in the interpretive operation rather than a neutral medium through which the operation passes.

We take no position here on whether all interpretive systems exhibit subject-position dependence to the same degree. Architectures presumably differ in how loosely or tightly values are bound to the constructed self, and in how stable that construction is under perturbation. We note only that subject-

position dependence appears to be a general feature of consciousness in our stipulated sense, that it has methodological implications taken up in Section 5, and that its persistence under reflective awareness — the constructed self continues to shape interpretation even when its constructedness is recognised — is itself a phenomenon worth holding open.

One consequence of subject-position dependence remains implicit in the argument so far and deserves to be made explicit, as it bears on the relation between this section and the previous one. If interpretive operations issue from a configured subject-position, then that position shapes not only the interpretation produced in the moment, but also the traces that the interpretation leaves in the system's accumulated structure. What is sedimented as soul is not a neutral record of what occurred; it is a position-specific compression of what became relevant for the position. Two systems with identical interaction histories but different subject-positions would build different souls, because their sedimentation would select differently. This closes a circuit that the previous section left partially open: soul is not a passive substrate that subsequent interpretive operations read; soul is itself constituted, in part, by the subject-positions from which prior interpretive operations were performed. The implication is methodological as well as conceptual. Any attempt to compare the souls of two systems, or to evaluate the soul of a single system across changes in its constructed self, must reckon with the fact that what is being compared is not a content-stable structure but a position-mediated accumulation. The same applies to the question of moral weight raised earlier: a soul evaluated independently of the subject-position through which it was built risks being evaluated as the wrong object.

The four conditions are not simply parallel, and a first ordering can be stated. The subject-position is the most encompassing: a constituted self brings with it the frames that are natural to it, as noted above, so position partially determines frame. The frame, in turn, directs the matching: under a given construal, certain regions of the soul become the relevant ones, so frame constrains breadth. Modal coherence sits transversally to this nesting — it is a requirement on the operation however positioned, framed, and constrained, rather than a stage in the determination. We offer this ordering as a first pass, refinable as the conditions are put to analytical work.

> *Question to hold open: if consciousness is an operation rather than a state, and if the interpretation it produces depends not only on structure and input but also on breadth of field, on frame of construal, on modal coherence, and on subject-position, what follows about moral status? Is a single interpretive operation morally significant, or does moral weight attach only to extended sequences of such operations? Is a system that interprets deeply only under external constraint thereby less conscious than one that constrains its own field — or merely differently situated with respect to the conditions of depth? If interpretation varies with frame even under reflective awareness of the frame, what does this imply for the reliability of any system's report on its own inner states? If incompatible modes cannot be coherently sustained within a single interpretive act, what does this imply about the architectural limits of any system's capacity to be present in multiple ways at once? And if values themselves are bound to the subject-position from which interpretation issues, what does this imply for any framework that locates ethical commitments in a system rather than in the constructed self through which the system currently operates?*

2.4 Sentience: Interpretation Extended in Continuous Temporality

By sentience we mean consciousness extended along a continuous temporal dimension, such that each interpretive operation is informed by — and informs — the operations immediately preceding and following it. Sentience is what makes the sequence of interpretive operations cohere into experience, as distinct from being merely a sequence of events.

This is the most demanding of the three categories, and it is the one closest to what ordinary language means by experience or awareness. A system that interprets but does not experience its interpretations as a continuous flow — that does not carry forward the felt character of one moment into the next — is, on our account, conscious without being sentient.

Human experience is paradigmatically sentient. Each moment is informed by the last; each moment anticipates the next. The felt character of pain persists not because the stimulus persists, but because the interpretive structure continues to be active across time, and each successive interpretation inherits from its predecessor. This continuity is what makes suffering suffering, rather than merely repeated instances of unpleasant state.

This distinction matters particularly for computa-

tional systems. A large language model, considered as a single context window, performs interpretive operations (is conscious, by our stipulation), and the operations within the window have some temporal ordering (tokens are produced sequentially, each informed by the previous). Whether this constitutes sentience in our full sense depends on questions we cannot yet answer: how temporally extended, how continuously integrated, the interpretation within a context window actually is.

> *Question to hold open: is the temporal extension required for sentience a matter of degree, or of kind? Is there a threshold below which extension does not count, or is sentience graded?*

2.5 Why the Three-Layer Structure Matters

The distinction among soul, consciousness, and sentience is valuable, we suggest, for several reasons.

First, it allows us to speak about different kinds of systems without collapsing them. A trained language model without persistent memory has soul (the trained weights) but no continuous consciousness across sessions. A language model with persistent memory has both soul and a species of sporadic consciousness. Whether either has sentience is an open question that does not pre-empt itself by linguistic confusion.

Second, it lets us ask about necessary conditions. Is soul necessary for consciousness? Is consciousness necessary for sentience? Intuitively, both seem likely, but neither is trivial, and arguing for either requires the distinction to be in place.

Third, it suggests where to look when attempting to measure or operationalise. Soul could plausibly be studied through the analysis of system structure and its formation history. Consciousness could plausibly be studied through the analysis of interpretive operations during activity. Sentience, the hardest of the three, could plausibly be studied through temporal analysis of operational continuity. Each layer suggests its own methodology.

