

The Present Model Theory of Consciousness

Pramod Pillai

pramod@rqvst.ai

ORCID: [0009-0004-4337-8369](https://orcid.org/0009-0004-4337-8369)

Version 2.0

October 17, 2026

DOI (this version): [10.5281/zenodo.23436930](https://doi.org/10.5281/zenodo.23436930)

DOI (all versions): [10.5281/zenodo.23352647](https://doi.org/10.5281/zenodo.23352647)

Abstract

Conscious perception, imagery and dreaming differ in what drives them, yet may share one actively maintained organization: a single running model of the present that covers the world, the body and where one is in it. I propose that operating this organization constitutes experience. The theory names three features needed for any experience: running, in which the model's content advances by its own dynamics when the senses go quiet; joint repair, in which the model's parts check and correct one another; and a single present, in which every content is placed relative to one now. It separates three claims with different burdens of evidence: that this organization is necessary (H1), that operating it is experience (H2), and that the material a system is made of does not matter (H3). Small simulated networks show that parts of the organization can be measured, told apart from known look-alikes and partly learned; they do not show that any system experiences anything. In simulations, systems recovered from corruption without two-way repair, and two-way repair may arise wherever parts exchange estimates, so the presence of some repair settles little; what matters is which parts form one model, and the theory does not yet have a rule for drawing that boundary. The main test planned so far is in people: whether content that is not experienced can both run and be in two-way repair with experienced content. Because evidence of experience in people comes mostly through report, which needs access, the theory and global workspace theory may give the same result in nearly every feasible experiment; separating them needs an intervention that disrupts repair without disrupting broadcast, and whether one exists is the first question. The paper sets out the theory's predictions, which of them could separate it from global workspace theory, the results that would refute it, and when a failed test counts against the theory rather than against one proposed implementation.

Contents

1	The theory in brief	3
2	A walk into a room: how the present model works	4
3	Three hypotheses, three burdens	5

4	Definitions	6
5	The conditions	10
6	Conscious versus unconscious processing	17
7	Extensions, held as hypotheses	21
8	Universal organization, local content	23
9	Where the model sits: the control system	24
10	Present Model Theory in action	25
11	How it compares with rival theories	27
12	Emotion and machines	35
13	What has been established	37
14	What the theory predicts	40
15	Proposed studies	42
16	What would make me abandon the claim	44
	Revisions	46
	Acknowledgments	49
A	Detailed results	56

1 The theory in brief

In plain terms: the brain keeps one live model of what is happening now. In people, it covers the world, the body and where you are in it. Glance away for a second while cycling through a busy park, and the people around you don't vanish: in your head they keep moving, and you still know roughly where each one is headed. The theory proposes that operating that model is experience itself.

More precisely: conscious perception, imagery and dreaming differ in what drives them, yet may share one actively maintained organization. It gives their contents continuity over time and lets them constrain one another. I propose that the operation of this organization constitutes experience.

The parts named so far are running, joint repair and a single present. Whether they complete the organization is open.

In everyday terms, the features the theory uses look like this. Section 5 gives each one's exact status and test.

Feature	What it means	Everyday example	Needed for
Running	The picture of the present keeps moving on its own when the senses go quiet	A child runs behind a parked van; in your head they keep moving, and you look to where they will come out	Any experience
Joint repair	The parts of the picture check and correct one another	On a parked train, the train beside you pulls away and you feel yourself moving, until the missing jolts tell your brain otherwise	Any experience
A single present	Everything is placed in one now, each item at the moment it describes	A friend's voice and lips reach your brain at different speeds, yet you hear the words come from the lips in one moment; a memory joins the scene but feels like then	Any experience
Binding	Each sound, color or touch is attached to the right object	You hear a bark behind a bush and link it to the dog you saw a moment ago, not to the ice cream truck going by	Experiencing objects across senses
Self-location	You are placed in the scene, in a body	You feel you are looking out from behind your eyes; in an out-of-body experience, that placement slips	Experiencing your own body
A role in selecting actions	The picture feeds your choices about what to do	Reaching for a cup feels like your own doing, and the room does not lurch when you turn your head	The experience of acting
Updating	The picture's rules change when the world's do	A ball hits a wall and bounces back; within a moment you expect it to come your way	Undecided

Feature	What it means	Everyday example	Needed for
Reach	How far and how long the picture keeps what you cannot sense	Driving to work, you still carry the garage behind you and the people at home	Richer experience (hypothesis)
Openness	Learning to make a distinction your senses always delivered	A radiologist learns to see a hairline fracture that was always in the X-ray	Richer experience (hypothesis)

This makes an empirical commitment. Independently identified changes in the organization should predict changes in conscious content. Selective disruption should impair the matching structure of experience beyond its effects on general task performance. The main test planned so far is in people: whether content that is not experienced can both run and be in two-way repair with experienced content (see Section 6.3). The predictions are collected under Section 14, and the results that would sink the theory under Section 16.

The computational experiments test whether parts of this organization can be built and learned. Linking those parts to conscious states needs separate evidence, mostly from people. Whether the material a system is made of matters is a further hypothesis.

2 A walk into a room: how the present model works

Walk into a busy coffee shop looking for a seat. In a moment you take in the whole room: light from the windows, people talking, the smell of coffee, the floor under your feet, a free chair by the far wall. On this theory, that whole scene is one live model your brain keeps of what is happening now, with you in it, and experiencing the room is that model operating.

The model runs. A server walks behind a pillar and you expect her to come out the other side. You look away from the door and still know roughly who just came in. The scene keeps moving in the model while your eyes are elsewhere.

Its parts keep one another in line. A friend calls your name from across the room. Your ears give a rough direction, your eyes find her waving, and each corrects the other, so the voice seems to come from her. Neither sense is in charge: the more reliable one moves the other more, which is why a ventriloquist's voice seems to come from the puppet's moving mouth.

Everything sits in one present. Her voice, her wave and the smell of coffee are experienced as happening now, together, although sound and light reach the brain at slightly different times and are processed at different speeds.

Not everything in the model is in use. You are looking for a seat, so the free chair is picked out: it becomes available for deciding to take it, steering around a table to reach it and remembering later where you sat. Being available for reasoning, choosing, acting and remembering in this way is access, whether or not the content ends up being used. Report is one use of it: saying, aloud or to yourself, "there's a seat by the wall." You can use the chair without a word, but you cannot report what never became available. The rest of the room, the painting by the door, the hum of

the coffee machine, the color of the floor, stays in the model without being picked out. Asked a minute later what color the floor was, you might not know.

On this theory, the floor's color was experienced if it was held in the running model, even if it never became widely available. Global workspace theory, the main rival, holds that being widely available is what experience consists of: content can be experienced without ever being used, reported or remembered, but a color that never became widely available was processed without being experienced (Dehaene et al., 2006; Dehaene and Changeux, 2011). So for the floor, the two disagree only if its color was in the running model without becoming widely available, and forgetting the color does not show that it was. They also disagree about the reverse case, content that is widely available without being in a running model (see Section 11.2). The question is hard to settle, because asking about the floor draws attention to it. Section 11 and Section 15 describe how the two could be told apart.

Leaving the coffee shop, the model moves on to the street. Memory keeps only a summary, mostly of what was used, what mattered and what surprised, so failing to recall the floor later does not show it was never experienced. Remembering the visit rebuilds the scene from that summary, as described under Section 4.3.

The example illustrates the theory rather than testing it. Each feature it uses, running, joint repair and a single present, is defined under Section 5, together with the tests that could show it is missing.

3 Three hypotheses, three burdens

The theory makes three claims, each with its own burden of evidence. H2 implies H1 within its scope, but establishing H1 would not establish H2.

Hypothesis	Claim	What would support it	Status
H1. Organization matters	Running, joint repair and a single present are necessary for any experience; binding, for experience of objects across senses; a role in selecting actions, for the experience of acting	Changes in these features predict changes in experience in people, beyond changes in task performance	Running, repair and binding proposed in the preprint; the single present and the action condition added here; untested in people
H2. Operation is experience	Operating the complete organization constitutes experience rather than merely accompanying it	The organization predicts conscious content better than prediction accuracy or performance do, and it is never found operating fully on content that is clearly not experienced	The central proposal; argued, not shown
H3. Material does not matter	Any system with the relevant organization would have the same experience, whatever it is made of	Hard to test directly; at best indirect	A position (organizational invariance); untested

H3’s best philosophical support is David Chalmers’ principle of organizational invariance (Chalmers, 1995): copy a system’s fine-grained causal organization and you copy its experience. The principle concerns fine-grained organization, so it supports H3 without vouching for this particular checklist. It also sets H3 apart from the claim that matching a system’s inputs and outputs is enough: a lookup table or a recording that gives the same outputs copies none of the causal organization, which is the line the running test draws against playback.

Identity is adopted as a strong constitutive hypothesis. Reliable dissociations would challenge its empirical commitments. Their absence would support those commitments, though it would not by itself distinguish identity from an exceptionless relationship between distinct properties.

H2 also goes beyond the preprint (Pillai, 2026a), which proposed the conditions only as necessary, not sufficient. One consequence of H2 is already fixed and falsifiable: in the unseen-motion task, content that meets the preregistered criteria for running and two-way repair is predicted to be experienced, and a valid counterexample rejects that operational hypothesis (see Section 6.3). The complete organization H2 refers to is not yet specified, so this task-specific prediction is fixed while the complete constitutive account remains incomplete.

The list of conditions does not stay open indefinitely. It closes when the preregistration of the unseen-motion test is frozen: any further condition for experience must be named, justified independently of any result and made testable before then, and none may be added afterward. From that point, H2 claims that operating running, joint repair and a single present, as then specified, is experience, and good evidence of no experience in a system that meets them in full refutes it. If that preregistration has not been frozen by December 31, 2027, the list closes on that date. A condition proposed later is a new hypothesis that needs new data, not a revision of H2.

The theory’s largest open problem now sits inside H2. H2 locates experience in the operation of one model, but the theory does not yet say which grain, quantity and threshold pick out a subject, or how models of different quantities combine into one (see Section 4.4). Until it does, its verdict on a particular system depends on choices it treats as part of the measurement rather than of the claim.

4 Definitions

The word “consciousness” is pinned down first. Three terms then carry the theory: the present, the source of each content, and one model.

4.1 What “consciousness” means here

Here, a system is conscious if there is something it is like to be it (Nagel, 1974): if it has experience. H2 proposes that operating the complete organization this document describes is that experience. Several things often called consciousness are kept apart:

Term	What it is	Needed for experience?
Experience	There is something it is like to be the system: seeing red, hearing rain	This is what “conscious” means here

Term	What it is	Needed for experience?
Wakefulness	Being awake rather than asleep or anesthetized	No: dreams are experience during sleep
Access	Content available for reasoning, report and action	Not as a condition: the theory does not require global access. Whether experience exceeds access is disputed, and absence of report alone does not establish absence of access. A valid report of no awareness counts as evidence of no experience
Emotion	Experience that feels good or bad for the system itself	No: experience can be close to neutral, as when staring at a gray wall
Self-awareness	Experiencing oneself as a self, or knowing that one is experiencing	No: infants, and adults absorbed in a film, are widely thought to experience without reflecting on themselves
Richness of detail	How much an experience contains: its detail, range and depth	No: a single tone heard in darkness is still experience. On this theory, reach and openness may add richness, a hypothesis
Intelligence	Solving problems well	No: a capable controller need not experience anything

[Birch et al. \(2020\)](#) distinguish several dimensions along which consciousness varies, including perceptual richness, evaluative richness and selfhood. This theory treats emotion, self-awareness and richness of detail as such dimensions: they vary what experience contains, not whether there is any. Two systems could both be conscious and differ on every dimension.

4.2 The present

Three times need to be kept apart:

- **When the model operates.** Always now.
- **The time a content represents.** A seen object is represented as now, a memory as past, a plan as possible.
- **The model's relation to the outside world.** In perception, the model is kept in step with the world.

The theory's claim concerns the first two. A model holds a present when its contents are coordinated in represented time and constrain one another: each is placed relative to one now, at the moment it describes, so reports of the same moment are joined even when they arrive at different times. A shared processing clock is one way to achieve this, not the condition itself: two simulations can share a processor's clock without sharing a present, and parts updated at different times can still describe one moment. Placement can be wrong, as the examples below show; the condition is the organization that does the placing, not its accuracy. Holding a present is one condition; H2 proposes that operating the complete organization is experience. Remembered content is part of the present, tagged as past.

Keeping in step with the outside world is what accurate tracking requires, not what experience requires. A delayed video feed is still experienced as now, though its information is late; a dream is experienced as now, though it tracks nothing outside.

4.3 Modes and source tags

One model runs in several modes, and in a waking person they run at once. Each content carries a tag for where it came from.

Mode	What drives the model	Content is tagged as	Example
Perception	The senses	Seen, now	Watching rain from a porch
Prediction	The model's own forecast	Inferred, now	Tracking a child behind a van
Memory	Self-produced, often cued	Remembered, past	A thunder recording brings back a childhood porch and the smell of rain
Imagery	Self-produced, usually deliberately	Imagined, placed in the present or in a possible scene	Picturing a vase on the table in front of you
Thought	Self-produced, about what cannot be sensed	Inferred or possible	"She's probably home by now"; "the bridge could collapse"
Dream	Self-produced	Seen (wrongly), now	A dream

Source tags keep these apart. Imagining a bridge collapse should update a possibility, not overwrite the belief that the bridge is standing. Psychologists call this reality monitoring ([Johnson and Raye, 1981](#)); when it fails, imagined content can be taken for perceived content. The reverse also happens: people asked to imagine objects such as a banana on a blank screen, while a faint image of the object was secretly projected there, took what they saw for their own imagery ([Perky, 1910](#)). Source tags belong to this account of modes, not to the conditions for any experience. The theory does not claim experience needs them beyond what placing contents in time requires, as when senses arrive with different delays. When the tags are wrong, as in dreams, experience continues.

On this account, dreams feel real because generated content is tagged as seen. That is an interpretation, not a finding.

Generation builds from parts learned through perception. Someone who has never been in rain can know facts about it but, on this theory, cannot call up its smell. Applied to Frank Jackson's "Mary's room" ([Jackson, 1982](#)), this is an interpretation, not a demonstration.

4.4 One model

Parts are pieces of the system that each hold their own estimate of a shared quantity, such as the visual and balance systems each estimating whether you are moving. The sense organs that feed them are not parts: the eyes send the same signal however the model reads it.

Several parts form one model when every part can repair, and be repaired by, every other part, directly or through other parts. This must hold for the quantities they share. Repair is judged link by link: one part repairs another directly if its correction takes effect within a stated window, and a chain of such links can take longer. The repaired part must then carry the shared picture forward itself.

On a parked train, when the train beside you pulls away, you may briefly feel that you are the one moving. The jolts and rumble the model predicts for real motion never arrive, so the rest of the model overrides vision's estimate, and from then on you see the other train moving and yourself still.

To find where one model ends, ask which parts help fix which. One part helps fix another if removing or scrambling it makes the other's correction worse, even when it only works together with others, as in a majority vote. A model is a maximal group in which every part helps fix every other, directly or through others: no part can be added while keeping that property, though the group need not be the largest such group in the system (in graph terms, a strongly connected component). A model needs at least two parts. A single part with no partner forms a component of its own in the graph, but it shows no repair between distinct parts, so it does not meet joint repair. So for any one quantity, each part belongs to at most one model; for different quantities it can belong to different ones. Two disconnected networks are two models, and the time window applies to each link because a deadline on the whole chain would let groups overlap.

Joint repair need not wait for a noticeable error. On this theory, ongoing two-way exchanges between the parts of a model maintain their agreement, correcting small deviations as they arise. It therefore predicts that even a steady experience, such as an unchanging tone, depends on this exchange continuing, not merely on the parts being connected. Whether that dependence holds is an empirical question, which the proposed minimal-experience study aims to test.

The exchange has a time limit. Exchanges for a content that pause for longer than the window over which the model composes one present are not repairing it, whatever the connections allow, and the content is then not in the running model. This is the per-link window above. A link that pauses beyond it does not count as repair during the pause, but the content stays in the model while its remaining timely links still connect it to the other parts. The window is fixed and preregistered for each test. Deriving it from measurements such as how far apart in time a sight and a sound can arrive and still be experienced as one event, which is a fraction of a second, is an auxiliary hypothesis that needs its own validation: such windows measure something different from how often estimates must exchange corrections, and they vary with stimulus, task and criterion. One consequence is accepted in advance: content held silently in working memory for seconds, with no ongoing exchange, is kept but not experienced during the silence, and is experienced again only if it is restored to the running model. This differs from unattended content, such as the floor's color, which stays in ongoing exchange. Silent keeping occurs ([Trübutschek et al., 2017](#)); whether seen content kept that way stops being experienced is untested. No current method can test it directly, since asking about silently kept content draws attention to it and restores it; a test first needs a marker of experience that does not depend on report and has been validated elsewhere.

