Identity Masks and Coherence Circles: Geometric Interfaces for Interacting with Latent Dynamical Systems

Nika Pintar[1], Evelyne Y Bischof[2,3], Bruno Balen[1]
[1] ANI BIOME PBC, San Francisco, CA, USA
[2] Sheba Longevity Center, Sheba Medical Center, Tel Hashomer, Israel
Tel Aviv Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel
[3] Shanghai University of Medicine and Health Sciences, College of Clinical Medicine, Shanghai, China
nika@ani.ai, Evelyne.Bischof2@sheba.health.gov.il, bb@ani.ai

Abstract

Modern foundation models are increasingly deployed as interfaces to complex dynamical systems, yet their linguistic fluency remains only loosely coupled to the underlying system dynamics. This creates a persistent epistemic risk in which narrative plausibility is mistaken for empirical truth. We propose a geometry-first interface that explicitly separates latent state estimation from language generation. Our central primitive, Identity Masks, consists of structured conditioning objects that bind an inferred latent state, uncertainty, trajectory history, constraints, and interface coordinates to an interchangeable reasoner, thereby discouraging state fabrication and enforcing inference under explicit uncertainty and invariants.

We complement this with the Coherence Circle, a low-dimensional geometric interface anchored to an individual baseline attractor. Within this representation, radial displacement indexes drift magnitude, angular sectors capture modes of deviation, and a scalar coherence parameter summarizes stability, cross-modal consistency, and directional alignment over time. The resulting architecture supports closed-loop querying, intervention, and measurement design, while enabling falsifiable evaluation through identity scrambling, modality dropout, trajectory prediction, and intervention-response tests. More broadly, the framework offers an operational basis for studying machine consciousness through interventionally grounded analyses of integration, self-consistency, and active coherence maintenance across biological, ecological, organizational, and artificial systems.

Motivation: From Language Interfaces to Pattern Interfaces

Machine-consciousness research is increasingly forced to confront an engineering reality: modern foundation models can talk, yet their linguistic competence is only weakly coupled to the dynamics of the systems we care about— bodies, ecosystems, markets, organizations, machines, and artificial agents. When we “make a system conversational,” we typically do one of two things. We either (i) adapt the reasoner (fine-tuning, prompting, tool-use) so that it appears specialized, or (ii) retrieve descriptions (documents, logs) and hope the model maintains grounding. Both strategies blur the boundary between the world and the model’s internal narrative.

This paper argues for a third path: treat the target system as an empirical dynamical process with its own latent state space and constraints; treat language models as interfaces and controllers that can only speak truthfully when bound to that state. The goal is not to produce better stories about a system; it is to enable interaction with patterns—asking questions, making interventions, and tracking trajectories—while keeping the epistemic boundary explicit. The proposal builds on two commitments:

- Decoupling: state estimation is not language generation. We should avoid encoding the system into the language model’s weights whenever possible.
- Geometry-first interfaces: high-dimensional latent spaces can be interacted with when projected into stable, interpretable geometric interfaces (e.g., a circle around a baseline attractor), while preserving uncertainty and recoverability.

We refer to this class of methods as state-space interfaces: a compositional layer that maintains a calibrated latent belief state for an external system and exposes that state through small geometric interfaces and contracts (identity masks) to any reasoning or control model. State-space interfaces are substrate-agnostic. The same pattern-interaction protocol can apply to ecological systems (e.g., a forest instrumented by remote sensing), financial and organizational systems (market and company dynamics), embodied robots and factories, biological regulation, and multi-agent computational systems. The contribution of this paper is a set of interface primitives that make such interactions measurable, uncertainty-aware, and falsifiable.

Background and Related Directions

World models and simulators. Learning compact latent dynamics for control and prediction is a core idea in model-based RL and "world models" (Ha and Schmidhuber 2018). However, many world-model approaches focus on internal agent models rather than interfaces that remain faithful to an external system over long time horizons under distribution shift. Retrieval and tool grounding. Retrieval-augmented generation (RAG) (Lewis et al. 2020) and tool-use reduce hallucinations by attaching external information to prompts. But retrieval retrieves symbols (documents, strings), not