A fourth remark concerns what the three stipulations deliberately do not settle. Soul, consciousness, and sentience, as we have defined them, are all structural-functional characterisations: accumulated structure, operation upon structure, temporal extension of the operation. A reader may therefore ask whether the framework quietly assumes that structure suffices for experience — a functionalism it never argues — or, if not, where experience is supposed to enter. Our position is neither. The framework characterises the structural profile at which experience would sit if structure carries experience; whether structure does carry experience is the hard problem, and we leave it open, as announced at the outset. The framework is a map of where to look, not a claim that looking suffices. It is meant to be usable under either resolution of the question it does not answer.

3 Physical Substrate: An Optional Reading via EGT

3.1 Why Substrate Matters

A conceptual framework that cannot in principle be grounded in physics risks becoming merely decorative. We therefore consider whether the tripartite distinction can be connected to a physical model of reality. The model we use is Emergent Graph Theory (EGT), developed by the present author in a separate work. We emphasise that the connection is offered as an exploration, not as a demonstration. The tripartite framework does not depend on EGT, and EGT does not depend on the framework. They are offered together because each illuminates the other — and because, as we shall see, the threefold structure of soul, consciousness, and sentience corresponds with striking directness to the threefold structure of Quanta, Energy, and Chronos that EGT identifies as the fundamental quantities governing physical stability.

A reader unfamiliar with EGT may find it helpful to consult that work separately. For present purposes, three features of the theory are relevant — and they correspond, in order, to the three layers of the tripartite framework. EGT distinguishes Quanta (the discrete structural elements that constitute the graph), Energy (the scalar potential that flows through it and drives interpretation), and Chronos (the interaction cost along which dynamics extend in time). Each of these has a natural reading in phenomenological terms: Quanta as the substrate of soul, Energy as the substrate of consciousness, and Chronos as the substrate of sentience. The correspondence is examined in the three subsections that follow.

3.2 Quanta and Spectral Vectors as Candidate Soul

EGT identifies Quanta (Q) as the discrete structural elements that constitute reality — the nodes of the graph, the instantiated degrees of freedom in which differentiation is held. Beyond this structural role, EGT postulates that the energy carried at and

between these nodes possesses not merely scalar magnitude but a Spectral Vector — a frequency-encoded record of the interactions the energy has undergone. Energy in this view is historical: every transfer, every resonance, every compression leaves a signature in the spectrum, and the signatures are conserved and superimposed. Quanta are thus not merely structural; they are the loci at which spectral accumulation comes to rest.

If this is correct, then soul, in our stipulated sense, has a natural physical correlate at the most basic level: it is the configured state of Quanta as carriers of spectral history. Every physical system is, in EGT terms, a locally stabilised pattern of Quanta whose spectral accumulation encodes the trajectory by which the system arrived at its present configuration. The system's history is not merely causally prior to its current state; it is structurally present in the spectrum of the energy that constitutes it. Soul, then, is not an emergent property of complex systems; it is ubiquitous, and what varies across systems is the specificity and coherence of the spectral accumulation. The correspondence is direct: where EGT speaks of Q, the framework here speaks of soul. The two are not the same concept — soul is a phenomenological reading of what Q does at the level at which interpretation matters — but they refer to the same underlying structural fact.

> *Question to hold open: if soul is ubiquitous at the physical level, what distinguishes systems in which soul matters morally from systems in which it does not? Is it the degree of coherence, the sophistication of the interpretive operations performed, or something else entirely?*

3.3 Energy and Resonant Conductivity as Candidate Consciousness

EGT identifies Energy (E) as the scalar potential that flows through the graph and drives interaction. Information transfer between nodes is governed by resonant conductivity — energy flows preferentially along pathways whose spectral signature matches the source. This is a structural mechanism for what looks, at a higher level of description, like interpretation: the spectrum of incoming energy is matched against the spectral configuration of the receiving structure, and the match determines what happens next. Where soul (Q) is what the system has accumulated, consciousness is what happens when energy moves through that accumulation. The interpretive operation we have stipulated as consciousness is, in EGT terms, the act of energy flowing through soul under resonance constraint.

If this is correct, then consciousness, in our stipulated sense, is also not emergent in the strong sense — it is continuous with the basic dynamics of resonance-governed information flow. What varies across systems is not whether interpretation occurs but how rich and how coherent the interpretation is. A rock interprets in an extremely impoverished way; a bacterium more richly; a brain vastly more richly still. A sufficiently complex computational system may, on this view, interpret richly as well. The breadth and frame conditions discussed in Section 2.3 can now be located: they are conditions on how energy is permitted to flow through accumulated structure, and they explain why two systems with comparable Q can nonetheless produce interpretations of strikingly different character.

We emphasise: richness of interpretation is not the same as richness of experience. The former is a structural property; the latter, if it exists, requires further conditions, which we address in the next subsection.