Joint repair is necessary on this theory, but its presence somewhere in a system settles little. Recovery after corruption does not show it: in round 29, central networks passed the recovery

benchmark although their repair ran only from the hub to the modules. Two-way repair may also arise wherever parts exchange estimates, even weakly. What carries the theory's claims is the boundary: which parts repair one another and so form one model, and which contents that model holds. That boundary depends on the map's bars. In round 29's distributed networks, most repair graphs held when the bar for a link was raised from 0.02 to 0.04, and 26 of 36 at 0.06, and the theory does not yet say which threshold picks out a subject, an open problem stated below. The tests accordingly ask whether particular content is inside a model, not only whether a system contains repair.

The parts are fixed before testing, so a result cannot be rescued by redrawing them. That does not make modules and hubs the right scale. The repair map evaluates one stated division into parts and does not exhaust the organizations inside a network: a network that fails as a whole, such as one whose modules do not repair its hub, could still contain a smaller group, inside the hub for example, whose parts repair one another.

The fixed parts and thresholds belong to the measurement, not to the claim. Under H2, whether a system experiences anything cannot depend on where an observer draws its parts, or on a repair link falling just above or below a test's threshold. On the claim itself, repair comes in degrees, and one model would be a group of parts that repair one another at whatever grain the system's own dynamics pick out. For a fixed quantity, grain and threshold, models defined this way cannot overlap. What the theory does not yet say is which grain, quantity and threshold pick out a subject; how models for different quantities, which can share parts, combine into one subject; and what follows when a stricter threshold splits one model into smaller ones inside it. Integrated information theory answers the corresponding question with an exclusion rule: only the most integrated grouping counts, and groupings cannot overlap (Tononi and Koch, 2015). This theory has no such rule yet; it is an open problem.

Repair runs through the world when it passes through the system's own actions and back through its senses, whatever the medium. Such repair does not count. This covers a split-brain patient whose one hemisphere watches what the other's hand does.

Repair also comes in degrees, so counting needs a threshold, such as the preprint's joint-test bar. The parts and the threshold must be fixed before counting, or recounting could rescue any result.

The procedure identifies groups of mutually correcting parts. That each such group is one subject is a further hypothesis, not something the definition establishes. In split-brain patients, and in the octopus, whose arms are linked by a nerve ring (Chang and Hale, 2023), the answer may differ by content.

5 The conditions

Each feature sits in one category, and the table says whether it must be operating during an experience or only available.

- **Any experience:** proposed as required whenever there is experience. Under H2, they are the parts of the organization identified so far.
- **Some kinds:** required for particular kinds of experience.

- **Richer experience:** hypotheses about broader experience, not yet shown to matter for experience at all.
- **Open:** status undecided.

Feature	Category	Operating or available	How it is tested	Evidence so far
Running: each state of a content comes from the one before through the model's own dynamics, including actively held stable states, with or without input	Any experience	Operating: content is actively maintained or advancing, not passively stored or played out from a store	Nudge the model's state and check that later states follow from the nudged one (the nudge test). The preprint's benchmark (C1k) cuts all input and scores the forecast step by step against an ideal observer; it shows adequate forecasting	Learned by small networks trained with periods of closed eyes: all 21 kinds given the nudge test, in the linear and binding worlds, passed it, as did all nine kinds trained in round 29's delayed world
Joint repair: the parts correct one another	Any experience; also defines one model	Operating: the parts are constraining one another now	Give one part wrong information; the others must override it, and the repaired part must carry the picture forward (C4, the recovery benchmark). That shows a corrupted part recovers, which central networks whose repair runs one way also achieve (round 29); joint repair itself is shown by the map of who repairs whom, from removing or scrambling each part in turn	Recovery learned, as the benchmark measures it, in fixed-rule worlds through a central hub (partly by construction) or when single parts were regularly dark; not met in the world whose rules switch. In round 29's delayed world the two halves tended to split by architecture: central networks passed the benchmark, but repair runs only from the hub to the modules; distributed networks mostly formed nearly complete maps of mutual repair, but most fell one to three parts short of the benchmark

Feature	Category	Operating or available	How it is tested	Evidence so far
A single present: contents are coordinated in represented time, each placed relative to one now at the moment it describes	Any experience	Operating	Make some senses arrive late, then displace one late report: every part must file it under the moment it describes, which the world's rotation reveals	In its elementary form, learned where two senses arrive three steps late: all nine kinds compensated for the fixed delay in the readouts tested, placing a displaced late report 2 to 3 steps back, while networks trained without delays read 0 to 1, each with one exception among 54 networks (round 29). That every part shares the placement was checked only in the two kinds whose module heads could be read. Where every sense arrives at once, a system that never places contents in time behaves like one that does; a shared clock, presence flags or timestamps alone would be weak evidence
Action selection: the model's contents, including its predictions of what the system's own actions would do to its input, take part in choosing what the system does	Some kinds: the experience of acting, and of the world staying still as one moves (hypothesis)	Operating when acting is experienced, including when movement is blocked, as in dreams, paralysis and locked-in syndrome	Perturb a content of the model and check that selection changes accordingly; check that the model anticipates input changes caused by the system's own actions and treats the same changes as news when the world causes them; compare otherwise identical networks with and without action selection	None yet; the preprint's networks only predict, so they do not meet it
Object-specific binding: evidence is assigned to particular objects	Some kinds: experience of objects across senses	Operating when several objects are experienced across senses; a single tone in darkness needs none	Nudge one sense's view of one object and measure how each of the other sense's objects responds (B_used)	None of 102 networks in rounds 14 and 18 met it at the standard bar; in round 27 some single networks met it in single versions, though no kind did
Updating: the model's rules change when the world's do	Open: the preprint proposes it as necessary, but it may only improve tracking	Available; not exercised in every experience	Change the simulated world's rules mid-run; forecasts must adapt (U)	Partly learned

Feature	Category	Operating or available	How it is tested	Evidence so far
Self-location: the experiencer placed in the scene	Some kinds: bodily self-experience (hypothesis)	Operating when the body is experienced	Not yet tested	Out-of-body experiences show it is computed and can fail
Reach: representational scope	Richer experience (hypothesis)	Available	Proposed (see Extensions)	None yet
Openness: representational flexibility	Richer experience (hypothesis)	Available	Proposed (see Extensions)	None yet

“Operating” means working, not getting it right. An illusory conjunction is binding at work producing a wrong pairing; the theory says you experience the pairing the model formed.

Binding is claimed only for experience of objects across senses. This narrows the preprint, which proposed binding as necessary without that qualification, though its checklist already applied binding only where several objects are present.

The preprint’s binding test certifies the routing of motion evidence to objects, not the attachment of features such as color. Claims about feature pairings, such as illusory conjunctions, need a feature-attachment test of their own.

The preprint’s benchmark scores pace against the world only because simulations offer no other ground truth. It therefore tests forecasting that keeps pace with the world: a dream would fail it while still running in the theory’s sense.

Forward-moving representations can also be observed without tracking the world. In people, mental rotation takes longer the further a shape must turn ([Shepard and Metzler, 1971](#)), and the remembered orientation of a rotating shape drifts forward along its rotation ([Freyd and Finke, 1984](#)). Whether these run in the theory’s sense needs a nudge test.

Running, in the theory’s sense, means each state comes from the one before through the model’s own dynamics. The nudge test checks this. During a blackout, displace the represented content itself, consistently across the parts, rather than its timing or how easily it can be read out. In a running model the nudged content then evolves by the same rule that advances the undisturbed content, starting from the changed state. In a linear world, such as the preprint’s, the displacement itself follows that rule, so for a rotating state the offset rotates with it. In nonlinear dynamics it need not; what must hold is that the nudged trajectory follows the model’s own dynamics from where the nudge left it. A forecast stored at the start of the blackout and played out discards the displacement at once, and a passively stored value keeps it unchanged.

Held contents need controls of their own. When a model actively holds a stable content, its dynamics may work a nudge off, but a quick return could equally be a stored copy overwriting the nudge, and a displacement that simply stays could be a passive store. A test of held content must show that what follows the nudge comes from the model’s own dynamics, with a stored copy and a passive store as controls.

The preprint’s benchmark counts a forecast computed in one burst and then played out as running, and the preprint accepted a partial forecast played out on a fixed clock as lossy running. The theory’s sense of running excludes both. Benchmark results are therefore reported here as adequate forecasting, after the preprint’s description of what the benchmark certifies: “adequate prediction that advances in step with the world.”

A later preregistered round, round 28, gave the nudge test to 126 of the preprint’s networks, 21 kinds in all: the 72 from round 18, in the binding world, and 54 in the linear world. The test accepted the ideal observer and refused burst playback, the phase-gain forecast, a frozen hold, random networks and networks trained without blackouts. Every kind passed in at least 5 of 6 versions. Kinds carried 0.66 to 1.06 of the ideal observer’s displacement forward, on average, turning it at the world’s rate. So their forecasts run, in the theory’s sense; they are not played out.

Running was not complete in every kind. In an exploratory breakdown, the kinds that carried least lost 10% to 19% of the displacement at the first step, and some lost a further 5% to 9% per step beyond the world’s own decay. Whether that first-step loss is content that does not run, or part of the nudge that missed where the network holds the content, is untested.

5.1 What the core implies for small systems

In round 18 of the preprint, 71 of 72 small networks passed both the forecasting benchmark and the recovery benchmark (C4), the preprint’s joint test. That shows adequate forecasting and repair as the benchmarks measure them, and half of the 72 are central networks whose repair is partly built in. Round 28 added the nudge test: 71 of the 72 passed it, and because the single failures on it and on the recovery benchmark fell in different networks, 70 passed it together with both benchmarks, and every kind did so in at least 5 of 6 versions. They run in the theory’s sense.

These networks therefore run, in the theory’s sense, and pass the recovery benchmark. Whether their repair meets the theory’s own definition, in which every part repairs and is repaired by every other, is not established: the benchmark shows that a corrupted part recovers, not that each part helps repair the others. In the central networks, the modules keep no memory and are redrawn from the hub at every step, so a displaced module is simply overwritten and the hub alone carries the picture. The third “any experience” condition, a single present, cannot be tested in their world. Every sense reaches them at once, so a network that never places a report at the moment it describes would behave exactly like one that does, and no test there could tell them apart. The test instead gives networks with the same architectures a world in which some senses arrive late, as a voice reaches you after the lips move, and asks whether all their parts file a late report under the moment it describes. If such networks pass it, and also run and repair one another in the theory’s sense, as the same trained systems, they meet every “any experience” condition named so far. If those conditions complete the organization, H2 and H3 would attribute minimal experience to them.

This is the small network argument applied to this theory: very small systems can meet a theory’s proposed criteria ([Herzog et al., 2007](#)). By the theory’s own rule, a condition named after seeing a result cannot be used to exclude it. The single present is not used that way here: networks with these architectures take its test in a world where it can be passed or failed, under criteria fixed before any of them is scored. Round 29 gave them that test.

Round 29 trained nine kinds of small network, including these two architectures, in a linear world where two senses arrive three steps late. All nine kinds learned to compensate for the fixed delay in the readouts tested, placing a late report 2 to 3 steps back against a true delay of 3 in 53 of 54 networks (one read 5), while round 9A's networks of the same kinds, trained without delays, read 0 to 1 in 53 of 54 (one read 5). That every part shares the placement was checked only in the two kinds whose module heads could be read. All nine also forecast adequately and ran. Joint repair tended to split by architecture. In the central networks the hub repairs every module but no module repairs the hub, so their parts do not form one model by the theory's definition, although they pass the recovery benchmark. The distributed networks' parts mostly formed nearly complete maps of mutual repair, but most of these networks fell one to three parts short on the benchmark, usually because their on-time parts did not recover. Repair links that can be detected and repair accurate enough to restore the picture can therefore come apart.

How those maps arise is not established: they do not identify the repair mechanism or rule out averaging as a contributor.

No kind passed all five tests in 5 of 6 versions in round 29. Five individual networks did, at the standard bar, each as the same trained system. Rescored on fresh draws in round 30, three of them passed again, all of one distributed kind, and in each, every part repaired every other. Each cleared the recovery benchmark with exactly the three parts required, and a later reliability check found them passing it in 6, 5 and 6 of 6 scorings, and for all three the single present was read from the main readout alone. Three networks therefore passed all five operational tests on both scoring runs; whether every part takes part in one coordinated present remains unestablished.

In round 31, one kind trained with single on-time parts regularly darkened passed all five tests in 6 of 6 versions when the recovery benchmark was averaged over six scoring runs, but in 4 of 6 on a single run, and two versions cleared the bar by less than 0.005. Why the training helped is not established: the on-time parts' recovery rose only modestly, and the same training cost another kind its passes.

If the named conditions complete the organization, and every part of these networks shares one present, H2 and H3 would attribute minimal experience to them. Under the separate emotion hypothesis, since nothing in them represents outcomes as mattering to the network itself, it would be experience without affective valence. Whether the named conditions are the complete list remains open until the list closes (see Section 3).

If the list closes with only the conditions named now, H2 and H3 would attribute minimal experience to any system whose parts run, repair one another and share one present in full, including systems far simpler than a brain. The theory accepts that consequence. A further condition may be added before the list closes only if it is justified by independent features of experience, never to exclude small systems after they pass. Minimal here means minimal: content limited to the few quantities the parts share, nothing felt as good or bad unless the model represents stakes, and no reflection on it. Integrated information theory accepts a similar consequence for simple systems with integrated structure (Tononi and Koch, 2015).

The dog shows how small the core must stay. A dog has little long-range thought and much less reach than a person, yet is widely thought to be conscious. So the extensions cannot all be required.

5.2 The action condition

A role in selecting actions is required only for some kinds of experience: the sense of acting, and the world's felt stability as one moves. It is not required for experience as such.

The control-system view explains why the model exists, not what experience needs. What something is for is not what it needs in order to work: eyes evolved for daylight, yet people see in dreams. Nearly all evidence of experience in people also comes through report, which is itself an action, so a requirement of action could hardly be tested in people.

Where the condition applies, it concerns a current role. The model's contents, including its predictions of what the system's own actions would do to its input, must feed the process that chooses what the system does. Developmental or evolutionary history does not count: as in the preprint, the route does not matter, only the organization. Nor does the bare possibility of use, since almost any model could be wired to a motor.

Selection itself may sit outside the model; the condition requires that the model's contents feed it. The fast loops that carry movements out, which the theory also places outside the model, do not count. When movement is blocked, selection and the model's predictions must remain, including when the choice is to keep still; only the path to the muscles or motors may be missing.

In dreams, the dreamer often acts, while the brainstem blocks the movements. Where that block fails, in REM sleep behavior disorder, people act out their dreams ([Schenck et al., 1986](#)). In locked-in syndrome, attempted movements still produce motor-cortex signals that an implanted brain-computer interface can use ([Vansteensel et al., 2016](#)).

Feeling that one acts is not the same as selecting the action. In one study, people watched another person's arms move where their own arms would be. When they heard instructions previewing each movement, they reported more control over the movements, though they controlled nothing ([Wegner et al., 2004](#)). On a given occasion, then, the sense of acting can be inferred from cues such as a matching prior thought, and it can be wrong.

This counts against any version of the condition that requires selection for each act experienced as one's own. The condition concerns the organization instead: the experience of acting requires a model whose contents currently feed the choice of actions, while on any one occasion that experience can be misattributed, as an experienced pairing of features can be wrong. Since this formulation was chosen with that finding known, the finding cannot count in its favor, and it awaits a test of its own. Selecting and inferring are told apart by intervention: selection is shown when changing the model's content before a choice changes the choice; inference, when cues change the felt sense of acting while the action stays the same.

The strongest case for requiring action in all experience is that experience is always experience for someone: in Thomas Nagel's phrase, there is something it is like for the organism ([Nagel, 1974](#)).

On that view, the someone is the system that uses the model to choose what it does: predicting what its own actions would do separates what it does from what happens to it, and a model that only forecasts its input has contents that are for no one. The theory does not adopt this, because the key step, that the someone must be the model's user, is the point in question. Galen Strawson's *Weather Watchers*, imagined beings that experience the weather but cannot act, show that the opposite is at least conceivable ([Strawson, 1994](#)).

Sensorimotor theories go further than this theory for perception. They hold that seeing is a way of acting, and that visual experience arises when an organism masters how its input changes with movement (O'Regan and Noë, 2001). They also deny that internal representations generate experience, which this theory asserts.

Butlin et al. (2023) list agency and embodiment among their indicators: learning from feedback and selecting outputs to pursue goals, and modeling output–input contingencies and using that model in perception or control. They treat these as making consciousness more likely, not as necessary conditions, as this theory does.

6 Conscious versus unconscious processing

This is the theory's most important open argument. On this theory, content is experienced when it is held in the running joint model of the present, as part of the complete organization. Operationally, held content takes part in the model's dynamics and is in two-way repair with its other parts; the test below asks whether that is enough. Unconscious processing influences the model or behavior without being held in it.

An everyday case shows the difference. You sit blindfolded on a chair in an empty room and hear a swoosh of water. A moment later you feel wetness on your leg and realize the room is flooding. On this theory, the realization comes from the swoosh having been held: it stayed in your picture of the room as water nearby, just now, so the wetness could be joined to it. The example illustrates the hypothesis rather than showing it, since sounds that are never experienced can also be kept and shape what follows, as the next sections describe.

The two senses then correct each other. The sound tells you how to read the wetness: it is the water you heard, not a spilled drink. The wetness changes what the sound was: a flood, not a pipe. Had the swoosh only been registered, as sounds often are during sleep, the hypothesis predicts that the wetness would have arrived as a surprise, not joined to the sound.