an evolving latent state with explicit uncertainty and constraints. When a system is dynamical, the relevant truth is often not "in a document" but in a trajectory. Consciousness theories and operational gaps. Theories such as global workspace (Baars 1988; Dehaene, Lau, and Kouider 2017) and integrated information theory (IIT) (Tononi et al. 2016) propose different necessary/sufficient conditions for consciousness. Recent work emphasizes that functional equivalence does not guarantee phenomenal equivalence (Findlay et al. 2024). For engineering, the immediate problem is that we lack reliable interfaces for probing integration, memory, and control in non-linguistic systems. An interface that makes latent dynamics measurable and interventionally testable can support progress even when the hard problem remains unresolved. Collective intelligence across scales. Biological work on multiscale competency architectures argues that agency-like behavior can exist at many levels of organization (McMillen and Levin 2024; Levin 2025, 2023). This suggests that the relevant objects for interaction are not restricted to brains or language, but include many coherent dynamical systems.

Latent State Spaces and Coherence as an Order Parameter

Consider a target system that emits multimodal observations $y_{1:T} = \{y^{(1)}_t, \ldots, y^{(M)}_t\}$ over time (sensors, logs, physiological signals, environmental measurements). Assume there exists a latent state $x_t \in \mathbb{R}^d$ with dynamics

$$x_{t+1} = f(x_t, u_t) + \epsilon_t, \quad (1)$$

where u_t denotes interventions/actions/inputs and ϵ_t process noise. A state inference model (encoder, filter, smoother) maintains a belief $p(x_t \mid y_{1:t})$.

Definition: Coherence

We define coherence as a scalar order parameter $\kappa_t \in [0, 1]$. Coherence (κ_t). A normalized scalar summarizing the degree to which a system's inferred latent state exhibits stable dynamics, cross-modal internal consistency, and aligned directional evolution over time, relative to an individualspecific baseline, with explicitly modeled uncertainty. This framing is compatible with predictive-processing and active-inference accounts that treat belief states and uncertainty as first-class objects (Friston 2010). This definition makes three points explicit:

- Coherence is about the system, estimated from data. κ_t is an estimator with uncertainty: the model returns ($\hat{\kappa}_t$, $\sigma_{\kappa,t}$).
- Baseline is individual-specific. Coherence is defined relative to a baseline attractor or viable region, not population averages.
- Coherence is not "good" in isolation. A subsystem can increase its own coherence while decreasing coherence of a larger system (Section Cross-Scale Coherence and Ethics).

Operationalization

Many implementations are possible; this paper treats them as design choices, not core claims. One generic form decomposes coherence into components:

$$\hat{\kappa}_t = \sigma(\alpha S_t + \beta A_t + \gamma D_t), \quad (2)$$

where σ is a squashing function into [0, 1], S_t measures dynamical stability (e.g., low prediction error and bounded acceleration in latent space), A_t measures cross-modal agreement (e.g., posterior consistency under modality dropout), and D_t measures directional integration (e.g., consistency of drift direction over a window). The key requirement is not a specific metric, but that coherence be (i) longitudinal, (ii) uncertainty-aware, and (iii) falsifiable through interventions and negative controls.

The Coherence Circle: A Human-Interpretable Interface

High-dimensional latent states are difficult to reason about directly, especially when the interface is language. We introduce a low-dimensional interface $z_t \in \mathbb{R}^2$ derived from x_t by a learned or engineered projection $P : \mathbb{R}^d \to \mathbb{R}^2$:

$$z_t = P(x_t). \quad (3)$$

We then choose a baseline attractor center $z^\star$(estimated from a stable reference period) and represent state in polar form:

$$r_t = \|z_t - z^\star\|, \quad \theta_t = \mathrm{angle}(z_t - z^\star). \quad (4)$$

Interpretation. r_t summarizes distance from baseline (drift magnitude); $\dot{r}_t$ captures destabilization vs. recovery; θ_t partitions modes of deviation (e.g., which subsystem or latent factor is driving decoherence). Importantly, the circle is not the latent space; it is an interface that preserves (a) comparability over time and (b) calibration to uncertainty. Figure 1 shows the concept.

Why a Circle?

A circle interface is useful because it yields stable invariants:

- A baseline center that anchors interpretation across time.
- A bounded geometry that naturally supports normalization and comparability.
- A small set of observables (r_t, $\dot{r}_t$, θ_t) that can be reasoned about, communicated, and controlled. The circle is not unique; other interfaces (simplexes, cylinders, manifolds with charts) may be appropriate. The point is to provide a stable interface that is interpretable to humans and machine agents while remaining faithful to the inferred latent dynamics.