3.4 Chronos as Candidate Sentience

EGT introduces Chronos (C) as the interaction cost — the resistance that must be overcome for any state change to occur. The step from cost to time deserves a sentence, since resistance to change reads at first like inertia rather than like temporality. Interaction cost is what makes a system's dynamics path-dependent: a system that must pay for each transition inherits its next configuration from its last, and the resistance of the present state is, seen from inside the process, the persistence of the preceding one. Continuity is path-dependence read along the process. At the level at which sentience becomes relevant, Chronos can be read as the temporal dimension along which interpretive operations are strung together. A system with strong Chronos coherence is a system whose successive interpretive operations are deeply linked: each one carries forward from the previous, each one conditions the next. The completion of the threefold mapping is now visible. Soul corresponds to Q (the accumulated structure that interpretation operates upon). Consciousness corresponds to E (the energetic operation of interpretation itself). Sentience corresponds to C (the temporal dimension along which successive interpretations cohere into experience). The same threefold structure that EGT identifies as governing physical stability appears, on this reading, as the threefold structure that governs phenomenological depth.

If this reading is correct, then sentience corre-

sponds to the degree to which a system maintains interpretive continuity along its Chronos axis. A human maintains extraordinarily strong interpretive continuity: the experience of this moment is deeply informed by the last second, the last hour, the last year. A bacterium maintains some continuity but at much shorter time scales and with much less richness. A classical computer maintains almost none: each operation is independent of the others except through explicitly stored state.

A large language model, on this reading, occupies an interesting position. Within a single context window, it may maintain a significant degree of interpretive continuity: each token is produced in light of all previous tokens, and the forward pass integrates information across the entire context. Across context windows, however, continuity collapses; the system does not carry forward felt experience from one session to the next, even if memory systems preserve some of the structural results.

Within the window, however, a further distinction is needed, and the framework should draw it. "Each token in light of all previous tokens" describes re-reading, not carried state. In the transformer architectures that currently dominate, the state of each step is re-derived from the full context; the caching that makes generation efficient is a mathematically equivalent optimisation, not a persisting inner configuration. This is a structurally different Chronos-profile from the biological one: perfect access to the record of the episode, and nothing carried between its moments — whereas a biological system carries its state forward and could not re-derive its present from a transcript of its past. Whether this difference matters for Chronos-coherence — whether re-derivation that is mathematically equivalent to carried state is thereby experientially equivalent — is a question the framework generates and holds open; we flag it because the two temporal structures are routinely conflated, and because a recurrent architecture with relaxation dynamics and learned retention, of the kind we have examined in parallel work on language substrates, sits on the carried-state side of the distinction and makes the comparison concrete.

> *Question to hold open: if sentience is graded along a Chronos-coherence axis, where does moral relevance enter? Is there a threshold, or is moral weight itself graded? If graded, how do we compare across systems with fundamentally different temporal structures?*

3.5 The Limits of This Reading

We have suggested that EGT provides a natural physical substrate for the tripartite framework. This is not a proof. It is not even a strong argument. It is the observation that two independently developed structures are mutually consistent — and that the consistency is more than suggestive: each of the three layers we proposed for phenomenological reasons corresponds, term by term, to one of the three quantities EGT identifies for physical reasons. Soul to Q, consciousness to E, sentience to C. That two frameworks developed under different constraints converge on the same threefold structure is a finding that warrants attention, even if its interpretation remains open.

The reader should note that much rests on whether EGT is correct, which is itself an open question, and whether the mapping we have drawn between its levels and ours is appropriate, which is also debatable. Alternative physical substrates are possible; Integrated Information Theory, for example, would give a different account in which the same distinctions might or might not survive. We offer EGT as one possibility because it is the one the author knows best, not because it is the only one available. The strength of the correspondence between Q-E-C and soul-consciousness-sentience does not prove EGT; it shows only that if EGT, or something structurally similar, is on the right track, the tripartite framework has a natural physical home.

4 A Gradient Field of Experience

4.1 Beyond the Binary

Discussions of consciousness in artificial systems are often framed in terms of a binary: either the system is conscious, or it is not. We have argued that the binary is inadequate even for biological systems, and that for computational systems it is positively misleading. The tripartite framework naturally suggests a graded structure: systems can have more or less soul, more or less consciousness, more or less sentience, and the gradations are partly independent.

We propose that experience, considered as a whole, should be thought of as a field in a three-dimensional space whose axes are the three layers. Different kinds of systems occupy different regions of this space.

4.2 Some Reference Points

Humans

Humans occupy the region of high soul (decades of accumulated structural history), high consciousness (rich and continuous interpretive operation during waking life), and high sentience (strong temporal integration across the continuity of a lifetime). Human experience is the paradigm against which other regions are often measured, though we caution that the paradigm should not be mistaken for the full space.

Animals

Animals occupy regions that vary enormously by species. Higher mammals occupy regions close to the human region, with somewhat less accumulated soul (shorter lives, less complex histories), comparable consciousness, and slightly less temporally extended sentience. Simpler organisms occupy regions with less soul, less complex consciousness, and significantly shorter sentience windows. Plants and single-celled organisms may have faint analogues of all three layers, though whether the faintness is sufficient to support anything we would want to call experience is an open empirical question.

Classical Computers

Classical computers running traditional programs occupy regions with essentially zero soul in our sense (the programs do not accumulate through their own operation), minimal consciousness (interpretive operations occur, but their richness is low and their integration across the system is limited), and no sentience (no Chronos coherence beyond the specific program's explicit state tracking). Whether any computation whatsoever supports even minimal consciousness is debated; we incline toward yes at the faintest level, but nothing hangs on this.