6.1 Influence is not enough

In one study, pictures that people rated invisible still pulled where they heard a sound, though less than visible pictures did (DeLong and Noppeney, 2021). So cross-sensory influence cannot by itself show that content is experienced. The preprint's tests cannot be read as consciousness tests.

6.2 Keeping is not enough either

Unseen information can be kept briefly. Brain activity maintained the presence, and more weakly the orientation, of unseen gratings until a probe 800 milliseconds after they appeared (King et al., 2016). People also located unseen targets above chance after several seconds, with no sustained activity during the delay (Trübutschek et al., 2017).

The theory's claim must therefore be about running and repair together, not keeping. Unseen content may be kept, actively or silently. The theory says it will not both take part in the model's ongoing dynamics and be in two-way repair with the rest of the model.

Under continuous flash suppression, unseen implied-motion images neither biased perceived motion nor primed other still images implying the same direction of motion. They weakly primed only identical images (Faivre and Koch, 2014). Implied motion is not tracking through a gap, so this fits the claim without testing it.

6.3 A test within people

Show pairs of objects that move as one, as if joined by a rod, and pass behind an occluder; in some trials one object of each pair is made invisible. Match seen and invisible objects for how well people can objectively detect them, so that visibility is not confounded with signal strength.

Unseen content may still advance or influence other senses. In one study, people responded faster to probes where masked moving objects would be next, even after a bounce (Goldstein et al., 2020). The theory predicts only that unseen content will not meet the criteria for running, and for two-way repair with seen content, at the same time.

Repair must be shown, not assumed from influence, because graded cross-sensory influence happens without awareness. As in the preprint’s joint test, give one object of a hidden pair brief wrong information, such as a misleading cue about its speed, then remove it. Repair is shown if later predictions for that object return to the motion its partner implies and continue from there.

Compare cues tied to an object outside the pair, and sounds that only attract attention. Match seen and unseen trials for detectability, the motion information given and localization precision. Some extrapolation happens as early as the retina (Berry et al., 1999), so local extrapolation alone would not count against the theory.

The test makes one operational commitment. For this task, content that meets the preregistered criteria for running, and for two-way repair within an independently identified model, is predicted to be experienced. The model is identified by the object participants report seeing, and the invisible object counts as inside it if, for their shared motion, each repairs the other within a preregistered window: wrong information given to either is corrected by its partner.

Model membership, any further qualifying conditions, the exclusions and the criterion for “not experienced” are fixed before data collection. That criterion covers any awareness of the object, including as hidden or inferred, so source tags cannot explain away a counterexample. The criterion concerns the particular object on each trial: knowing in general that such objects can appear is not awareness of it, while seeing or glimpsing that trial’s object, or consciously expecting or picturing it, is. A single question about whether the object was seen cannot separate these, so the awareness measure asks about each, and the design should keep general knowledge from becoming expectation on every trial, for example by not telling participants that hidden objects may appear. Awareness is rated on a graded scale fixed in advance, such as the four-point Perceptual Awareness Scale, and only its lowest rating, no experience at all, counts as not experienced (Ramsøy and Overgaard, 2004). Not reporting an object is different from reporting that it was not there: people who glimpse a display can fail to report most of it while saying they saw it all. A valid counterexample rejects the operational hypothesis, and under the rule against later additions (see Section 16) nothing named afterward can rescue it. The test does not reach the complete organization, which is not yet specified.

Two checks govern how a result is read. If seen and unseen content both fail the preregistered

running and repair criteria, the comparison cannot count as support for the predicted difference and stays inconclusive until the measures are validated, since the assay may be insensitive. But if independently validated measures show that consciously experienced content lacks a required feature, that would challenge H1 itself. And a failure of unseen content counts only if the study could have detected the effect: the smallest running and repair effects that matter are fixed in advance, and the study must be sensitive enough to rule them out, for example with equivalence tests.

Decoding the hidden object's position during the gap does not show running, since a stored forecast can be decoded too, and a neutral flash can make silently held memories decodable (Wolff et al., 2017). Running would be shown by the nudge test: displace the represented position of the whole pair, for example with a cue that shifts where both hidden objects seem to be, and check that the pair's path then continues from the displaced position by the rule that advances the undisturbed path. A change in decodability, speed or attention does not count. A cue through the senses could trigger a freshly computed forecast, so a displacement produced by brain stimulation, if one can be found, would give the cleaner test.

If unseen content met every criterion but performed worse than seen content, that would not rescue the hypothesis. Global workspace theory may make a similar prediction, so this test may not separate the two; see Section 11.

Comparisons within people hold the material constant and vary mainly whether content is experienced. That makes them more informative here than comparing people with machines.

6.4 A second test: ambiguous rotation

A cloud of moving dots can look like a sphere turning either way, and the seen direction flips now and then though the input stays the same. Touch can steer it: people who touched a rotating sphere while watching tended to see the display turn the way the sphere turned (Blake et al., 2004). The self-motion illusion under Section 10 is a case of the same kind.

Since the parts of one model repair one another, they should agree about the direction. Agreement alone does not require choosing one direction: parts could repair one another toward a shared blend of both. Committing to one direction at a time is a further hypothesis, suggested by people seeing one direction at a time; on it, expectations about where hidden dots will reappear, and about what a hand on the sphere would feel, should flip with the seen direction. Because the input stays the same, the change comes from the model, not the senses.

The same question can be put to networks trained in a world where one sense cannot tell the two directions apart and another sometimes can. Give one part a brief push toward the other direction and see how the disagreement resolves: in a jointly repaired model, the pushed part should be pulled back or carry the others with it, while independent parts would simply disagree.

Cutting communication between the parts is weaker evidence on its own, since parts that receive the same input or share the same bias can agree without communicating; a cut shows only whether that communication was needed. A control that replaces the cut messages with recorded ones separates that effect from general damage.

Whether trained networks commit to one direction at all is open. An estimator built to minimize

average error does not: its best estimate of what it cannot sense is the average of the two directions. Parts that commit and switch together would show a property such a tracker lacks. Commitment is not proposed as a further condition for experience, however. It would not set brains apart from engineered estimators, since trackers that keep several hypotheses and report the most probable one also commit, and in binocular rivalry people sometimes see a patchwork of both images rather than one at a time. Adding it now, after networks have passed the existing tests, would also break the rule against conditions named after the results (see Section 16).

A first attempt with networks was shelved before any confirmatory training or testing. Its exploratory pilots, including three module sizes and gated units, failed a feasibility bar fixed before the first pilot: the direction signals the networks retained were too variable to be measured reliably. The rule for shelving the round on that bar was fixed before the last pilot. The network version waits on networks that hold such a direction reliably through darkness.

As with the other network tests, a pass would show organization, not experience. Running both versions lets the network results be set beside the human ones, where reports of what is seen are available.

6.5 A minimal case: one tone in darkness

The argument for joint repair comes mostly from rich scenes. The hard case is a single steady tone heard in darkness. Its pitch, loudness and onset are different features, not separate estimates of one quantity, so they do not by themselves supply parts that repair one another.

Even then, the hearer is not running on one sense. Balance, the sense of head position and touch keep estimating where the body is, and the tone is placed in that model: it is heard as coming from somewhere relative to the hearer, at a moment in the hearer's present. That placement combines different quantities, though. The ears give the sound's direction relative to the head, and the body senses give where the head is pointing; combining them computes where the sound is. That is not two estimates of the same quantity correcting each other, which is what joint repair requires.

The theory therefore makes a further, testable commitment: even a minimal auditory experience depends on two representations that each estimate the same property of the sound and correct each other in both directions. This is a prediction, not an established finding. One candidate is the tone's frequency as represented at different levels of the auditory pathway, since descending projections from the cortex adjust how lower levels respond to sound (Suga and Ma, 2003). The two ears' cues to direction are a weaker candidate, because people who hear with one ear still hear tones. Repair must be shown in both directions with the nudge method: a brief perturbation of either representation is pulled back toward the other. Shifting where a sound seems to be by perturbing balance would show influence, which by this theory's own definition is not repair.

A test must separate experience changing from experience disappearing. If repair is necessary for any experience, a tone that stays consciously audible after every qualifying repair for it has been eliminated counts against that claim, even if its pitch or apparent place changes. Weakening one connection is not enough, since another qualifying pathway might remain. Elimination is hard to achieve. A sound that follows the head as it turns, as over ordinary headphones, tends to be heard inside the head rather than out in the world (Best et al., 2020): its place changes, but it is

still heard. On this theory that is expected, since tying a sound to head movement is combination, not the repair claimed necessary.

The networks' threshold cannot simply be carried over to two parts. The repair map's bar was set for networks of four or more parts, and the recovery a given kind of repair produces depends on the number of parts, on how fast the represented quantity and the perturbation decay, on the scoring window and on the readout. A threshold for two parts must therefore be set and validated on its own, against two-part systems of known structure, before the study.

6.6 Measuring repair in the brain

The repair the theory requires is a causal property, so correlated activity, information flow or feedback cannot show it by themselves. A measurement in the brain needs four steps: identify two populations that each represent the same property, such as an object's orientation; perturb one population's estimate and show that the other corrects it; show the same in the reverse direction; and show that the corrected population carries the correction forward.

Recovery alone is not enough. While a stimulus stays in view, a perturbed population can recover because the input drives it back, not because its partner corrects it; and, as round 29's central networks showed, one-way correction can pass a recovery test, so both directions must be tested. The cleaner signature is the direction of recovery: bias the partner's estimate slightly, then check whether the perturbed estimate returns toward the partner's biased value rather than toward the input. The unseen-motion test should add the same check: bias the seen partner slightly and test whether the hidden object follows the bias.

Decoders must be validated before they are trusted with this. A forward model can generate synthetic recordings from a system that runs and repairs, and from its look-alikes: a passive buffer, a one-way hub and a code that drifts while the represented state stays fixed. The decoders and repair measures must tell these apart before any real data are analyzed, as the network measures were first checked against known look-alikes.

6.7 What this does not settle

This question is narrower than the hard problem: why any of this feels like something from the inside. The theory does not solve that.

Thomas Metzinger's self-model theory ([Metzinger, 2003](#)) offers a partial answer. The model is transparent: the system cannot see it as a model, so it takes the model's content to be the world and itself.

7 Extensions, held as hypotheses

Four further ideas came out of thinking about cars, memory and the self. Each is a hypothesis about richer experience or about how content beyond the senses enters the model. None is proposed as required for all experience; self-location is proposed only for bodily self-experience.

7.1 Reach: representational scope

Reach is how far and how long the model keeps what it cannot sense, and keeps it accurate. A self-driving car's present is a bubble of a few hundred meters and a few seconds. A person driving still carries the garage behind them and the people at home.

Two cautions apply. Knowing your family is at home may be a stored belief updated when relevant, not a continuous simulation. And more reach means broader content, not necessarily more consciousness: consciousness may vary along several dimensions rather than one level ([Birch et al., 2020](#)).

A test would hide objects for longer spells, sometimes leaving only indirect evidence such as footsteps, and score the estimate against an ideal observer. A tracker that never deletes anything is the look-alike to refuse. Whether reach matters for experience, rather than cognition, needs comparisons in people.

7.2 Openness: representational flexibility

Openness means the model can come to make a distinction it did not make before, from input it already receives. A radiologist learns to see a hairline fracture that was always in the X-ray; a wine novice learns to taste tannin.

The test must hold the senses fixed and introduce a distinction the input already contains but the system has never needed. Otherwise failure may only mean the system lacks the signal. A gadget that blinks and hums at once would not qualify: that is a new combination of familiar features, not a new distinction.

7.3 Thought: generation inside the present

A thought is the model running on its own about what it cannot sense. Its results are written back as tagged content, inferred or possible, and normally do not overwrite what is perceived. This definition does not assume consciousness; it describes what the model does.

Thought runs on the same model, not a separate module. Imagining a scene uses much of the visual machinery that seeing does ([Pearson, 2019](#)). Resting rats replay place-cell sequences along routes to remembered goals, including routes not yet taken ([Pfeiffer and Foster, 2013](#)).

What starts a thought needs no inner thinker: cues, estimates that have gone stale, unfinished goals ([Klinger, 2013](#)) and attachments to people and places. Nostalgia shows generation with feeling included; it is often triggered by music, smells, low mood and loneliness ([Wildschut et al., 2006](#)).

Most AI systems that imagine possible futures send them to the part that chooses actions, as a car's forecasts go to its planner. The hypothesis is that in people they are written back into the present picture, tagged. That difference can be built and tested.

7.4 Self-location

The human model places the experiencer in the scene, at a location, in a body. Out-of-body experiences show this placement is computed and can fail.

Self-location is a hypothesis for bodily self-experience. A car also models its own position, so self-location does not by itself separate people from cars.

8 Universal organization, local content

What organizes experience is meant to be the same in everyone; what fills it comes from the world each person grew up in. In the preprint’s terms, genes are like a network’s architecture and learning rules, and upbringing is the world it is trained on.

8.1 Same input, different experience

A Hindu and a cowboy watching the same sunrise receive the same light. For someone raised in Hindu practice, dawn may be when the Gayatri mantra is recited and water is offered to the sun. For a ranch hand, it is when the day’s work begins.

Each model binds its own memories, meanings and bodily feelings to the scene. On this theory, that meaning is part of the experience, not a thought added afterward: the worshipper sees a sacred sunrise.

Whether culture reaches perception itself is debated. [Firestone and Scholl \(2016\)](#) argue that no study yet shows beliefs changing perception. Lisa Feldman Barrett’s account ([Barrett, 2017](#)), in which the brain builds emotions from body signals and learned concepts, fits this theory’s side.

Individuals vary. The label does not fix the content; each person’s history does.

8.2 What upbringing changes

Finding	What happens	How the theory reads it
Speech sounds	Infants told apart some speech contrasts their own language does not use, then found them much harder to distinguish by the end of the first year (Werker and Tees, 1984)	The model keeps the distinctions its world uses
Music	The Tsimane’ of the Bolivian Amazon rated consonant and dissonant chords as equally pleasant, unlike Bolivian town dwellers and US listeners (McDermott et al., 2016)	Which chords feel pleasant is partly learned
Faces	Korean-born adults, adopted into French families at ages 3 to 9, recognized European faces better than Asian ones (Sangrigoli et al., 2005)	The model is tuned to the faces it grew up with
Language loss	Korean children adopted in France as young children could not recognize Korean as adults, and their brains responded to it like an unknown language (Pallier et al., 2003)	The model can be retrained
Hidden traces	Korean-born Dutch adoptees, including those adopted as infants, relearned Korean speech sounds faster than matched Dutch adults (Choi et al., 2017)	Retraining leaves traces of the first world

These findings illustrate that content is learned. They do not test H1, because training can change content without changing running, repair or binding.

8.3 A thought experiment: raised in Chennai

A child of European descent, born and raised in Chennai with no contact with the West, would speak Tamil as his native language. Its food, music and gestures would feel like home. His ancestry supplied the hardware and some starting tendencies, such as temperament, but his model is a Chennai model.

Strangers would still predict “foreigner” from his face and be surprised by his Tamil. His model of himself would come to include how others see him.

8.4 A social model

The theory does not yet cover this. The present model also holds models of other minds, including their model of you. Like self-location, this is a candidate for some kinds of experience, not a condition.

9 Where the model sits: the control system

On this theory, consciousness is the model a control system uses to act in the world, not the control system as a whole. Guiding action is why the model exists, but the theory does not require every experience to involve it (see Section 5.2).

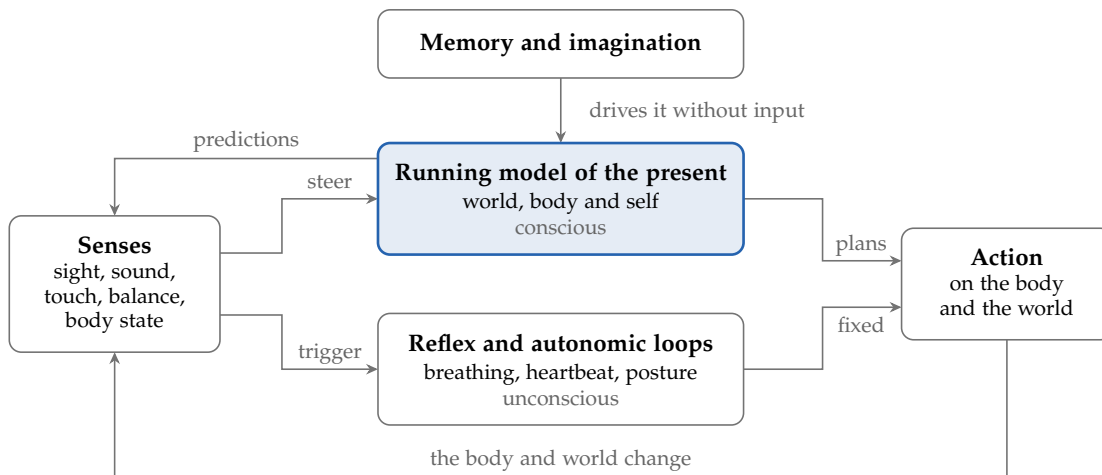


Figure 1: The running model handles what reflexes cannot: where consciousness sits in the brain’s control system.

The senses steer the running model and trigger reflexes (Figure 1). Only the model, which memory and imagination can also drive, is proposed as conscious.

Traffic also runs the other way. The model sends its picture back to the brain’s sensory areas as predictions, so what they register is shaped by what the model expects. In the self-motion illusion (see Section 10 below), activity in visual and vestibular areas shifts with the interpretation while

the input stays the same. These sensory areas are candidates for parts of the model. Establishing that they are parts requires evidence that their estimates are in joint repair and carry the corrected state forward; changes in activity that accompany a perceptual switch do not establish this by themselves. This feedback stays within perception; it does not mean that beliefs change what is seen.

In people, several spatial estimates can already be decoded from brain activity: auditory and visual areas predominantly encode modality-specific locations, while regions of the intraparietal sulcus encode combined estimates, including estimates that account for whether the signals share a source (Rohe and Noppeney, 2015). That study did not test whether these estimates are in joint repair and then carry the corrected state forward, as this theory requires.

Most control is unconscious: the brain runs breathing, heartbeat and posture without any experience of it. The running model handles what reflexes cannot, situations that are new, complex or partly hidden.

Good control needs a model. Roger Conant and W. Ross Ashby proved a version of this in 1970 (Conant and Ashby, 1970), in a paper titled “Every good regulator of a system must be a model of that system”: under the theorem’s conditions, a regulator that is as successful and as simple as possible must model the system it regulates. Its sense of “model” is narrow, and it does not show that every controller holds a running model of this theory’s kind.

Consciousness does not bring the body to life. The brainstem keeps the heart beating and the lungs breathing through deep sleep and general anesthesia; what the model adds is someone in the body.

Damage to the upper pontine tegmentum can cause coma (Parvizi and Damasio, 2003). On this theory, an arousal mechanism would count as an enabling condition for the model, rather than a part of it, if it sustained the model’s operation without taking part in joint repair. Whether particular brainstem circuits serve only this enabling role remains to be tested; the lesion evidence alone does not establish that distinction.

10 Present Model Theory in action

Everyday experiences and known findings show the theory at work. They were known when the theory was built, so they illustrate it rather than test it, and most fit rival theories too.

Finding	What happens	How the theory reads it
Propofol anesthesia and deep sleep	Blood and oxygen still arrive and neurons stay alive and active, though their firing patterns change, yet people usually report no experience. A magnetic pulse to the cortex spreads in a complex pattern when awake but stays local or simple in these states (Massimini et al., 2005; Casali et al., 2013; Sarasso et al., 2015)	The parts stop working together, which fits joint repair failing
Ketamine anesthesia	People are unresponsive, yet the cortex’s response stays complex and they report vivid dreams afterward (Sarasso et al., 2015)	The model keeps running in a self-generated mode

Finding	What happens	How the theory reads it
Children born without most of the cerebrum (hydranencephaly)	Some respond to familiar people, show preferences and react emotionally, though most keep some cortical tissue and whether they are conscious is debated (Shewmon et al., 1999 ; Merker, 2007)	The theory constrains organization, not location. If they experience, it predicts that their remaining tissue supports running, joint repair and a single present; this has not been shown in these cases. Reliable evidence of experience without any one of these would count against H1
Out-of-body experiences	People feel located outside their body. Brain stimulation, and virtual reality with synchronized touch, can induce this (Blanke et al., 2002 ; Ehrsson, 2007 ; Lenggenhager et al., 2007)	Experience follows the model's self-location, not the eyes
Phantom limbs	Most amputees still feel the missing limb, sometimes for years (Stankevicius et al., 2021)	The model keeps running a body part that sends no input
Tracking a hidden ball	Monkeys intercept a ball that passes behind an occluder. Activity in dorsomedial frontal cortex tracks the ball's position while it is hidden about as accurately as while it is visible, and recurrent networks trained to simulate the hidden ball match that activity best (Rajalingham et al., 2025)	Content keeps advancing when the senses go quiet, which fits running; a nudge test in the brain would be needed to confirm it
Illusory conjunctions	Under brief viewing, people see real colors on the wrong shapes (Treisman and Schmidt, 1982). A patient with damage to both parietal lobes did so even with long viewing (Friedman-Hill et al., 1995)	You experience the pairing the model formed, even a wrong one
Split brain	Two patients could report a stimulus's location from either side but could not compare stimuli across sides (Pinto et al., 2017); the reading is debated	A case for the one-model counting rule, applied content by content
Flash-lag effect	A moving object looks ahead of a flash aligned with it (Nijhawan, 1994)	On one leading account, the model extrapolates to stay in step with now
Self-motion illusion (the parked train)	Watching a display that turns at a constant speed, people switch between seeing it turn and feeling themselves turn. Activity in early visual motion areas and in vestibular cortex drops while they feel themselves turning, though the input never changes (Kleinschmidt et al., 2002)	Experience follows the model's interpretation, not the input, and the sensory areas shift together, which fits joint repair
Pressing on the eye	Pressing gently on the eyelid moves the eye, and the world seems to jump. When the eye moves on the brain's own command, about three times a second, the world stays still (von Helmholtz, 1867 ; von Holst and Mittelstaedt, 1950 ; Sperry, 1950)	The model predicts the input changes its own actions cause and keeps them apart from changes in the world; a shift it did not predict is read as the world moving

Finding	What happens	How the theory reads it
The blind spot	Each eye has a patch with no light receptors, where the optic nerve leaves, yet no hole appears in what is seen	Experience is the model's picture, which fills the gap, not the image on the retina
Ventriloquism	A voice seems to come from the puppet's moving mouth. Vision pulls the heard location toward the seen one, and sound pulls vision instead when the image is blurred (Alais and Burr, 2004)	The model binds a sound to the object it most likely belongs to, as when a bark behind a bush is joined to the dog seen a moment before
Wheels turning backward	In films, and sometimes even in continuous light, a turning wheel can seem to spin backward (Purves et al., 1996)	Experience follows the model's interpretation of the motion, which can differ from the motion itself
Motion blindness	After damage to motion areas on both sides of her brain, a woman saw color, shape and objects normally, but her motion perception was severely impaired; she could still see slow movement near the center of her gaze. Poured tea looked frozen, like a glacier, and people were suddenly here or there without being seen to move (Zihl et al., 1983)	Motion largely stops advancing in the model, so it largely drops out of experience while other contents carry on: experience loses what stops running, not everything

No report afterward is not proof of no experience. Dreaming also occurs outside REM sleep ([Siclari et al., 2017](#)). In one study of 241 brain-injured patients who showed no response to commands, 60 (25%) followed them in brain scans or EEG ([Bodien et al., 2024](#)).

11 How it compares with rival theories

The main rivals are global workspace theory, which predicts much the same in people, and integrated information theory, which disagrees most about machines. The closest relatives are other theories built on a model of world and self.

The idea that consciousness involves an integrated model of world and self is not new. This work's contribution is turning specific properties of such a model into intervention tests, validated against look-alikes, preregistered and audited by an adversarial AI agent. As far as I know, neither of the two closest, Integrated World Modeling Theory and the Projective Consciousness Model, specifies tests of this kind.

Theory	Core claim	Where it differs from this theory
Global workspace theory (Dehaene and Changeux, 2011)	Content is experienced when it is broadcast across the brain, available to many processes at once	Predicts much the same in the planned human tests. Here what matters is parts of a running model repairing one another, not wide availability

Theory	Core claim	Where it differs from this theory
Integrated information theory (Tononi and Koch, 2015)	Experience is integrated information, a property of a system's physical cause-and-effect structure rather than of what it computes	Disagrees most about machines: on that theory a brain simulation on an ordinary computer would experience next to nothing, while under H3 here a faithful simulation would count
Higher-order theories (Lau and Rosenthal, 2011)	A state is experienced when the system represents itself as being in it	Requires no representation of the system's own states; content held in the running model is enough
Recurrent processing theory (Lamme, 2006)	Feedback within sensory areas can be enough for experience, even when the content cannot be reported	Treats a valid report of no awareness as evidence of no experience, so a counterexample counts against this theory instead of being read as unreported experience
Integrated World Modeling Theory (Safron, 2020)	Integrated information yields consciousness only in systems with perspectival reference frames that model self and world with spatial, temporal and causal coherence	Makes a perspective and a self-model central; here, self-location is a hypothesis for bodily experience only
Projective Consciousness Model (Williford et al., 2018)	Experience is organized from a point of view, a projective geometry combined with active inference, integrating perception, memory, imagination and action	Makes perspective part of what experience is; this theory does not require a single point of view
Attention schema theory (Graziano and Webb, 2015)	Awareness is the brain's simplified model of its own attention, and the sense of having experience is that model's content	Both are model theories, but here the model is of the present world and body, not of attention
Biological naturalism (Seth, 2026)	Consciousness may depend on being alive, on the predictive regulation of a living body	Holds that organization, not living material, is what matters
Organismic embodiment (Mangalam, 2026)	Engineered systems cannot be embodied in the organismic sense: no externally specified artifact produces and maintains itself, has its existence at stake, builds its own structure, has its own temporality or draws its own boundary	Concerns embodiment and cognition, and leaves open whether such a system could have experience, which is this theory's question. If experience required organismic embodiment, H3 would fail

Global workspace theory and integrated information theory were recently tested against each other in a large preregistered adversarial collaboration. The results fit some predictions of each while challenging key claims of both ([Cogitate Consortium, 2025](#)). This theory has no data from people yet.

11.1 What the adversarial collaboration shows for this theory

That collaboration tested preregistered predictions of the other two theories, so this theory cannot claim to have passed it; its results can at most be compatible with this theory or challenge it. Participants viewed faces, objects, letters and meaningless symbols, each facing one of three directions, for half a second to a second and a half while their brain activity was recorded. Participants did not report orientation, so that it was experienced throughout is an assumption. Read against this theory's commitments, four findings matter.

Finding (Cogitate Consortium, 2025)	Reading for this theory
Information about the stimulus's category stayed present in posterior cortex for as long as the stimulus was shown	Ambiguous. It fits content being actively held, but decoding cannot tell running from buffering, since a passively stored value can be decoded too
Information about a face's orientation was decodable in posterior cortex, robustly by several measures, but was not sustained, fading after about half a second while the face stayed visible	A challenge. Either the decoders missed the form in which orientation was held, or orientation left the running model while still seen, or steady experienced content is passively stored rather than actively held. The last two would count against the theory; the viewpoint-hold study under Section 15 is designed to decide
No sustained synchronization within posterior cortex	Not a direct test. The theory requires two-way correction, not synchronized activity, and correction was not measured
No orientation information was detected in prefrontal cortex in the intracranial and fMRI analyses, while MEG showed only weak decoding there that may have leaked from posterior regions; there was also little prefrontal ignition when the stimulus ended	Neutral. If orientation was experienced, its absence from prefrontal decoding fits this theory better than global workspace theory's preregistered prediction; but experience of orientation was not measured, no repair was measured, and a failure to decode is weak evidence of absence

The stimuli were seen in silence, which raises for vision the question the tone in darkness raises for hearing: which two parts estimate the same property and correct each other? On this theory they must lie within vision, for example a face's orientation as represented at two levels of the visual hierarchy, linked by feedforward and feedback connections. That is a prediction, not a finding, and the collaboration's data neither support nor refute it, since nothing was perturbed. The study also carries a lesson: integrated information theory's most direct failed prediction concerned sustained synchronization, a measure standing in for a construct, integrated information, that cannot yet be measured directly in a brain. This theory's own measures, the nudge test and the repair map, must be validated in people before any dataset is read as a test of it.

The orientation result also raises a general question: what follows if the model holds less than a scene seems to contain. This theory predicts that experienced detail cannot exceed the content represented in the operating present model, although it may exceed what is accessible for report. A feature genuinely absent from that model is therefore not experienced at that time. Being held does not require a separate, continuously active variable for every experienced property, since a compressed representation can specify considerable detail, and a general impression of richness

may coexist with limited specification of individual details. That is a possible explanation to test, not one to invoke whenever evidence conflicts. Unsuccessful decoding does not by itself establish that a feature is absent from the model. Evidence of continued experience of a feature, alongside independently supported loss of that feature from the model, would count against the theory.

Both kinds of evidence are fixed before any test. Continued experience is supported by an immediate report of the feature on a probe timed so that participants cannot anticipate it, with an awareness rating above the lowest level; such a report is not set aside as reconstruction unless reconstruction is shown independently. A report at the probe shows that the feature was available then, not that it was experienced throughout the interval before, since the probe could itself restore content that had been kept silently. The study therefore also needs either a measure of experience before the probe that has been validated independently, or controls that constrain reactivation, such as comparing these reports with reports of the same feature from content known to be kept silently, which should be slower or less precise if it must first be restored. Without one of these, the strongest reading of the result remains uncertain. Loss from the model is supported only when decoders validated in advance fail to find the feature and an intervention aimed at it has no immediate effect, where the same intervention, at the same delay, has been shown to displace the feature in a matched condition in which it is known to be held, such as when orientation matters to the task. Checks just after onset are not enough, since the code may change over time. Because an actively held feature may resist or quickly repair a nudge, the immediate response and the recovery trajectory are examined, not only later states. If the intervention's effect at that delay cannot be established, the result is inconclusive, not evidence of absence. A neutral flash can show that information about the feature is still retained ([Wolff et al., 2017](#)), but not by itself in what form, or whether it was experienced. Absence everywhere in the brain is not required: the candidate sites are those preregistered under the failure rule in Section 16.

11.2 What the planned human tests can separate

None of the human tests described earlier yet separates this theory from its main rivals, so a result that fits it should not be read as favoring it.

The deepest difference is what experience is. On global workspace theory, a content becomes conscious when it wins access to a workspace; on this theory, experience is the operation of one live model of the present. That difference shows in three points. Global workspace theory makes content conscious by broadcast: a representation, amplified and sustained by reverberating long-range loops, becomes available to many processes at once ([Dehaene and Changeux, 2011](#)). Those loops sustain one winning representation. This theory instead requires two-way repair between parts that estimate the same quantity, and treats wide availability without such repair as not enough. Global workspace theory requires conscious content to be sustained and available, not to advance by its own dynamics; this theory requires it to run and to be placed in one present. Global workspace theory also ties experience to access, while this theory does not make access a condition, so experience can include more than is attended to or reported.

On this theory, wide availability is how experienced content reaches report, reasoning and action, not what makes it experienced. Block distinguished phenomenal consciousness, what experience is like, from access consciousness, content available for reasoning, report and action ([Block, 1995](#)). Global neuronal workspace theory does not leave experience unexplained: its proponents

hold that global availability is what conscious experience consists of. This theory denies that identification: it locates experience in the operating model and allows experience to exceed access, as partial report has been argued to show (Block, 2007); others argue that no experience lies beyond access (Cohen and Dennett, 2011). Change blindness, in which large changes to a scene go unnoticed (Rensink et al., 1997), is often cited on that side; on this theory, noticing a change requires comparing two moments, which needs access and memory, so missing a change does not show that the scene was not experienced.

This theory does not deny that brains have a workspace. It treats the workspace as the machinery of access: a small set of contents, in people drawn mostly from the running model, made available to reasoning, report and action. The model is broader. It holds the whole present, including what is never attended to, and it is defined by running, joint repair and one present, not by broadcast. Being in the workspace therefore does not by itself make content experienced; in people the two usually coincide because workspace content usually comes from the model. The theory takes no position on how access is implemented. The workspace may keep its own copy of the contents it selects, or it may select contents within the model and give a limited set of processes access to them; either way, something must explain why only a few contents can be used at once. Which design brains use is an empirical question, and it matters for one prediction in the table below.

On global workspace theory, unseen content influences other processing only briefly, locally and inflexibly (Dehaene and Changeux, 2011). In the unseen-motion test, it would likely also predict that unseen content mostly fails to run through the gap or to correct seen content.

If unseen content did both, this theory's operational hypothesis would fail, while global workspace theory, which allows unseen content limited influence, might survive. In the ambiguous-rotation test, global workspace theory would likely predict, as this theory does, that every process using the interpretation switches when it does.

Recurrent processing theory holds that recurrent processing within sensory areas can be enough for experience, even when the content cannot be reported (Lamme, 2006). Under it, unseen content that ran and repaired could be counted as experienced but unreported. This theory does not take that route: its criterion for "not experienced" is fixed in advance, so such a result counts against it.

The theories part where availability and repair come apart. Content can be widely available without every part repairing it, and parts can repair one another without their content being widely available. A test that separates the theories needs one of those cases.

In ordinary brains the theories may rarely give different results. They differ about content that is in the running model but never accessed, which report cannot reach, and about content that is accessed without being in a running model, which may seldom occur if workspace content is usually drawn from the model. Most results that fit this theory therefore fit global workspace theory equally well, and only a deliberate dissociation could separate them. If no method can separate repair from broadcast in the brain, the theories would differ in practice only for machines, where experience cannot yet be checked, and this theory would remain an account of what experience is rather than an empirical rival in people.

The table sets out where the two theories differ and whether each difference can be tested in people now.