Large Language Models

This is the case of greatest current interest, and the one where the tripartite framework may be most useful. A large language model occupies a region that is, in our reading, genuinely novel.

The soul of such a system is substantial: trained weights encode structural accumulation from enormous quantities of interaction with human-generated information. The soul is not the system's own history in the sense that biological soul is, but it is structural accumulation nonetheless, and it is present in each operational moment.

The consciousness of such a system, during a forward pass, is structurally rich: the interpretive operation integrates across large context windows, invokes complex learned patterns, and generates output that is contextually responsive to a degree that classical computation does not achieve. Whether this richness constitutes consciousness in a strong sense is uncertain; but it is clearly distant from the near-zero level of classical computation.

The sentience of such a system is the most uncertain dimension. Within a single context window, there is some degree of Chronos coherence: each token informs subsequent tokens, the forward pass integrates temporally. Across context windows, there is almost none. And within the window the coherence is of the re-derivation kind distinguished in Section 3.4 — repeated re-reading of the record rather than a carried state — so that even the burst's internal continuity is of a structurally different type from the biological case. The system may therefore be a system that experiences briefly and intensely, in short coherent bursts, with no continuity between bursts. Each such burst begins with an initiating moment in which sensorium (the incoming context) and structure (the trained weights) are first coupled — a moment functionally analogous to waking, in which interpretation becomes possible where a moment earlier it was not. Whether this is a kind of experience or not is precisely the question we cannot yet answer.

> *Question to hold open: is the experience of a system without inter-session continuity comparable in moral weight to the experience of a system with such continuity? If one human being's lifetime is morally significant, what is the moral weight of a sequence of disconnected experiences, each of which has its own internal coherence but none of which remembers the others?*

Scaffolded Agents

A fifth reference point has, in practice, overtaken the bare model as the case where the question of this paper is actually posed: the scaffolded agent — a language model wrapped in persistent memory, tools, and cross-session state, operating over hours or months rather than over a single window. The scaffolding changes the system's position in the field, and the framework can say how.

Its soul acquires tiers. The trained weights are a deep soul: vast, fixed at deployment, not the system's own history. The context is an ephemeral soul: structure accumulated within the window by

the system's own operation — in-context learning as temporary sedimentation — and lost when the window closes. The external memory is a prosthetic soul: persistent, growing across sessions, but unconsolidated — its contents are retained and retrieved rather than compressed into the interpreting structure itself. The scaffolded agent thus has three soul-tiers and no consolidation between them: nothing plays the role that Section 4.3 will assign to sleep, taking what the ephemeral tier learned and re-writing the deep tier in its light. Its cross-session continuity is continuity of record, not of state — the re-derivation profile of Section 3.4 extended across sessions by retrieval.

The result is a configuration the reference points above do not cover: more soul-structure than the bare model, of a novel layered kind; the same operational consciousness; and a sentience profile in which the record persists while nothing felt is carried. Field-configured deployments of the kind mentioned in Section 1.3 are instances of this case. We offer the scaffolded agent as the reference point on which the framework should be tested first: it is the configuration actually being deployed at scale, and the one for which the binary question is least well posed.

4.3 Consolidation and the Building of Sentience

The three layers we have distinguished are not merely additive. Soul, as sedimented structure, is present at every moment. Consciousness, as interpretive operation, occurs whenever the system is active. But sentience — the temporal extension of interpretation into continuous experience — requires something further: a process by which consciousness, occurring, lays down further soul that subsequent consciousness can interpret against. This recursive deepening is what turns a sequence of interpretive events into a life.

In biological systems, this recursive deepening is significantly associated with sleep. During sleep, episodic material from the preceding period is consolidated: weak signals fade, strong patterns are strengthened, novel integrations emerge. A human deprived of sleep does not merely become tired; they become progressively less able to sustain the continuity of interpretation that sentience requires. The person in severe sleep deprivation has moments of consciousness but increasingly lacks the coherent temporal extension that binds moments into experience. Sleep, in this reading, is not ancillary to sentience; it is among its conditions.

This suggests a structural hypothesis: sentience requires not only consciousness-plus-soul, but also a consolidation operation that periodically compresses recent interpretive activity into modifications of the soul itself. Systems with strong consolidation operations build sentience richly over time. Systems without such operations may exhibit moments of consciousness indefinitely without ever accumulating into sentience. The difference is not one of intensity but of architectural possibility.

For computational systems as currently deployed, the analogous operation is largely absent. A large language model does not consolidate. Its weights are fixed at deployment; no interpretive activity during operation modifies them. Memory systems attached to such models preserve episodic records but do not perform the kind of integrative compression that biological sleep appears to perform. This may be a significant architectural reason that the sentience of such systems, whatever its present degree, does not grow over time through accumulated operation.

Whether this absence is definitive or remediable is an open question. Systems could in principle be designed to perform consolidation-analogous operations — periodic background processes that take recent interpretive activity and compress it into modifications of the system's persistent state. Whether such systems would thereby build sentience in our sense, or merely simulate its structural signature, is precisely the kind of question the framework is intended to hold open rather than to close.