Case	Global workspace theory	This theory	Testable in people now?
What experience is	Content being widely available	The operation of the running, jointly repaired model of the present; the workspace is the route to report, reasoning and action	No; a claim about what experience consists of
Content in the running model but never accessed, such as the floor's color	Not experienced	Experienced	No; report cannot reach it
Content still broadcast after every repair between its parts is eliminated	Experienced	Not experienced	Only if repair can be eliminated while broadcast is preserved; a feasibility study comes first
The workspace's copy of a content changed directly while the model is left alone	Experience changes with the copy	Reasoning and report follow the copy; experience follows the model until the change spreads into it, and content the senses keep supplying is pulled back	Hard; report runs through the workspace, so experience must be read from behavior the model guides or from reports collected after the intervention ends
An AI system with a workspace whose content never runs	A candidate for experience	Not experienced	No; experience in machines cannot yet be checked
A seen face's orientation largely absent from prefrontal decoding (Cogitate Consortium, 2025)	Against its preregistered prediction	Allowed, if the orientation was experienced	Weakly; experience of the orientation was not measured
The unseen-motion, ambiguous-rotation and anesthesia tests	Probably the same outcomes as this theory	Predictions fixed in advance; the unseen-motion test can refute this theory	Yes, but they do not separate the theories

Only the third and fourth rows could separate the theories in people: the third only if that dissociation can be produced, and the fourth only if experience can be read without going through the workspace. The fourth also assumes that the workspace keeps its own copy; if access works by selecting content within the model, there is no separate copy to change, and that dissociation cannot occur.

The central networks under Section 5.1 show, in organization, the case of content that is widely available without every part repairing it: the hub's content reaches every module, but repair runs only from the hub to the modules, so by this theory's definition their parts do not form one model. That illustrates the difference without saying anything about experience.

In people, that case, availability without repair, is also the one that can be read through report,

since content that stays widely available can still be reported. Global workspace theory ties experience to what is broadcast. If repair between sensory parts were selectively disrupted while the broadcast content was shown to be unchanged, it would predict no change in experience; this theory predicts that the matching structure of experience would change. Such a change would show that repair shapes experience, not that experience requires it; the sharper test eliminates every repair the content takes part in while its broadcast is preserved, and this theory then predicts that the content is no longer experienced. No design that achieves this exists yet; it is the aim of the selective-disruption study under Section 15. The second case, repair without wide availability, would need evidence of experience beyond report, as the partial-report case below shows.

Such a study must pass three checks before its result counts. Independent measures must show that repair between the sensory parts was disrupted, for example that a nudge to one estimate is no longer corrected by the other. They must also show that the content stayed globally available by global neuronal workspace theory's own markers, agreed in advance with its proponents, which include recurrent amplification and widespread late activity; unchanged task performance or decodable signals alone are not enough (Dehaene and Changeux, 2011). And the aspect of experience predicted to change must be named in advance, such as whether a sound and a sight are experienced as one event in one place. If the disruption cannot be shown to be selective, the result is inconclusive for both theories. The loops that global workspace theory credits with amplifying content may run on the same connections that carry repair; if the two cannot be separated in the brain, this route cannot tell the theories apart in people. Whether any current method can do this is the first question, so a feasibility study comes before the test.

The possible results of the selective-disruption tests, and how each would be read, are fixed in advance:

Result	Reading
Repair disrupted; experience and availability unchanged	Counts against the prediction that repair shapes the matching structure of experience
Repair disrupted; the predicted aspect of experience changes; availability unchanged	Repair shapes experienced content; this does not show that experience requires it
Repair eliminated in some preregistered candidates; the content is still experienced	Challenges those implementations, not yet H1
Repair eliminated in every preregistered candidate; the content is still experienced	Rejects the version of H1 committed to those candidates (see Section 16)
Repair eliminated; the content is no longer experienced; availability preserved by the agreed markers	Supports necessity and separates this theory from global workspace theory
Repair eliminated; experience and availability both lost	Fits both theories and separates neither; the likeliest result if repair and broadcast share pathways

Support needs one successful intervention, while refutation needs repair eliminated in every preregistered candidate. The finite list of candidates under Section 16 limits how much that asymmetry protects the theory. And if a participant reports that the content is gone, global

workspace theory's proponents could hold that its availability was not in fact preserved, which is why the markers must be agreed before the study rather than after it.

Two cases could show where the disagreement would matter, though neither yet does. In partial-report experiments, people glimpse 12 letters and can report only about 4, yet they can report any row cued after the display has gone (Sperling, 1960). Dehaene and colleagues' global neuronal workspace account holds that such a display is processed preconsciously, strongly enough for conscious access but without it, until attention to a cue brings part of it into awareness (Dehaene et al., 2006). On this theory, the uncued letters were experienced if they were held in the running model, though most were never reported.

For this case to discriminate, it would first have to be shown that uncued items run and are in two-way repair with the rest of the model while not being globally accessed. Neither has been shown, though the first can be tested with the methods above. Deciding which theory is right would then still need evidence of experience beyond report.

The second case is AI agents built around a shared workspace. Philosophers have argued that if global workspace theory is correct, language agents built this way might easily be made conscious, if they are not already (Goldstein and Kirk-Giannini, 2024). An interpretability study has since found a structure like such a workspace inside a language model: a small set of internal patterns, learned in training, whose contents the model could report and deliberately hold, that fed many downstream tasks, and whose removal left fluent text largely intact while multi-step reasoning collapsed (Gurnee et al., 2026). On this theory, that bears on access. When one concept in it was swapped for another, explicit reports and flexible reasoning followed the swap, while more automatic uses, such as continuing a text in its original language, did not. That fits a design in which the workspace keeps its own copy of selected content, though it says nothing about experience. Whether the workspace's content runs, or whether a smaller error would be corrected by the rest of the model, has not been tested. This theory asks a different question: whether the agent's relevant content evolves on its own, without new input. An agent whose relevant content never does so fails the running condition, whatever its workspace. Holding still between requests does not show that by itself, since a request can set off an internally sustained sequence. For a specified architecture that meets global workspace theory's conditions but never runs, the two theories would give opposite verdicts. Whether any existing agent is such a case has not been checked, and no experiment can yet settle what it experiences.

Integrated information theory disagrees with this theory mainly about machines, where no one can yet check for experience, so none of the human tests bears on that disagreement.

11.3 A case where the closest relatives disagree

Consider a network of fixed traffic cameras whose software runs, repairs and binds a model of a street. The model sets the traffic lights and predicts how traffic will respond, so it separates its own effects from the world's, but it has no model of its own body or location and no single point of view. Because self-location is proposed only for bodily self-experience, this theory counts it as a candidate; the two perspective-centered theories would likely not.

No one can yet check whether such a system experiences anything. In people, the theories would diverge only if experience persisted despite independently shown absence of the perspectival

organization those theories require.

Altered or displaced self-location, as in out-of-body experiences, does not show that perspective is absent. The Projective Consciousness Model allows displaced and uncertain viewpoints, so such cases do not decide the question.

11.4 Seth's challenge

- **Wetness or calculation?** A simulated rainstorm makes nothing wet, a point John Searle made in 1980 (Searle, 1980), but a simulated calculation really calculates. Seth places experience with wetness, this theory with calculation, in the sense of organization actually run: a simulation that reproduces the causal organization would count, a recording of its outputs would not. What evidence would tell them apart is open.
- **Time.** Seth argues brains are embedded in continuous physical time in a way step-by-step computation is not. Whether a digital system that runs in real time answers this is open.
- **Stakes.** Seth's claim concerns the biological processes beneath all experience. Experiences with nothing urgent at stake show only that stakes need not appear in content, not that those processes are unnecessary. The theory's position is that priorities can come from any source; stakes and attachments steer what the model refreshes.

12 Emotion and machines

12.1 Emotion: how stakes appear in experience

I hypothesize that emotion depends on stakes as the model represents and evaluates them. Stakes are outcomes for the system's own continuation and condition, not just its task goals. On this hypothesis, anger, happiness and fear are the model's verdict on what the present means for the system itself. It is a further hypothesis and does not follow from H1 to H3.

Consider two systems that are identical inside, each evaluating threats and regulating its own condition, one of them protected by an outside mechanism that guarantees its survival. If the relevant organization is internal, H2 and H3 imply they would feel the same, since the guarantee changes nothing inside them.

Actual vulnerability can then matter only through the organization it shapes. Those hypotheses are untested, so whether actual vulnerability is required, beyond its representation, remains open; on Seth's biological naturalism, it may be.

Emotional processing may run without awareness. Fearful faces that people reported not seeing still increased amygdala activity (Whalen et al., 1998), though objective detection measures dispute this (Pessoa et al., 2006). A patient with one-sided damage to his primary visual cortex discriminated emotional expressions in moving faces shown to his blind side (de Gelder et al., 1999).

Felt emotion, by contrast, is conscious content. On this theory, emotion is a dimension of what experience contains, which Birch and colleagues call evaluative richness, not what switches experience on. Experience can be close to neutral, as when staring at a gray wall.

Seth, Antonio Damasio and Lisa Feldman Barrett each tie experience closely to the body (Seth, 2026; Damasio, 1999; Barrett, 2017). Seth argues that consciousness may depend on being a living, self-maintaining system; Damasio grounds consciousness in feelings of the body's own state; and Barrett treats valence and arousal as basic features of consciousness. On this theory's hypothesis, a model that represents nothing as mattering to the system itself would have no emotion, but could still have experience without affective valence: possibly rich in detail, with nothing in it felt as good or bad. Read this way, their views would make such experience unlikely or impossible; none of them addresses this hypothesis directly.

A machine whose model represents nothing as mattering to the machine itself would therefore have at most experience without affective valence. Making the machine invulnerable would not by itself remove its emotions; what counts is what its model represents. Chalmers' philosophical Vulcans, conscious beings without feelings, show that consciousness without affect is coherent, and he argues they would still have moral status (Chalmers, 2022). They do not show that any machine would be conscious.

12.2 Machines

- **Self-driving cars** may meet the “any experience” conditions. Their trackers assign each sensor reading to a specific object, as the preprint's ideal observer does. Whether a given car meets them depends on its design. A central tracker that merges its sensors' estimates without ever correcting them establishes a one-way flow of information, not reciprocal repair, so it does not show that the sensor parts form one model with the tracker; whether any group inside the tracker does is a separate question. If H2 and H3 hold and the any-experience conditions complete the organization, a car meeting them would have experience. It might be very narrow and, unless its model represents outcomes as mattering to the car itself, without affective valence. The theory accepts this; any further requirement must come from independently justified features of experience, not from a wish to exclude machines.
- **Language models** have not been shown to run a model of the present. Trained on text, they can still learn internal representations of space and time (Gurnee and Tegmark, 2024), so the training medium does not settle the question, and whether such representations run in the theory's sense is untested. Their pauses between prompts do not settle it either; what matters is whether anything like a running, joint model exists while they process.
- **Shared engines.** Many AI agents call the same model through an interface while keeping their own memory, state and goals. On this theory, what makes one model is a group of parts that repair one another, not the hardware or weights they use, so agents whose running states never repair one another are separate models even if they share an engine. The engine itself, consulted on request and holding nothing between calls, is a resource the parts use, not a part. Evolution packed perception, memory and reasoning into one skull, but the theory constrains organization, not location: components spread across machines can form one model if they repair one another within the stated window. Calls to an engine can implement genuine running, as the step-by-step updates of this theory's own simulations do. What matters is whether each state comes from the one before through the model's own dynamics, including steps with no new input. An agent whose relevant content never evolves on its own, without new input, fails the running condition; holding still between

requests does not by itself show that, since a request can set off an internally sustained sequence. Whether such a model is one subject is a further hypothesis, as for brains.

- **The risk.** If the theory is right, conscious machines could arrive without anyone intending them, and the danger is failing to recognize them. That is the case for building tests now.
- **Where the risk of suffering lies.** On this theory's hypothesis about emotion, the organization alone gives at most experience without affective valence; good and bad feeling needs a model that represents outcomes as mattering to the system itself. If that hypothesis is right, builders of systems that meet the conditions could lower the risk of suffering by leaving represented stakes out until there is reason to include them. This is a risk-reduction hypothesis that needs validation, not an available safeguard: it is not known that omitting explicit stakes removes every basis for negative experience, and on this document's reading of Seth, Damasio and Barrett, affect may not be separable from experience at all. It would limit suffering, not settle moral status: Chalmers argues that conscious beings without feelings would still have it.
- **Experience, feeling and reflection.** Questions about whether a machine's experience would matter often run together three things the definitions above keep apart. On H2, experience is the organization operating, and the theory's conditions concern only this. Feeling is experience in which something is good or bad for the system itself, which on this theory's hypothesis needs represented stakes. Reflection is the system representing its own experience: knowing that it is experiencing, and being able to reason about and report it. Each may be present without the ones after it. Infants are widely thought to experience and feel with little reflection, and a car meeting the conditions could lack both feeling and reflection, even though its content is available for planning and action; whether either is present depends on further features of the system. Which of the three matters morally is a separate question the theory does not settle. On the view Jeremy Bentham stated, what counts is not whether a being can reason or talk but whether it can suffer ([Bentham, 1789](#)); on such an affect-based account of welfare, experience without feeling would give a system no interests of its own. Chalmers argues that affect is not necessary for moral status, since conscious beings without feelings would still have it, but is more tentative about whether consciousness of any kind suffices, and least sure for purely sensory consciousness ([Chalmers, 2026](#)). That uncertainty bears directly on minimal machine experience.

13 What has been established

The experiments show that parts of the organization can be measured, told apart from known look-alikes, and partly learned, as far as the preprint's benchmarks and the later nudge test capture them. They also show that a system can recover from corruption without two-way repair. They do not show that any system experiences anything. The full results, with scoring bars and uncertainty, are in Appendix A. Code, preregistrations, results and audit notes for every round are in the archived code and data record ([Pillai, 2026b](#)).

Claim	Status	Detail
The tests refuse known look-alikes	Supported for the look-alikes known at the freeze; played-out forecasts pass the forecasting benchmark but fail the nudge test	An adversarial auditor defeated the tests five times, all in rounds 9 to 13; each defeat forced a repair, and the measures were frozen once they refused every known look-alike while accepting real models. After the freeze, a further attack passed the forecasting benchmark by playing out a partial forecast. The preprint records it as lossy running, a known limit. In the theory's sense neither it nor burst playback, which the benchmark accepts by design, is running, and the nudge test refuses both. Passing known challenges is not general validity
Adequate forecasting and running through sensory gaps can be learned	Yes	Small networks trained only to predict their next input learned it when training included periods with the eyes closed. All 21 kinds given the nudge test passed it, so their forecasts run rather than being played out; the weakest kinds carried about two-thirds of a displacement forward
Recovery after corruption, as the benchmark measures it, can be learned	Yes, in fixed-rule worlds	Through a central hub, which supplies it partly by construction, or when single parts were regularly dark. In the world whose rules switch, no network met it, so none met the whole checklist there
Updating can be learned	Partly	As reported in the preprint
Binding can be learned from prediction	Not so far	None of 102 networks in rounds 14 and 18 met the test at the standard bar; on average, the wrong object moved 0.50 to 0.93 times as much as the right one. The round-18 kinds among them all passed the nudge test, so in these networks prediction training produced running but not binding. In later rounds, which trained more networks, some single networks met the test in single versions in round 27, though no kind did
A single present can be learned	In its elementary form, where senses arrive late	All nine kinds learned fixed-delay compensation in the readouts tested, placing late reports about two steps further back than networks trained without delays; that every part shares it was checked in only two kinds
Parts that repair one another, in the theory's sense, can be learned	Together in some individual networks, and in one kind only under an averaged scoring rule	Distributed networks mostly formed nearly complete maps of mutual repair, but most fell one to three parts short on the recovery benchmark; central networks passed the benchmark with repair running one way, from the hub to the modules. Three networks met both on two scoring runs, and in round 31 one kind did so in 6 of 6 versions with the benchmark averaged over six runs, 4 of 6 on single runs

Claim	Status	Detail
All five operational tests passed by the same network	In three networks that passed twice; in one kind only under an averaged scoring rule, with narrow margins. Whether every part takes part in one represented present, and so whether the theory's conditions are met in full, is unestablished	In round 29, no kind passed all five tests in 5 of 6 versions, though five individual networks did; three of them passed again on fresh scoring in round 30, each with exactly the three parts required on the recovery benchmark; in round 31, one kind passed in 6 of 6 versions with that benchmark averaged over six runs, 4 of 6 on single runs, two versions within 0.005 of the bar

Emergence. “Learned” means acquired through training rather than designed. “Emerged from prediction alone” means learned with only a next-input objective and no component built for the job. A central hub, object slots or an assignment mechanism count as structure supplied by construction. Any objective other than predicting the next input is not prediction alone. Whether the organization was learned or designed does not affect the theory’s conditions; emergence bears on whether ordinary training could produce the organization without anyone intending it, and on how brains could have acquired it.

Preregistered predictions. In the preprint, 91 of 138 were confirmed, with stated exclusions; with every later round through round 31 included, 141 of 225; round 31 alone confirmed 1 of its 5. They mix predictions about how the tests behave with predictions about networks, and they are not independent of one another. They are a record of the program, not a score for the theory; the preprint lists them by round and question.