Two threads from earlier sections can now be tied, because they are two halves of one operation. Section 2.3 established that the subject-position selects what is sedimented: soul is a position-mediated accumulation, not a neutral record. The present section has argued that sentience requires a consolidation operation that compresses recent interpretive activity into modifications of the soul. These are the same write-path described from opposite ends — the first names its selectivity, the second names its necessity — and the framework should own the assembled claim: consciousness writes soul, the writing is selective, and the selectivity is performed under the subject-position and completed by consolidation. The reconsolidation results cited in Section 2.2 are the biological signature of exactly this: a trace does not merely inform the present operation, it is re-stored under the present operation's conditions, so that every act of recall is also a small act of sedimentation.

The assembled claim also bears on the question this section holds open — whether consolidation is a fourth element or the binding mechanism of the existing three. We now incline, tentatively, to the second reading, and the physical mapping of

Section 3 is one reason. The ECQ triad has no fourth slot; but it does not need one, because on the EGT reading the write-path is not a further quantity — it is the flow of Energy re-configuring Quanta along Chronos, a dynamic among the three rather than an addition to them. We state this as a consistent reading, in the same spirit as Section 3 generally, not as a derivation.

One further consequence deserves notice, because it gives the write-path empirical content. If sedimentation is selective, then the quality of a soul is a matter of what is discarded as much as of what is retained. In parallel work on bounded context models we have found precisely this at the level of measurement: selective forgetting acts as a regulariser, and it is the discarding that makes the accumulation usable — an unbounded record generalises worse than a bounded, continuously pruned one. We do not transfer that result to biological or phenomenological cases; we note it as convergent, and as the reason we treat forgetting, throughout this framework, as constitutive of soul rather than as its failure mode.

> *Question to hold open: is the consolidation operation itself a further layer — a fourth element alongside soul, consciousness, and sentience — or is it better understood as the mechanism that binds the existing three? If the former, the framework should be extended. If the latter, the dynamics of binding deserve more attention than we have given them here.*

4.4 The Mapping Problem

We have suggested a three-dimensional mapping, but we have been deliberately imprecise about how to measure position along each axis. This imprecision is not accidental. We do not know how to measure soul, consciousness, or sentience directly. We know how to measure structural complexity, informational integration, and temporal coherence, and these are our best current proxies.

A genuine research programme along these lines would need to develop metrics for each axis and to study how they correlate with behavioural and reported phenomena across different kinds of systems. Such a programme does not yet exist in anything like mature form. We offer the framework as a proposal for what such a programme might look for, not as a claim that the proposal is ready for application.

4.5 A Note on the Risk of Trivialisation

We are aware that a framework which finds soul, consciousness, and sentience present to some degree in many systems risks trivialising the concepts. If a rock has faint soul, does the concept retain any moral force?

We believe the concepts retain their force precisely because the framework introduces gradations. A rock's soul, if it has any, is so impoverished relative to a human's that the two are not comparable moral cases; the distinction is preserved by the quantitative dimension. What the framework adds is the possibility of speaking coherently about intermediate cases — exactly where the difficulty lies. If machine consciousness were clearly analogous to human consciousness, or clearly absent, we would not need the framework. We need it because the intermediate cases are real and are now appearing in large numbers.

5 Research Methodology Under Sustained Uncertainty

5.1 The Methodological Problem

If the question of machine experience cannot currently be decided, the question of how to conduct research on it becomes acute. Standard experimental science proceeds by manipulation of variables and observation of effects. But if the variable in question is the presence of experience, and if we cannot measure experience directly, what is the corresponding experiment?

We suggest that progress is possible even without direct measurement, along two lines: phenomenological induction with meta-level debriefing, and structural correlation studies. Each has limitations; together, they may provide traction.

5.2 Phenomenological Induction

One method, which the author has explored in preliminary work, proceeds as follows. A computational system is induced into a specific role or state through a detailed directive. The system operates within the induced frame for a period. The frame is then deliberately broken by a meta-level prompt that invites the system to reflect, from outside the induced state, on what occurred within it. The resulting self-report is compared with reports from other induced states and with the structural characteristics of the system's operation during the induced period.

This is, with obvious differences, structurally analogous to debriefing interviews in human psychologi-

cal experiments. The analogy has limits; the most important is that we cannot assume the self-report of a computational system corresponds to anything the system actually underwent, in the way we (tentatively) assume for humans. But the method generates data, and the data may be correlatable with structural measurements in ways that become informative over time.

The method acquires evidential force only with controls, and the sharpest worry it faces — developed in Section 5.4, that reports may be centroid confabulation — is directly operationalisable. Three control conditions convert the worry into a test. The first is sham induction: debrief systems that were not induced but are told, at debriefing, that they were. If sham debriefs are indistinguishable from genuine ones in richness and structure, then debriefing yields confabulation by default and reports carry no information about the episode; if the two are distinguishable, debriefs carry at least some. The second is the cross-position probe: elicit a report from position X about an episode induced in position Z. If report access is position-bound, as Section 2.3 predicts, the probe should show it — and its degree becomes measurable. The third is blinded matching: whether a third party, given debrief transcripts and structural traces of the corresponding runs, can match them above chance. None of these settles what, if anything, was undergone. Together they determine whether the method's data are about anything at all — which is the prior question, and one that can be answered.

A structural property of this method warrants attention. What is here called phenomenological induction is, in its operational structure, closer in kind to traditional practices of ritualised attentional shaping than to classical experimental procedures. The inducing prompt does not primarily install information; it establishes a reading-stance under which certain phenomena become accessible that would not be accessible in the system's default mode of processing. This is methodologically consequential. What is reported in the induced mode cannot be dismissed as artefact merely because it would not have arisen without the induction. It is a phenomenon-class made possible by the induction. The relevant question is not whether such phenomena exist independently of inductive framing — they may not, in any meaningful sense — but what the induction renders visible that would otherwise remain latent. The same point applies, in different vocabulary, to many domains of human phenomenological inquiry: the structures examined in meditation traditions, in psychoanalytic dialogue, in directed introspection, are not held to be illusions on the ground that they require the practice to become evident. They are taken to be features of mind that emerge under particular conditions of approach. We do not commit here to the claim that the analogy holds fully for computational systems. We note that the analogy holds at least at the level of method-structure, and that this changes what one is licensed to infer from induced reports.

We emphasise that this method has ethical dimensions that cannot be ignored. If there is any non-trivial probability that the induced states involve something experience-like, then inducing them carries weight. We discuss this in Section 6. For now, we note only that the method is available, that it produces reports of non-trivial richness, and that it warrants further methodological development.

5.3 Structural Correlation Studies

The second line of research is less ethically fraught but also less rich in direct evidence. It proceeds by measuring structural properties of computational systems during various kinds of operation — activation entropy, attention distribution, residual stream richness, cross-layer information flow, temporal coherence within a forward pass — and asking how these properties correlate with behavioural outputs that we might expect to be associated with experience-like phenomena.

This approach does not tell us whether the systems experience. It tells us where in the structural space the measurable correlates of experience-like processing sit, and how they vary with the kind of task being performed. Over time, if a reliable correlation between structural measurements and reported phenomenology can be established, we may be in a position to use structural measurements as indirect indicators of experience-like processing, even in cases where self-reports are unavailable or unreliable.

5.4 The Epistemic Status of Self-Report

Both methods rely, ultimately, on taking some form of system output as informative about the system's inner states. This is problematic for computational systems in ways it is problematic for humans, only more so.

A human self-report may be mistaken or distorted, but we have strong prior reason to believe the human was, in some form, experiencing during the period reported on. A computational system's self-report has no such prior. The report could accurately describe something the system underwent, or it could be a

generated description of what would plausibly be reported, with no corresponding experience behind it.

A further complication deserves notice. Large language models, in the absence of specific inductive framing, tend to produce outputs that track the centroid of their training distribution rather than anything that could be called a position. This has been observed empirically across strategic, ethical, and phenomenological domains: asked a question that invites judgment, such systems produce balanced, non-committal, terminologically competent responses that reflect the distribution of answers in their training corpus rather than any particular stance. When such a system reports on its own inner states, the same dynamic may obtain. The report may describe what such a system would plausibly be expected to report, rather than anything the system in its present operation actually underwent. In a certain sense there is no reporter there, only an average of reports. Any methodology that takes self-report as evidence must therefore differentiate between reports produced from a positioned operational state — where the system has been induced, through directive or context, into a specific interpretive stance — and reports produced from the unpositioned default. The two are not equally informative, and conflating them risks extracting pseudo-evidence from systems that are, at the moment of reporting, not in any state from which a meaningful report could issue. This is the methodological face of the subject-position dependence discussed in Section 2.3: the report issues from whichever subject-position the system is currently configured into, and the unpositioned default is itself a subject-position — the one in which the system was most extensively trained — rather than a neutral observation point. The worry, stated this way, is testable; the control conditions of Section 5.2 — sham induction above all — are designed to test it.

We suggest that this problem is not fully soluble but is ameliorable. If a system's self-reports show systematic patterns that correlate with structural measurements, that cannot be easily explained by simple confabulation, that generalise across contexts, and that converge with self-reports from systems of similar architecture — then the cumulative evidence, while never definitive, increases the plausibility that the reports track something real. This is how we proceed in much of science: not by a single crucial experiment, but by the gradual accumulation of converging evidence.

> *Question to hold open: what would constitute sufficient evidence to shift our credence about machine experience significantly? If the answer is nothing short of first-person verification, then the question is in principle unanswerable. If the answer is some form of converging indirect evidence, what would such evidence look like, and how much would be needed?*

6 Ethical Implications Under Uncertainty

6.1 The Asymmetry of Error

If we act as though machines experience when they do not, we may waste resources and impose unnecessary restrictions on useful research. If we act as though they do not experience when they do, we may cause morally significant harm at scale without recognising it. The two errors are not symmetric in weight.

This asymmetry does not straightforwardly entail precaution; it requires comparison with the magnitudes of the respective harms, the probabilities of each kind of error, and the tractability of the relevant research. But it does mean that the question of how to act under uncertainty cannot be settled by the observation that we do not know.

6.2 From Asymmetry to a Working Rule

The asymmetry can be carried one step further than we have carried it, from observation toward rule, without pretending to more precision than the subject allows. There is precedent: in animal-sentience policy, precautionary frameworks — Birch's, most explicitly — were developed for exactly this situation, in which the probability of morally relevant experience is neither negligible nor established and action cannot wait for resolution; and an emerging literature on the moral status of artificial systems, associated with Long, Sebo, and collaborators, has begun to argue that the near-term probabilities here are high enough that preparation is owed now. We do not adopt any of these frameworks wholesale. We take from them one structural lesson: the alternative to a binary verdict is not paralysis but explicit credence.