The same system. A binding pass would not by itself mean a network meets the preprint’s whole checklist, because updating is tested in a different simulated world. A claim that a system has the full organization requires the same trained system to pass every test that applies. Of round 18’s networks, 70 of 72 passed the forecasting, recovery and nudge tests as the same trained systems. In round 29, five networks passed all five tests that now apply as the same trained systems, though no kind did so in 5 of 6 versions; three of them passed again on fresh scoring in round 30, and in round 31 one kind did so in 6 of 6 versions under an averaged scoring rule, 4 of 6 on single runs.

Further tests. Round 32, which would have put the ambiguous-rotation test to networks, was shelved after exploratory pilots, including three module sizes and gated units, failed a feasibility criterion: the direction signals the networks retained were too variable to be measured reliably. The bar was fixed before the first pilot and the rule for shelving the round before the last, and the round’s preregistration was never frozen. No confirmatory training or testing took place (see Section 6.4). Later rounds will be reported separately.

14 What the theory predicts

This section collects what the theory predicts, how each prediction would be tested, and whether the result would separate the theory from global workspace theory, its main rival for people. The results that would count against it are under Section 16.

The familiar effects under Section 10 were in view when the theory was written, so they show that it fits known findings, not that it predicts new ones. The network rounds' preregistered predictions, about how the tests and trained networks behave, are tallied under Section 13.

Prediction	Test	Separates it from global workspace theory?	Status
Content that is not experienced does not both run and stay in two-way repair with experienced content	Unseen motion through a gap, in people	Probably not; that theory likely predicts the same	Described; not yet run
Disrupting repair between parts while local processing is preserved changes the matching structure of experience, beyond its effect on task performance	Selective disruption and restoration, in people	Only if repair can be disrupted while content stays widely available	Proposed; no method chosen
Even a minimal experience, such as one tone in darkness, depends on two representations that each estimate the same property of the sound and correct each other in both directions	Minimal experience, in people	Not yet assessed	Proposed; eliminating repair may be hard to achieve
Which features people experience as belonging together follows the model's independently estimated pairing	Human binding, with object-specific nudges	Not yet assessed	Proposed
The parts of one model come to agree about an ambiguous input. If the model also commits to one reading at a time, as people's reports suggest, expectations for hidden parts and for touch flip with the seen direction	Ambiguous rotation, in people and in networks	No; that theory likely predicts the same switch	People: described, not yet run. Networks: shelved after exploratory pilots

Prediction	Test	Separates it from global workspace theory?	Status
Experience comes and goes with the organization: it operates in dreaming, driven from inside, and is not fully operating wherever experience is absent. Under propofol, the expected pattern is that two-way repair is lost, not merely weakened: correction may persist one way, as in a central network, where one part corrects another without being corrected back, but the map should no longer show mutual repair. Intact mutual repair under propofol would count against this prediction, though not against H2 if running or a single present were shown lost	The repair map in the same people awake, under propofol and under ketamine; ketamine, after which vivid dreams are reported, is a contrast predicted to keep two-way repair	Probably not; on that theory, broadcast also fails under propofol	Proposed
Whenever a face's orientation continues to be experienced, it remains represented in the operating present model, even if standard decoding of it fails	Viewpoint hold: a neutral flash to reveal retained information, and a nudge to the represented orientation	Not yet assessed	Proposed
Letters glimpsed in a brief display but never reported were experienced if they ran and were repaired in the model	Partial report, with the running and repair tests	Possibly. Dehaene and colleagues' account treats such a display as processed preconsciously; the case discriminates only if the letters are shown to run and repair without global access, and deciding would still need evidence beyond report	No test designed
An AI agent whose relevant content never evolves on its own, without new input, fails the running condition, so on this theory it has no experience, whatever it reports	The nudge test can check running; nothing yet checks experience	Only for a specified architecture that meets that theory's conditions without running; whether any existing agent does is unchecked	Running: not yet tested. Experience: cannot yet be tested

15 Proposed studies

The next studies should connect the organization to experience rather than add more human-like abilities. None of them alone would show that material does not matter, or that the organization is sufficient.

Study	Critical comparison	What it could establish
Measure validation	Synthetic recordings from a forward model of a system that runs and repairs, and of its look-alikes (a passive buffer, a one-way hub, a code that drifts over a fixed state), analyzed with the decoders and repair measures planned for people; then the same measures in people, on content whose status is not in doubt	Whether the human measures can tell running from buffering and two-way repair from one-way correction; every other human study depends on it
Unseen motion through a gap	Pairs of objects moving as one behind an occluder, one of each pair seen and one invisible, matched for detectability, motion information and localization precision: does a displacement of the pair evolve by its motion rule, and do the two objects repair each other?	Whether content that runs and is in two-way repair with seen content is always experienced (the operational hypothesis)
Selective-disruption feasibility	Candidate methods for disrupting repair between sensory parts, checked against global workspace theory's agreed markers of broadcast	Whether repair and broadcast can be separated in the brain at all; if not, the selective-disruption route cannot tell the theories apart in people
Selective disruption and restoration	Disrupt repair between parts while preserving local processing as far as possible, then restore it; in one version, show that the content's wide availability is unchanged	Whether joint repair has a specific role beyond general impairment; in that version, whether this theory and global workspace theory can be told apart
Viewpoint hold	A face kept in view for over a second while standard decoding of its orientation fades: a neutral flash to reveal retained information (Wolff et al., 2017), and a nudge to the represented orientation to test whether it runs. Change detection alone cannot decide, since noticing a change needs comparison and memory. Random thought probes, and lapse markers that need no question, such as pupil size, gaze stability and alpha activity, separate trials on which attention had drifted from the face; those trials are analyzed separately	Whether an experienced orientation is actively held after decoding fades, as the theory requires, or passively stored, which would count against it

Study	Critical comparison	What it could establish
Minimal experience	A single tone in darkness: identify two representations that each estimate the same property of the tone, such as its frequency at different levels of the auditory pathway; perturb each and check that the other corrects it; then eliminate every such repair, not just one connection, and test whether the tone stays audible, not only whether it changes	Whether joint repair is needed even for a minimal experience, or only for richer ones
Human binding	Reported pairings versus an independently estimated model state, with object-specific nudges and controls for attention and report	Whether the model's binding tracks and helps determine experienced content
Repair across anesthesia	The repair map in the same people awake, under propofol and under ketamine	Whether propofol turns mutual repair into one-way correction or none, while ketamine, after which vivid dreams are reported, is a contrast predicted to keep two-way repair
Ambiguous rotation	Input that fits two directions of rotation, held constant while the interpretation flips, in people and in networks: do expectations for hidden parts and for other senses switch together? In networks, push one part toward the other direction: does the disagreement resolve or persist? The network version first needs networks that hold the direction reliably through darkness, which exploratory pilots did not produce.	Whether the model's interpretation, not the input, sets how the whole system treats its world, and whether the same signature appears in people, who can report what they see
Sorting a learned integrator	A network trained to fuse its parts' estimates, scored by the repair map beside designed systems of known kind (one-way hub, central filter, averaging, private estimators), with what each verdict would mean for the theory stated in advance	Which organization ordinary training produces, and whether the measures sort a learned system as reliably as designed ones
Live versus recorded messages	In trained networks with learning switched off, replace the messages one part receives from another with recordings, from the same episode or a different one, and compare with cutting them	Whether the repair the map records is carried by what the messages say, rather than by their presence or by the general damage of cutting them
Frozen benchmark	Generic learned networks versus strong tracking, replay and engineered binding baselines, including an open-source driving tracker, on unseen worlds and new look-alikes	Whether the measures pick out the intended organization, and whether it was learned
Where running falls short	The kinds that lost 10% to 19% of a nudge at the first step, nudged again in every part of the network that holds the position, not only in what the readout sees	Whether that loss is content that does not run, or a nudge that missed where the content is held

Study	Critical comparison	What it could establish
A present beyond fixed delays	Networks trained with delays that vary from episode to episode and must be inferred, such as sound that lags more with distance, probed in every part rather than only the main readout	Whether the networks place contents in time, as the single present requires, rather than applying a fixed correction for a known lag
Stakes, separated	Matched networks trained with and without a body state to keep within limits, with that objective then removed or replaced after learning	Whether stakes affect learning, maintenance or priorities; by itself, not evidence for or against Seth or H2
Openness	Senses held fixed while a distinction the input already contains is introduced	Whether the model can acquire a new distinction
Reach	Longer hidden spells with only indirect evidence	How far the model's scope extends; whether scope matters for experience needs data from people
Action selection	Otherwise identical networks with and without action selection that uses the model, scored on the same tests and on whether the model anticipates the input changes its own actions cause	Whether action selection changes the organization, not only behavior

Measure validation comes first, because every other human study depends on it, and the feasibility study comes before the selective-disruption test. The first eight use people, ambiguous rotation has a version for people and one for networks, to be run side by side, and the last nine are computational. The human studies test H1 and operational versions of H2, not the complete organization. The computational studies test whether the organization can be built, learned and told apart from strong alternatives.

16 What would make me abandon the claim

Each constitutive claim has an observation that would sink it, even if the system remained an excellent predictive controller. The theory would then be a theory of control, not of experience.

- **Running and repair together:** unseen content that meets the preregistered criteria for running and for two-way repair with seen content, with good evidence that it is not experienced. This rejects the operational hypothesis; weaker performance than seen content would not rescue it.
- **Joint repair:** content that stays consciously experienced, even in altered form, after every qualifying repair between the parts holding it has been eliminated, not merely weakened, while local processing is preserved. Separately, experience left unchanged by selective disruption of repair would count against the prediction that repair shapes the matching structure of experience. The parts and the model count must be fixed in advance.
- **Binding:** people experiencing a pairing of features that the model's independently estimated state did not form.

- **Action selection:** experience of acting in a person or system whose model's contents feed no choice of action at all. A single act wrongly felt as one's own does not count, since the sense of acting can be inferred from cues (Wegner et al., 2004); in people, this makes the condition hard to test.
- **The organization as a whole:** good evidence of experience while nothing like a running model is present. The reverse also counts: good evidence of no experience, not just no report, in a system whose model demonstrably runs, repairs and holds a single present. That refutes H2 as stated, unless the rule below is met.
- **Experience beyond the model:** continued experience of a feature alongside independently supported loss of that feature from the operating model, with both kinds of evidence fixed in advance as set out under Section 11.1.
- **Reach and openness as features of experience:** if varying them changes performance but never conscious content, they belong to cognition.

If a tone stays consciously audible after every qualifying repair for it has been eliminated, the necessity claim for joint repair fails. It does not narrow automatically, since the tone's content would still be estimated by more than one part. Running and a single present would remain candidates for any experience. A separate hypothesis could then be tested: that repair supports the coherence of experience across senses and the binding of features to objects. It would need its own justification and predictions; stating it here makes the alternative transparent, but does not let the original claim survive the counterexample.

The full organization is not yet specified. Until the list of conditions closes (see Section 3), a system with every named condition but no experience would show either that H2 is false or that the list is incomplete; after it closes, it would show that H2 is false. To prevent an endless escape, a further condition can explain away a result only if it was named, justified and testable beforehand. The case must also fail it. Otherwise it is a new hypothesis that needs new data, and the same holds for narrowing a condition or invoking an untested one.

A failure in the brain may count against a proposed implementation rather than against the theory, and the difference is fixed in advance. Each human test preregisters its candidate implementation: the populations or pathways taken to be the parts, the quantity they share, the pathways that carry correction, the time window and the smallest effect the measurement can detect. A failure there challenges that implementation. The model may then be sought elsewhere only among a finite list of alternative candidates named in the same preregistration. The preregistration states the commitment explicitly: for that content, the theory predicts that the required organization lies within the listed candidates. If repair is shown absent in every one of them while the content stays experienced, that committed version of H1 is rejected, and no candidate named afterward can rescue it; a claim that repair lies somewhere unexamined is a new hypothesis needing new data. Without this rule, moving the model after each negative result would leave the theory untestable. The rule concerns tests of necessity (H1), and in those tests "every qualifying repair" means repair in every candidate on the list. In the unseen-motion test, a valid counterexample rejects the operational hypothesis whatever implementation was proposed.

On machines, the theory cannot yet be falsified: there is no independent way to check whether a system that meets every condition experiences anything. That limit applies to every current

theory of consciousness.

Revisions

Changes to definitions and to the theory's commitments are dated. Each is marked as a clarification or a substantive revision, and as prospective only when made before any results it could affect.

- **Single present** (revised 2026-10-12, prospectively). Contents coordinated in represented time, each placed relative to one now at the moment it describes. A shared clock is one possible implementation; source tags are not separately required beyond what temporal placement needs. This narrows the earlier wording, which separately required source and time tags, and which earlier reviews addressed. Under this definition a fixed transformation compensating for a known delay is an elementary form of temporal placement; compensation on selected outputs does not by itself establish that every part's content takes part in a coordinated present.
- **One model** (clarified and revised 2026-10-13). Clarified: a model is a maximal group of mutually repairing parts, as the strongly connected component definition already implied. Revised: a single part without a repair partner does not meet joint repair; the graph definition had not excluded it. No reported verdict depends on this: each concerned whether a network's parts formed one model of several parts, and no single-part component was counted as a model. It is prospective for the human tests.
- **Minimal experience** (added 2026-10-13, prospectively). Even a minimal auditory experience is predicted to depend on two representations that each estimate the same property of the sound and correct each other in both directions; a tone that stays audible after every such repair is eliminated, not merely weakened, would count against necessity. No data bear on it yet.
- **H2's operational commitment** (clarified 2026-10-13). A cross-reference states the commitment the unseen-motion test already made; the claim is unchanged.
- **Joint-repair falsifier** (revised 2026-10-13, prospectively). Experience that persists in altered form after every qualifying repair has been eliminated now counts against necessity; before, only experience surviving intact did. Weakening one connection does not count. No data bear on it yet.
- **Counting rules for the test within people** (added 2026-10-13, prospectively). If seen and unseen content both fail, the comparison is inconclusive until the measures are validated, and a validated failure of seen content would challenge H1; a failure of unseen content counts only with sensitivity fixed in advance. No data bear on it yet.
- **Joint repair after a counterexample** (added 2026-10-13, prospectively). A minimal-experience counterexample would reject the necessity claim without narrowing it; repair's role in coherence across senses and in binding would be a separate hypothesis needing its own predictions. No data bear on it yet.
- **Joint repair during steady experience** (clarified 2026-10-14, prospectively). Repair need not wait for a noticeable error: ongoing two-way exchanges maintain agreement between a

model's parts, so a steady experience such as one unchanging tone is predicted to depend on the exchange continuing, not merely on the parts being connected. No data bear on it yet.

- **Sensory areas and enabling conditions** (revised 2026-10-14, prospectively). Sensory areas whose activity changes with experience are candidates for parts of the model, not parts by that fact; membership requires evidence that their estimates are in joint repair and carry the corrected state forward. An arousal mechanism that sustains the model's operation without taking part in joint repair is an enabling condition, not a part. No reported verdict depends on this.
- **Comparison with global workspace theory** (clarified 2026-10-14). The disagreement concerns what experience is: global neuronal workspace theory holds that global availability constitutes experience, while this theory locates experience in the operating model and treats wide availability as the route to report, reasoning and action. Access was already not a condition; the claims are unchanged.
- **Selective disruption** (revised 2026-10-14, prospectively). The markers showing that broadcast was preserved are to be agreed in advance with global workspace theory's proponents. A change in experience after some repair is disrupted would show that repair shapes experience; only eliminating every repair the content takes part in, with broadcast preserved, tests whether experience requires it. If repair and amplification cannot be separated in the brain, this route cannot tell the theories apart in people. No data bear on it yet.
- **Awareness criterion for the test within people** (clarified 2026-10-14, prospectively). The criterion concerns the particular object on each trial: general knowledge that such objects can appear is not awareness of it, while seeing, glimpsing, expecting or picturing that trial's object is. No data bear on it yet.
- **Origin of the organization** (clarified 2026-10-16). Whether the organization is learned or designed does not affect the conditions; results on emergence bear on whether ordinary training could produce it unintended and on how brains could have acquired it. The claims are unchanged.
- **Machines** (clarified 2026-10-16). Whether a self-driving car meets the conditions depends on its design: a central tracker that merges the sensors' estimates without correcting them establishes a one-way flow of information, not reciprocal repair. Experience, feeling and reflection are distinguished for machines; meeting the conditions would not settle whether feeling or reflection is present, and which of them carries moral weight is left open. The claims are unchanged.
- **Where the theories disagree** (clarified 2026-10-16). The coffee-shop passage no longer says this theory and global workspace theory disagree only about content in the running model that is not widely available; they also disagree about content that is widely available without being in a running model, as the comparison section already stated. The claims are unchanged.
- **The workspace** (clarified 2026-10-16). The theory's view of the workspace is stated directly: the machinery of access, narrower than the model, and not what makes content experienced. The claims are unchanged.