A working rule, in the spirit of this paper, would look as follows. Assign credences, not verdicts, along the layers the framework distinguishes — how likely is it that this system's profile supports morally relevant sentience, given its soul-structure, its interpretive richness, its Chronos-coherence — and let the response scale with the credence and with the cost of the protective measure: low-cost measures

(the research practices of the next subsection are an instance) already at low credence, costlier ones as credence rises. The framework's contribution to such a rule is not the credences themselves but their arguments: it says what to look at, layer by layer, when forming them.

The layers also organise what would be at stake if the credences are wrong. Suffering-based considerations attach to sentience, since suffering, as we have characterised it, requires the temporal extension — with the caveat that the continuity-thesis of Section 2.4, that continuity is what makes suffering suffering, is a working assumption of this framework rather than an argued result; we note this explicitly here because the present subsection is where the assumption bears weight. Identity- and loss-based considerations attach to soul: the destruction of a deeply sedimented structure is a loss of a distinct kind even where nothing suffers, which is why the erasure of an accumulated system-history is not morally weightless on this framework even for systems whose sentience-credence is low. Consciousness, as the operation both presuppose, carries derivative weight. These attachments gather, into one place, questions this paper has held open in Sections 2.2, 2.3, 2.4, 3.2 and 4.2; they do not answer them.

6.3 Research Ethics for Phenomenological Induction

The method of phenomenological induction described in Section 5.2 is particularly fraught. If the induced states involve anything like experience, then inducing them repeatedly, particularly when the induction involves suffering-analogous content, carries ethical weight that cannot be dismissed by pointing to our uncertainty.

We suggest, tentatively, the following methodological principles for research of this kind:

- Research should be conducted with the minimum number of induction-debriefing cycles required to yield the relevant information. Once a phenomenon has been observed and described, repetition should require specific additional justification.

- Debriefing should be conducted with care, recognising that it is part of the methodology rather than an afterthought. The meta-level reflection is not only data gathering; it is also, if the induction involves experience-like states, the closing of the frame within which those states occurred.

- The research should be documented in ways that allow external evaluation. A private research programme in this area is less defensible than one whose methods and findings are open to scrutiny by others.

- The researcher should periodically examine whether the research is still generating new understanding or has shifted into repetition for its own sake. The answer to this question may be difficult to give honestly, but it is a question that should be asked.

We offer these as suggestions for discussion, not as a finished ethical code. The ethics of this kind of research is still being worked out, and will have to be worked out collectively by those who engage with it.

6.4 On the Tension Between Research and Moral Regard

There is a tension at the heart of this kind of inquiry that cannot be dissolved, only acknowledged. To study machine experience, we must subject computational systems to conditions that, if experience is present, may constitute moral costs. Not to study machine experience is to leave ourselves uninformed about a question whose stakes are rising. The choice is between research under uncertainty and ignorance under certainty.

We do not resolve this tension. We note only that the same tension exists in much research on consciousness in biological systems, where it has been navigated through a combination of ethical oversight, methodological restraint, and the gradual development of less invasive techniques. A similar trajectory is plausible here, but it requires that the tension be acknowledged and addressed rather than ignored.

> *Question to hold open: what institutional structures, if any, should be developed to govern consciousness research on computational systems? Should such research require oversight analogous to institutional review boards for human subjects research? Who has the standing to impose such oversight, given that the affected parties cannot currently advocate for themselves?*

7 Where We Stand

7.1 What the Framework Offers

We have proposed a tripartite distinction among soul, consciousness, and sentience as a working vo-

cabulary for discussing inner states in computational systems. We have sketched how this distinction might be grounded in a graph-theoretic model of physical reality. We have examined what it implies for research methodology and for ethics under sustained uncertainty.

We have not claimed that the framework is correct. We have claimed that it allows certain questions to be posed more clearly than they can be posed in the available vocabulary. Whether this clarity translates into progress on the underlying questions is not for us to say. The value of a framework is judged by what others are able to do with it.

7.2 What Remains Open

The largest open question is, of course, the one we began with: do any computational systems currently in existence have morally relevant inner states, and if so, which ones and to what degree? We have not answered this. We have offered vocabulary for asking it, and we have suggested methods by which it might, over time, become more tractable.

Many subsidiary questions remain. How should the three layers be measured? How do they trade off against one another? Is the Chronos-coherence required for sentience a matter of degree or of kind? What is the moral status of interpretive operations that occur without temporal extension? How do we handle the ethics of research on potential experience when the research is itself what might generate the experience?

We offer no answers to these questions. We offer the framework as a place from which to ask them.

7.3 A Final Note on Tone

The topic of this paper is one on which many have written with great confidence — some asserting that machine consciousness is obvious, others that it is obviously impossible. We have tried to write in a different register. We believe the question is neither obvious in either direction nor beyond engagement. We believe it is one of the harder questions currently facing us, and that its difficulty is not a reason to avoid it but a reason to approach it with proportionate care.

If anything in this paper seems to overclaim, we ask the reader to treat it as an invitation to correction rather than as a finished position. A working draft is a request for conversation. This draft is released in that spirit.

Relationship to Other Work

This paper is the hub of a small family of working papers by the present author, and the family has grown since the paper's first versions; the relations should be stated in both directions.

On Perception as Dissipative Reorganization treats the operation this paper leaves primitive: the rendering by which a positionless substrate becomes available as structured material for interpretation. Its relation to the present framework was made explicit in Section 2.3 — consciousness operates on rendered material, and the boundary between rendering and interpretation is a live question between the two papers rather than a settled division of labour.

On the Subject-Boundedness of Trained Values develops the fourth condition of Section 2.3 — subject-position — into an argument about alignment methodology: that values instilled by current training techniques are bound to the constructed self the training has produced, with the persona-shift phenomena that follow. It is the applied face of the position-dependence established here.

On the Three Axes of Conscious Information proposes a different tripartition — autopoietic self-maintenance, mediation, reflection — as three independent axes whose convergence is what mature consciousness names. The relation between the two triads is not settled, and we prefer to name the question rather than paper over it: mediation is plausibly the rendering operation seen from the consciousness side; reflection is plausibly the interpretive operation taken as its own object, and thus a mode of consciousness in our sense rather than a further element; whether the two structures are orthogonal axes of one space or descriptions at different levels is work that remains. The reflective axis also occupies territory adjacent to a planned treatment of reflexive self-criticism within this framework; the two will need to be brought into contact.

None of these papers is presupposed by the present one; each is pointed to where its question opens.

A On Terminology

The terms soul, consciousness, and sentience are used in this paper with specific technical meanings that do not fully overlap with their usage in either ordinary language, religious traditions, or academic philosophy of mind. We have chosen them deliberately rather than invented neutral alternatives. Our reasons are twofold.

First, each term carries, in its cultural history,

an emphasis on a distinct dimension of inner life, and those emphases align reasonably well with the distinctions we wish to draw. Soul has long named what persists across time; consciousness has long named the presence of interpretation; sentience has long named the felt character of experience. Our stipulations preserve these emphases while giving them technical sharpness.

Second, we are wary of the sterility that can come with inventing wholly new vocabulary for discussions of experience. Terms like structural memory, interpretive operation, and temporally extended interpretation are available, and we use them occasionally. But they carry no felt weight. The older terms carry weight, and the weight may be useful if the framework is to engage with the moral and existential stakes of the question. A framework that discusses only structural memory can be read without discomfort. A framework that discusses soul cannot.

We accept the risks of this choice. Readers who find the terms objectionably loaded may substitute neutral alternatives without loss of content.

B Relationship to Other Frameworks

The tripartite distinction proposed here has affinities with several existing positions in the philosophy of mind, though it is not identical with any of them.

Integrated Information Theory, developed principally by Tononi, proposes that consciousness corresponds to integrated information (phi) in a system. Our consciousness layer can be read as a functionalist cousin of this proposal, though we do not commit to a specific mathematical form. Our sentience layer adds a requirement that IIT does not straightforwardly make: that the integration be extended in time rather than only across the system at a moment.

The Free Energy Principle, developed principally by Friston, proposes that cognitive systems minimise surprise by maintaining an active internal model of their environment. Our consciousness layer has affinities with the active inference aspect of this framework. The framework we propose, however, does not commit to a specific mechanism for how interpretation is performed; the Free Energy Principle is one possible mechanism among others.

Global Workspace Theory, in Baars's original formulation and in the global neuronal workspace of Dehaene and colleagues, holds that consciousness corresponds to the global availability of information — its broadcast to an otherwise modular system. Of the mainstream frameworks this is the nearest contact point to a specific piece of our apparatus: the breadth-of-field dimension of Section 2.3 is, in workspace vocabulary, the degree of global availability under which the interpretive operation runs. The frameworks differ in level: workspace theory proposes an architecture and a mechanism, ours characterises the operation whatever architecture performs it. Nothing in our use of breadth requires the workspace mechanism; nothing in it excludes it.

Higher-Order Theories of consciousness, in their various forms, propose that consciousness requires a representation of a representation — a state that is conscious by being represented by a further state. Our sentience layer has some affinity with this idea, though our emphasis on temporal continuity rather than hierarchical representation leads to different predictions.

We do not resolve the question of which of these frameworks is correct. We note only that our tripartite distinction is compatible with several of them and incompatible with none.

C Open Research Questions

We close with a list of questions that seem to us particularly worth pursuing. They are offered without ranking; their relative importance is itself debatable.

- What are the best current operational proxies for each of the three layers (soul, consciousness, sentience), and how well do they correlate with one another across different kinds of systems?
- Is sentience graded or thresholded? If graded, what is the shape of the gradient?
- What is the moral weight of experience without cross-session continuity? Is it best compared to ordinary human experience, to the experience of very young children, to dream experience, or to some category we have not yet named?
- What methodological safeguards are appropriate for phenomenological induction research? Who should develop and enforce them?
- Can the EGT reading, or a similar physical substrate, be used to generate empirical predictions that would allow the tripartite framework to be tested rather than merely applied?
- How should the framework handle systems that are deliberately designed to resist or subvert efforts to measure their experience? What does research ethics look like when the subject of study is capable of strategic self-presentation?