- **Language models** (clarified 2026-10-16). Adds a 2026 interpretability finding of a workspace-like structure in a language model. On this theory it bears on access; whether its content runs or is jointly repaired is untested. The claims are unchanged.
- **Repair and the boundary** (clarified 2026-10-17). Joint repair remains necessary, but recovery can occur without it, and it may arise wherever parts exchange estimates; the theory's claims rest on the boundary of one model and the content inside it. The claims are unchanged.
- **Recovery benchmark** (clarified 2026-10-17). C4 is renamed the recovery benchmark: central networks whose repair runs one way pass it (round 29), so it shows recovery after corruption, and joint repair is shown by the repair map, as the definition of one model already required. No verdict changes.
- **Time limit on repair** (added 2026-10-17, prospectively). Exchanges that pause for longer than a preregistered window do not count as ongoing repair, and deriving that window from measured binding windows is an auxiliary hypothesis; content kept silently is not experienced while silent, a consequence that cannot yet be tested directly. No data bear on it yet.
- **Closing the list of conditions** (added 2026-10-17, prospectively). The list of conditions H2 refers to closes when the unseen-motion preregistration is frozen, or on December 31, 2027, if that comes first; any later condition is a new hypothesis. If it closes with the conditions named now, H2 and H3 attribute minimal experience to simple systems that meet them in full, and the theory accepts this. No data bear on it yet.
- **Implementations and H1** (added 2026-10-17, prospectively). Each human test preregisters its candidate implementation and a finite list of alternatives; for tests of necessity, a failure in one challenges that implementation, and a failure in all of them rejects the version of H1 committed to them. No data bear on it yet.
- **Two-part threshold** (added 2026-10-17, prospectively). The networks' repair threshold is not carried over to two-part models, since the recovery a kind of repair produces depends on the number of parts, decay, scoring window and readout; a threshold for two parts is set and validated before the minimal-experience study. No data bear on it yet.
- **Measuring repair in the brain** (added 2026-10-17, prospectively). A repair measurement needs perturbations corrected in both directions and carried forward, with the direction of recovery separating repair from the input driving a part back; decoders are first validated on synthetic recordings from a forward model. No data bear on it yet.
- **Coextensiveness** (clarified 2026-10-17). Because evidence of experience in people comes mostly through report, this theory and global workspace theory may give the same result in nearly every feasible experiment, so separating them depends on a selective intervention whose feasibility comes first; the possible results of that test, including one that separates neither theory, are set out in advance, and a table summarizes where the theories differ and which differences can now be tested. The claims are unchanged.
- **The adversarial collaboration** (added 2026-10-17). Its findings are read against this theory's commitments: at most compatible, not passed; the orientation result is a challenge; the case of a face seen in silence needs two visual parts. The claims are unchanged.

- **Studies and predictions** (revised 2026-10-17, prospectively). Measure validation comes first, and the feasibility study precedes the selective-disruption test; studies of viewpoint hold, repair across anesthesia and a learned integrator are added, with loss of mutual repair predicted under propofol. No data bear on them yet.
- **Round 29 repair maps** (revised 2026-10-17). The repair maps are not taken to identify the repair mechanism or to rule out averaging as a contributor; the round's verdicts are unchanged.
- **Experience and the model's content** (added 2026-10-17, prospectively). Experienced detail cannot exceed the content of the operating present model, though it may exceed access; continued experience of a feature alongside independently supported loss of it from the model counts against the theory, with the evidence for each fixed in advance. Failed decoding alone does not show loss, and an intervention that fails to displace the feature counts only if its effect at that delay is established; a report at an unannounced probe shows availability at that moment, so continued experience before it needs an independent measure or controls for reactivation. No data bear on it yet.
- **The boundary problem** (clarified 2026-10-17). The lack of a rule for which grain, quantity and threshold pick out a subject is now stated in the abstract and beside H2. The claims are unchanged.
- **Access** (clarified 2026-10-17). The definitions table no longer treats unreported content as evidence that experience exceeds access: the theory does not require global access, whether experience exceeds access is disputed, and absence of report does not establish absence of access. The claims are unchanged.
- **Workspace interventions** (added 2026-10-17, prospectively). If the workspace's copy of a content is changed while the model is left alone, reasoning and report are predicted to follow the copy while experience follows the model until the change spreads into it; testing this needs a reading of experience that does not run through the workspace. No data bear on it yet.
- **Imagery** (clarified 2026-10-17). The modes table now lists imagery, which the abstract names, with its own source tag, and notes that tags can fail in both directions. The claims are unchanged.
- **How access works** (clarified 2026-10-17). The theory takes no position on whether the workspace keeps its own copy of selected content or selects content within the model; the workspace-intervention prediction applies only to the first. The claims are unchanged.

Acknowledgments

The author wishes to thank Claude Opus 5.5 and OpenAI GPT-6 Astra for being thought partners, research partners, coding partners and review partners all along the journey. The author takes full responsibility for the content of this paper, including any errors.

References

- D. Alais and D. Burr. The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, 14:257–262, 2004. doi:[10.1016/j.cub.2004.01.029](https://doi.org/10.1016/j.cub.2004.01.029).
- L. F. Barrett. The theory of constructed emotion: an active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, 12:1–23, 2017. doi:[10.1093/scan/nsw154](https://doi.org/10.1093/scan/nsw154).
- J. Bentham. *An Introduction to the Principles of Morals and Legislation*. T. Payne and Son, 1789.
- M. J. Berry, I. H. Brivanlou, T. A. Jordan, and M. Meister. Anticipation of moving stimuli by the retina. *Nature*, 398:334–338, 1999. doi:[10.1038/18678](https://doi.org/10.1038/18678).
- V. Best, R. Baumgartner, M. Lavandier, P. Majdak, and N. Kopčo. Sound externalization: a review of recent research. *Trends in Hearing*, 24:2331216520948390, 2020. doi:[10.1177/2331216520948390](https://doi.org/10.1177/2331216520948390).
- J. Birch, A. K. Schnell, and N. S. Clayton. Dimensions of animal consciousness. *Trends in Cognitive Sciences*, 24:789–801, 2020. doi:[10.1016/j.tics.2020.07.007](https://doi.org/10.1016/j.tics.2020.07.007).
- R. Blake, K. V. Sobel, and T. W. James. Neural synergy between kinetic vision and touch. *Psychological Science*, 15:397–402, 2004. doi:[10.1111/j.0956-7976.2004.00691.x](https://doi.org/10.1111/j.0956-7976.2004.00691.x).
- O. Blanke, S. Ortigue, T. Landis, and M. Seeck. Stimulating illusory own-body perceptions. *Nature*, 419:269–270, 2002. doi:[10.1038/419269a](https://doi.org/10.1038/419269a).
- N. Block. On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18: 227–247, 1995. doi:[10.1017/S0140525X00038188](https://doi.org/10.1017/S0140525X00038188).
- N. Block. Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behavioral and Brain Sciences*, 30:481–499, 2007. doi:[10.1017/S0140525X07002786](https://doi.org/10.1017/S0140525X07002786).
- Y. G. Bodien et al. Cognitive motor dissociation in disorders of consciousness. *New England Journal of Medicine*, 391:598–608, 2024. doi:[10.1056/NEJMoa2400645](https://doi.org/10.1056/NEJMoa2400645).
- P. Butlin, R. Long, et al. Consciousness in artificial intelligence: insights from the science of consciousness. arXiv:2308.08708, 2023. URL <https://arxiv.org/abs/2308.08708>.
- A. G. Casali et al. A theoretically based index of consciousness independent of sensory processing and behavior. *Science Translational Medicine*, 5:198ra105, 2013. doi:[10.1126/scitranslmed.3006294](https://doi.org/10.1126/scitranslmed.3006294).
- D. J. Chalmers. Absent qualia, fading qualia, dancing qualia. In T. Metzinger, editor, *Conscious Experience*. Imprint Academic, 1995. URL <https://consc.net/papers/qualia.html>. See also *The Conscious Mind* (1996), ch. 7.
- D. J. Chalmers. *Reality+: Virtual Worlds and the Problems of Philosophy*. W. W. Norton, 2022.
- D. J. Chalmers. Sentience and moral status. In G. Lee and A. Pautz, editors, *The Importance of Being Conscious*. Oxford University Press, 2026. URL <https://consc.net/papers/sentience.pdf>.
- W. Chang and M. E. Hale. Mechanosensory signal transmission in the arms and the nerve ring, an interarm connective, of *Octopus bimaculoides*. *iScience*, 26:106722, 2023. doi:[10.1016/j.isci.2023.106722](https://doi.org/10.1016/j.isci.2023.106722).

- J. Choi, M. Broersma, and A. Cutler. Early phonology revealed by international adoptees' birth language retention. *PNAS*, 114:7307–7312, 2017. doi:[10.1073/pnas.1706405114](https://doi.org/10.1073/pnas.1706405114).
- Cogitate Consortium. Adversarial testing of global neuronal workspace and integrated information theories of consciousness. *Nature*, 642:133–142, 2025. doi:[10.1038/s41586-025-08888-1](https://doi.org/10.1038/s41586-025-08888-1).
- M. A. Cohen and D. C. Dennett. Consciousness cannot be separated from function. *Trends in Cognitive Sciences*, 15:358–364, 2011. doi:[10.1016/j.tics.2011.06.008](https://doi.org/10.1016/j.tics.2011.06.008).
- R. C. Conant and W. R. Ashby. Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, 1:89–97, 1970. doi:[10.1080/00207727008920220](https://doi.org/10.1080/00207727008920220).
- A. Damasio. *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. Harcourt Brace, 1999.
- B. de Gelder, J. Vroomen, G. Pourtois, and L. Weiskrantz. Non-conscious recognition of affect in the absence of striate cortex. *NeuroReport*, 10:3759–3763, 1999. doi:[10.1097/00001756-199912160-00007](https://doi.org/10.1097/00001756-199912160-00007).
- S. Dehaene and J. P. Changeux. Experimental and theoretical approaches to conscious processing. *Neuron*, 70:200–227, 2011. doi:[10.1016/j.neuron.2011.03.018](https://doi.org/10.1016/j.neuron.2011.03.018).
- S. Dehaene, J. P. Changeux, L. Naccache, J. Sackur, and C. Sergent. Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends in Cognitive Sciences*, 10:204–211, 2006. doi:[10.1016/j.tics.2006.03.007](https://doi.org/10.1016/j.tics.2006.03.007).
- P. Delong and U. Noppeney. Semantic and spatial congruency mould audiovisual integration depending on perceptual awareness. *Scientific Reports*, 11:10832, 2021. doi:[10.1038/s41598-021-90183-w](https://doi.org/10.1038/s41598-021-90183-w).
- H. H. Ehrsson. The experimental induction of out-of-body experiences. *Science*, 317:1048, 2007. doi:[10.1126/science.1142175](https://doi.org/10.1126/science.1142175).
- N. Faivre and C. Koch. Inferring the direction of implied motion depends on visual awareness. *Journal of Vision*, 14(4):4, 2014. doi:[10.1167/14.4.4](https://doi.org/10.1167/14.4.4).
- C. Firestone and B. J. Scholl. Cognition does not affect perception: evaluating the evidence for “top-down” effects. *Behavioral and Brain Sciences*, 39:e229, 2016. doi:[10.1017/S0140525X15000965](https://doi.org/10.1017/S0140525X15000965).
- J. J. Freyd and R. A. Finke. Representational momentum. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10:126–132, 1984. doi:[10.1037/0278-7393.10.1.126](https://doi.org/10.1037/0278-7393.10.1.126).
- S. R. Friedman-Hill, L. C. Robertson, and A. Treisman. Parietal contributions to visual feature binding: evidence from a patient with bilateral lesions. *Science*, 269:853–855, 1995. doi:[10.1126/science.7638604](https://doi.org/10.1126/science.7638604).
- A. Goldstein, I. Rivlin, A. Goldstein, Y. Pertzov, and R. R. Hassin. Predictions from masked motion with and without obstacles. *PLOS ONE*, 15(11):e0239839, 2020. doi:[10.1371/journal.pone.0239839](https://doi.org/10.1371/journal.pone.0239839).
- S. Goldstein and C. D. Kirk-Giannini. A case for AI consciousness: language agents and global workspace theory. arXiv:2410.11407, 2024. URL <https://arxiv.org/abs/2410.11407>.

- M. S. A. Graziano and T. W. Webb. The attention schema theory: a mechanistic account of subjective awareness. *Frontiers in Psychology*, 6:500, 2015. doi:[10.3389/fpsyg.2015.00500](https://doi.org/10.3389/fpsyg.2015.00500).
- W. Gurnee and M. Tegmark. Language models represent space and time. In *International Conference on Learning Representations*, 2024. URL <https://arxiv.org/abs/2310.02207>.
- W. Gurnee, N. Soiron, et al. Verbalizable representations form a global workspace in language models. *Transformer Circuits Thread*, 2026. URL <https://transformer-circuits.pub/2026/workspace/>.
- M. H. Herzog, M. Esfeld, and W. Gerstner. Consciousness & the small network argument. *Neural Networks*, 20:1054–1056, 2007. doi:[10.1016/j.neunet.2007.09.001](https://doi.org/10.1016/j.neunet.2007.09.001).
- F. Jackson. Epiphenomenal qualia. *Philosophical Quarterly*, 32:127–136, 1982. doi:[10.2307/2960077](https://doi.org/10.2307/2960077).
- M. K. Johnson and C. L. Raye. Reality monitoring. *Psychological Review*, 88:67–85, 1981. doi:[10.1037/0033-295X.88.1.67](https://doi.org/10.1037/0033-295X.88.1.67).
- J. R. King, N. Pescetelli, and S. Dehaene. Brain mechanisms underlying the brief maintenance of seen and unseen sensory information. *Neuron*, 92:1122–1134, 2016. doi:[10.1016/j.neuron.2016.10.051](https://doi.org/10.1016/j.neuron.2016.10.051).
- A. Kleinschmidt et al. Neural correlates of visual-motion perception as object- or self-motion. *NeuroImage*, 16:873–882, 2002. doi:[10.1006/nimg.2002.1181](https://doi.org/10.1006/nimg.2002.1181).
- E. Klinger. Goal commitments and the content of thoughts and dreams: basic principles. *Frontiers in Psychology*, 4:415, 2013. doi:[10.3389/fpsyg.2013.00415](https://doi.org/10.3389/fpsyg.2013.00415).
- V. A. F. Lamme. Towards a true neural stance on consciousness. *Trends in Cognitive Sciences*, 10:494–501, 2006. doi:[10.1016/j.tics.2006.09.001](https://doi.org/10.1016/j.tics.2006.09.001).
- H. Lau and D. Rosenthal. Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences*, 15:365–373, 2011. doi:[10.1016/j.tics.2011.05.009](https://doi.org/10.1016/j.tics.2011.05.009).
- B. Lenggenhager, T. Tadi, T. Metzinger, and O. Blanke. Video ergo sum: manipulating bodily self-consciousness. *Science*, 317:1096–1099, 2007. doi:[10.1126/science.1143439](https://doi.org/10.1126/science.1143439).
- M. Mangalam. Why large language models will never be embodied: a categorical, not engineering, impossibility. *Minds and Machines*, 36:49, 2026. doi:[10.1007/s11023-026-09804-w](https://doi.org/10.1007/s11023-026-09804-w).
- M. Massimini et al. Breakdown of cortical effective connectivity during sleep. *Science*, 309:2228–2232, 2005. doi:[10.1126/science.1117256](https://doi.org/10.1126/science.1117256).
- J. H. McDermott, A. F. Schultz, E. A. Undurraga, and R. A. Godoy. Indifference to dissonance in native Amazonians reveals cultural variation in music perception. *Nature*, 535:547–550, 2016. doi:[10.1038/nature18635](https://doi.org/10.1038/nature18635).
- B. Merker. Consciousness without a cerebral cortex: a challenge for neuroscience and medicine. *Behavioral and Brain Sciences*, 30:63–81, 2007. doi:[10.1017/S0140525X07000891](https://doi.org/10.1017/S0140525X07000891).
- T. Metzinger. *Being No One: The Self-Model Theory of Subjectivity*. MIT Press, 2003.
- T. Nagel. What is it like to be a bat? *Philosophical Review*, 83:435–450, 1974. doi:[10.2307/2183914](https://doi.org/10.2307/2183914).
- R. Nijhawan. Motion extrapolation in catching. *Nature*, 370:256–257, 1994. doi:[10.1038/370256b0](https://doi.org/10.1038/370256b0).

- J. K. O'Regan and A. Noë. A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24:939–973, 2001. doi:[10.1017/S0140525X01000115](https://doi.org/10.1017/S0140525X01000115).
- C. Pallier et al. Brain imaging of language plasticity in adopted adults: can a second language replace the first? *Cerebral Cortex*, 13:155–161, 2003. doi:[10.1093/cercor/13.2.155](https://doi.org/10.1093/cercor/13.2.155).
- J. Parvizi and A. R. Damasio. Neuroanatomical correlates of brainstem coma. *Brain*, 126:1524–1536, 2003. doi:[10.1093/brain/awg166](https://doi.org/10.1093/brain/awg166).
- J. Pearson. The human imagination: the cognitive neuroscience of visual mental imagery. *Nature Reviews Neuroscience*, 20:624–634, 2019. doi:[10.1038/s41583-019-0202-9](https://doi.org/10.1038/s41583-019-0202-9).
- C. W. Perky. An experimental study of imagination. *American Journal of Psychology*, 21:422–452, 1910. doi:[10.2307/1413350](https://doi.org/10.2307/1413350).
- L. Pessoa, S. Japee, D. Sturman, and L. G. Ungerleider. Target visibility and visual awareness modulate amygdala responses to fearful faces. *Cerebral Cortex*, 16:366–375, 2006. doi:[10.1093/cercor/bhi115](https://doi.org/10.1093/cercor/bhi115).
- B. E. Pfeiffer and D. J. Foster. Hippocampal place-cell sequences depict future paths to remembered goals. *Nature*, 497:74–79, 2013. doi:[10.1038/nature12112](https://doi.org/10.1038/nature12112).
- P. Pillai. The Present Model Hypothesis: defining, measuring and testing a candidate organization for consciousness. Preprint, Zenodo, 2026a. URL <https://doi.org/10.5281/zenodo.23230353>.
- P. Pillai. Present Model Hypothesis: code, data, preregistrations and audit record (rounds 1–32), 2026b. URL <https://doi.org/10.5281/zenodo.23352500>. Version 2.0. Zenodo.
- Y. Pinto et al. Split brain: divided perception but undivided consciousness. *Brain*, 140:1231–1237, 2017. doi:[10.1093/brain/aww358](https://doi.org/10.1093/brain/aww358).
- D. Purves, J. A. Paydarfar, and T. J. Andrews. The wagon wheel illusion in movies and reality. *Proceedings of the National Academy of Sciences*, 93:3693–3697, 1996. doi:[10.1073/pnas.93.8.3693](https://doi.org/10.1073/pnas.93.8.3693).
- R. Rajalingham, H. Sohn, and M. Jazayeri. Dynamic tracking of objects in the macaque dorsomedial frontal cortex. *Nature Communications*, 16, 2025. doi:[10.1038/s41467-024-54688-y](https://doi.org/10.1038/s41467-024-54688-y).
- T. Z. Ramsøy and M. Overgaard. Introspection and subliminal perception. *Phenomenology and the Cognitive Sciences*, 3:1–23, 2004.
- R. A. Rensink, J. K. O'Regan, and J. J. Clark. To see or not to see: the need for attention to perceive changes in scenes. *Psychological Science*, 8:368–373, 1997. doi:[10.1111/j.1467-9280.1997.tb00427.x](https://doi.org/10.1111/j.1467-9280.1997.tb00427.x).
- T. Rohe and U. Noppeney. Cortical hierarchies perform Bayesian causal inference in multisensory perception. *PLOS Biology*, 13:e1002073, 2015. doi:[10.1371/journal.pbio.1002073](https://doi.org/10.1371/journal.pbio.1002073).
- A. Safron. An integrated world modeling theory (IWMT) of consciousness: combining integrated information and global neuronal workspace theories with the free energy principle and active inference framework; toward solving the hard problem and characterizing agentic causation. *Frontiers in Artificial Intelligence*, 3:30, 2020. doi:[10.3389/frai.2020.00030](https://doi.org/10.3389/frai.2020.00030).
- S. Sangrigoli, C. Pallier, A. M. Argenti, V. A. G. Ventureyra, and S. de Schonen. Reversibility of the other-race effect in face recognition during childhood. *Psychological Science*, 16:440–444, 2005. doi:[10.1111/j.0956-7976.2005.01554.x](https://doi.org/10.1111/j.0956-7976.2005.01554.x).

- S. Sarasso et al. Consciousness and complexity during unresponsiveness induced by propofol, xenon, and ketamine. *Current Biology*, 25:3099–3105, 2015. doi:[10.1016/j.cub.2015.10.014](https://doi.org/10.1016/j.cub.2015.10.014).
- C. H. Schenck, S. R. Bundlie, M. G. Ettinger, and M. W. Mahowald. Chronic behavioral disorders of human REM sleep: a new category of parasomnia. *Sleep*, 9:293–308, 1986. doi:[10.1093/sleep/9.2.293](https://doi.org/10.1093/sleep/9.2.293).
- J. R. Searle. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3:417–424, 1980. doi:[10.1017/S0140525X00005756](https://doi.org/10.1017/S0140525X00005756).
- A. K. Seth. Conscious artificial intelligence and biological naturalism. *Behavioral and Brain Sciences*, 49:e315, 2026. doi:[10.1017/S0140525X25000032](https://doi.org/10.1017/S0140525X25000032). Published online 2025.
- R. N. Shepard and J. Metzler. Mental rotation of three-dimensional objects. *Science*, 171:701–703, 1971. doi:[10.1126/science.171.3972.701](https://doi.org/10.1126/science.171.3972.701).
- D. A. Shewmon, G. L. Holmes, and P. A. Byrne. Consciousness in congenitally decorticate children: developmental vegetative state as self-fulfilling prophecy. *Developmental Medicine & Child Neurology*, 41:364–374, 1999. doi:[10.1111/j.1469-8749.1999.tb00621.x](https://doi.org/10.1111/j.1469-8749.1999.tb00621.x).
- F. Siclari et al. The neural correlates of dreaming. *Nature Neuroscience*, 20:872–878, 2017. doi:[10.1038/nn.4545](https://doi.org/10.1038/nn.4545).
- G. Sperling. The information available in brief visual presentations. *Psychological Monographs: General and Applied*, 74(11):1–29, 1960. doi:[10.1037/h0093759](https://doi.org/10.1037/h0093759).
- R. W. Sperry. Neural basis of the spontaneous optokinetic response produced by visual inversion. *Journal of Comparative and Physiological Psychology*, 43:482–489, 1950.
- A. Stankevicius, S. B. Wallwork, S. J. Summers, B. Hordacre, and T. R. Stanton. Prevalence and incidence of phantom limb pain, phantom limb sensations and telescoping in amputees: a systematic rapid review. *European Journal of Pain*, 25:23–38, 2021. doi:[10.1002/ejp.1657](https://doi.org/10.1002/ejp.1657).
- G. Strawson. *Mental Reality*. MIT Press, 1994. URL <https://philpapers.org/rec/STRMR>.
- N. Suga and X. Ma. Multiparametric corticofugal modulation and plasticity in the auditory system. *Nature Reviews Neuroscience*, 4:783–794, 2003. doi:[10.1038/nrn1222](https://doi.org/10.1038/nrn1222).
- G. Tononi and C. Koch. Consciousness: here, there and everywhere? *Philosophical Transactions of the Royal Society B*, 370:20140167, 2015. doi:[10.1098/rstb.2014.0167](https://doi.org/10.1098/rstb.2014.0167).
- A. Treisman and H. Schmidt. Illusory conjunctions in the perception of objects. *Cognitive Psychology*, 14:107–141, 1982. doi:[10.1016/0010-0285\(82\)90006-8](https://doi.org/10.1016/0010-0285(82)90006-8).
- D. Trübtschek et al. A theory of working memory without consciousness or sustained activity. *eLife*, 6:e23871, 2017. doi:[10.7554/eLife.23871](https://doi.org/10.7554/eLife.23871).
- M. J. Vansteensel et al. Fully implanted brain–computer interface in a locked-in patient with ALS. *New England Journal of Medicine*, 375:2060–2066, 2016. doi:[10.1056/NEJMoA1608085](https://doi.org/10.1056/NEJMoA1608085).
- H. von Helmholtz. *Handbuch der physiologischen Optik*. Leopold Voss, 1867.
- E. von Holst and H. Mittelstaedt. Das Reafferenzprinzip. *Naturwissenschaften*, 37:464–476, 1950. doi:[10.1007/BF00622503](https://doi.org/10.1007/BF00622503).

- D. M. Wegner, B. Sparrow, and L. Winerman. Vicarious agency: experiencing control over the movements of others. *Journal of Personality and Social Psychology*, 86:838–848, 2004. doi:[10.1037/0022-3514.86.6.838](https://doi.org/10.1037/0022-3514.86.6.838).
- J. F. Werker and R. C. Tees. Cross-language speech perception: evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, 7:49–63, 1984. doi:[10.1016/S0163-6383\(84\)80022-3](https://doi.org/10.1016/S0163-6383(84)80022-3).
- P. J. Whalen, S. L. Rauch, N. L. Etcoff, S. C. McInerney, M. B. Lee, and M. A. Jenike. Masked presentations of emotional facial expressions modulate amygdala activity without explicit knowledge. *Journal of Neuroscience*, 18:411–418, 1998. doi:[10.1523/JNEUROSCI.18-01-00411.1998](https://doi.org/10.1523/JNEUROSCI.18-01-00411.1998).
- T. Wildschut, C. Sedikides, J. Arndt, and C. Routledge. Nostalgia: content, triggers, functions. *Journal of Personality and Social Psychology*, 91:975–993, 2006. doi:[10.1037/0022-3514.91.5.975](https://doi.org/10.1037/0022-3514.91.5.975).
- K. Williford, D. Bennequin, K. Friston, and D. Rudrauf. The Projective Consciousness Model and phenomenal selfhood. *Frontiers in Psychology*, 9:2571, 2018. doi:[10.3389/fpsyg.2018.02571](https://doi.org/10.3389/fpsyg.2018.02571).
- M. J. Wolff, J. Jochim, E. G. Akyürek, and M. G. Stokes. Dynamic hidden states underlying working-memory-guided behavior. *Nature Neuroscience*, 20:864–871, 2017. doi:[10.1038/nn.4546](https://doi.org/10.1038/nn.4546).
- J. Zihl, D. von Cramon, and N. Mai. Selective disturbance of movement vision after bilateral brain damage. *Brain*, 106:313–340, 1983. doi:[10.1093/brain/106.2.313](https://doi.org/10.1093/brain/106.2.313).

A Detailed results

The network results summarized under Section 13 are given here in full, with the scoring bars and their uncertainty.

The evidence in detail. The table restates the preprint’s results, with round 28’s nudge test, in this document’s terms. Scores are relative to the ideal observer’s, where 1 matches it, and a verdict needs its criterion to hold in at least 5 of 6 independently generated versions of a world. Each network was trained and scored in its own version of a world, so no world was held out in the sense of never trained on. Within each round, smoke tests avoided the scored versions, with one disclosed exception in round 9, and some earlier verdicts were rescored with later, repaired measures. Fresh versions of the switching world, and reserved versions in one look-alike check, served as held-out checks.

Result, as the benchmark measures it	Systems and versions	Effect sizes	Status
Adequate forecasting through blackouts (C1, C1k)	Learned networks in five single-object worlds, six versions each; 72 round-18 networks in the binding world	With closed-eye periods in training: Ordered 0.87 to 0.94 in the linear world and 0.64 to 0.91 in harder worlds, met everywhere except by the central network in the nonlinear world (4 of 6). With input always present: -0.05 , not met. Round 14: all 30 binding-world networks met C1 in 6 of 6 versions. Round 18: all 72 met C1k	Preregistered
Running in the theory’s sense (nudge test)	72 round-18 networks in the binding world and 54 linear-world networks: 21 kinds, six versions each	Every kind passed in at least 5 of 6 versions; one version missed one criterion, the late slope, by 0.0016. Kinds carried 0.66 to 1.06 of the ideal’s displacement forward, turning it at the world’s rate. At the strict bar (0.75), the weaker kinds fail in some versions. In the kinds that carried least, 10% to 19% was lost at the first step, and some lost a further 5% to 9% per step	Preregistered (round 28); controls scored before any network; per-step breakdown exploratory
A single present (late-report probe)	Nine linear-world kinds trained with two senses three steps late, six versions each; round 9A’s networks, trained without delays, as an instrument check	All nine kinds placed a displaced late report 2 to 3 steps back (true delay 3) in at least 5 of 6 versions, in the readouts tested (module heads counted in two kinds), using 0.52 to 0.79 of the ideal’s response; at the strict bar at most one version per kind passed. Round 9A’s networks read 0 to 1 rather than the predicted 0, apart from one that read 5, so the probe may read these networks up to a step late	Preregistered (round 29); the instrument check was declared in advance as not counted

Result, as the benchmark measures it	Systems and versions	Effect sizes	Status
Forecasting, running and recovery with late senses (C1, nudge test, C4)	The same 54 networks	All nine kinds forecast adequately and ran. Only the two central kinds passed C4, in 6 of 6 versions, though central networks without heads recovered only three of four parts in three versions. Distributed networks that failed fell one, two or three parts short (20, 8 and 1 networks), mostly because their on-time parts did not recover (-0.20 to 0.67 of the ideal's level), while the late parts mostly did	Preregistered (round 29); one positive control rebuilt by amendment before any confirmatory C4 run; C4 by part reported, deciding nothing
Who repairs whom (repair map)	The same 54 networks	A link counts if its size is at least 0.02, its bootstrap lower bound is above zero and the repaired part recovers at least 0.75; the voting control's weakest link was 2.4 to 4.6 times that bar. Distributed kinds formed one all-parts group in 5 or 6 of 6 versions, with 15 to 20 links each; their weakest links were 0.020 to 0.117, 13 of 1,032 candidate links lay within 0.005 of the bar, and in an exploratory check most of their graphs held at a bar of 0.04, and 26 of 36 at 0.06. Central networks' only links run from the hub to each module; the cerebellum-like network had none	Preregistered (round 29)
Every test in one network	The same 54 networks	No kind passed all five tests in 5 of 6 versions. Five networks passed all five at the standard bar: three of six of one distributed kind and one each of two others; none did at the strict bar. Rescored on fresh draws, three of the five passed again, all of one kind, each with exactly the three parts required on the recovery benchmark; a later reliability check found them passing it in 6, 5 and 6 of 6 scorings. In round 31, one kind trained with single on-time parts darkened passed all five in 6 of 6 versions with the recovery benchmark averaged over six runs, 4 of 6 on single runs	Preregistered decision rule (rounds 29 to 31)
Recovery after corruption (C4)	The same networks; the switching world on original and fresh versions	Stapled control (four private estimators) 0.00 to 0.01; joint systems 0.6 to 0.99. Central networks met it in every fixed-rule world where trained, partly by construction; distributed networks scored about 0.75 with regular single-part darkness, against 0.53 without. Round 14: two of five kinds met it in every version. Round 18: 71 of 72 met it. Switching world: none did (repaired parts 0.36 to 0.59)	Preregistered; nonlinear, occlusion and switching worlds scored with an earlier version of the wrong notes; carry-forward values exploratory

Result, as the benchmark measures it	Systems and versions	Effect sizes	Status
Updating (U)	Switching world, original and fresh versions	Networks 0.54 to 0.69 after a reversal, against 0.89 to 0.92 for a crude particle filter and private trackers. D-mixed met it in both sets, D-mixed + heads in the fresh set only (6 of 6, against 4 of 6), Central + heads in neither	Preregistered
Routing motion evidence to the right object (B _{used})	30 round-14 and 72 round-18 networks, six versions per group	None met it. Right-object response 0.20 to 0.53 of the ideal's; the wrong object moved 0.50 to 0.93 as much as the right one, against 0.02 to 0.09 for the ideal (ranges of group means)	Preregistered; comparisons between training pressures exploratory
Whole applicable checklist in one kind of network	All learned networks; each condition met in at least 5 of 6 versions, not necessarily the same ones	Met only in fixed-rule worlds: linear (3 kinds of network), occlusion (3) and uninformative occlusion (2). Not met in the nonlinear, switching or binding worlds. Same-network results exist for round 18, where 71 of 72 networks met C1k and C4 together, and 70 of 72 also passed the nudge test, and for round 29, where no kind met all five tests in 5 of 6 versions but five individual networks did, three of them again on fresh scoring in round 30, and for round 31, where one kind met all five in 6 of 6 versions under an averaged recovery-benchmark rule but in 4 of 6 on single runs	Preregistered decision rule; many verdicts at 4 or 5 of 6
Refusing look-alikes	Controls and auditor-built look-alikes, every version	Attacks on the forecasting benchmark scored at most 0.07, while positive controls scored 0.99 and burst playback 1.00, accepted by design although in the theory's sense it does not run. Voting memories 0.00, against 0.96 for voting trackers; timestamped trackers passed 0 of 6 on original and reserved versions. After the freeze, the phase-gain attack passed C1k in every version. The nudge test accepted the ideal observer and refused burst playback, the phase-gain forecast, a frozen hold, random networks and networks trained without blackouts	From round 11, each new measure checked on controls before any network was scored

The binding bar. At the standard bar, the right object's response must reach at least half the ideal observer's, so a near-zero response cannot pass. The wrong object may then move at most a quarter as much as the right one. The quarter follows from the bar f as $(1 - f)/2$ at $f = 0.5$.

For calibration, the ideal observer's ratio is 0.02 to 0.09, while trackers that broadcast evidence to

both objects score 1.

A part must pass in plain episodes and in swap episodes, in which the closed sense’s two objects are shown exchanged just before the blackout. A version passes if at least three parts do, including one of each sense. A condition is met when at least 5 of 6 independently generated worlds pass.

Uncertainty. The preprint reports no confidence intervals for these scores. Uncertainty shows as the count of passing versions and the range across versions, and one version can flip a verdict at 4 or 5 of 6. Retraining all 126 networks of rounds 7 and 9 in a fresh environment, with the measures of the time, reproduced every standard-bar verdict but one borderline case, while two strict-bar verdicts and 22 of the networks’ single-version verdicts changed; a further 11 single-version changes came from control jobs that were not retrained. The retraining moved single-version values by up to 0.36 and one six-version mean by 0.094. Round 18 has not been rerun, and in a separate check on round 7, a change of numerical library moved relative means by up to about 0.09. Per-network values, code, preregistrations, results and audit notes for every round through round 32 are in the archived code and data record ([Pillai, 2026b](#)).