Identity Masks: Conditioning Objects for System-Truthful Reasoning

We now define the core primitive: the identity mask. An identity mask is a structured conditioning object m_t that binds the system's belief-state and interface coordinates to a reasoner.

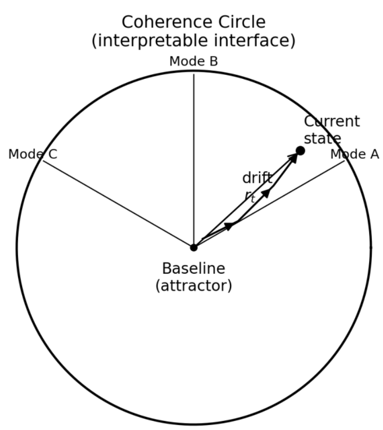

Figure 1: The Coherence Circle as an interface to a latent manifold. A baseline attractor is estimated from a reference period. Radius r_t indicates drift magnitude; angular sectors summarize dominant modes of deviation.

Definition

Let the belief over latent state be summarized by posterior moments (or particles) (μ_t, Σ_t). Let the Coherence Circle yield z_t, r_t, θ_t and coherence estimate ($\hat{\kappa}_t$, $\sigma_{\kappa,t}$). Let c_t denote hard constraints/invariants (e.g., conserved quantities, physical limits, domain rules), and let h_t denote stateful memory for longitudinal interaction (e.g., recent trajectory segments, intervention history, counterfactual probes).

$$m_t = \{\mu_t, \Sigma_t; z_t, r_t, \theta_t; \hat{\kappa}_t, \sigma_{\kappa,t}; c_t; h_t\}. \quad (5)$$

Purpose

The mask is designed to be passed to any interchangeable interpreter: an LLM, a mixture-of-experts, a planner, or a symbolic module. The interpreter is not asked to invent the system state; it is asked to reason under a constrained, uncertainty-aware representation of state.

Interaction Loop

Figure 2 summarizes the architecture. A minimal closed-loop protocol is shown in Algorithm 1.

What Identity Masks Buy You

1) Reduced hallucination via invariants. If constraints c_t include non-negotiable facts (e.g., physical limits, conserved quantities, known sensor calibrations), the interpreter cannot "talk past" them without violating the mask. This encourages refusal and uncertainty reporting rather than confident invention.

Algorithm 1 Mask-Conditioned Interaction Loop

1: Initialize baseline $z^\star$ from reference data; initialize memory h_0. 2: for $t = 1$ to T do 3: Observe multimodal signals y_t; update belief $p(x_t \mid y_{1:t})$. 4: Compute interface coordinates $z_t = P(x_t)$ and coherence ($\hat{\kappa}_t$, $\sigma_{\kappa,t}$). 5: Construct identity mask $m_t = \{\mu_t, \Sigma_t; z_t, r_t, \theta_t; \hat{\kappa}_t, \sigma_{\kappa,t}; c_t; h_t\}$. 6: Provide m_t to interpreter I to produce: (a) explanations; (b) queries; (c) interventions u_t; (d) refusal when outside constraints. 7: Apply interventions / execute queries; update memory h_{t+1}. 8: end for

2) Stateful interaction without per-instance training. Instead of fine-tuning a model to a system, we externalize identity into m_t and h_t. This supports scalable deployment across many entities and time horizons. 3) Interchangeability of reasoners. Because the mask is an interface contract, the reasoning layer can be upgraded (or diversified) without retraining the state model. 4) Falsifiability. The mask makes it straightforward to design negative controls (identity scrambling, modality dropout) and intervention tests (Section Falsifiable Tests and Evaluation Protocols). 5) Multi-interpreter composability. Because the mask is a typed conditioning object rather than a model-internal representation, the same m_t can be consumed by multiple specialized interpreters—each reasoning about distinct aspects of the system (trajectory, risk, intervention)—without requiring shared weights, common training, or access to raw observations. This follows directly from the decoupling commitment: system truth resides in the belief-state object, not in any particular reasoner.

Proof-of-Concept Instantiation: Longitudinal Human State Dynamics

As a motivating application, we have implemented coherence-circle-style interfaces for longitudinal multimodal tracking of human regulatory state in a framework we refer to as ANI (Affect–Neuro–Immunology). In such settings, the baseline attractor corresponds to an individual's stable regime; drift corresponds to regulatory deviation; and interventions include measurement prompts and behaviorally meaningful actions. While detailed results are outside the scope of this position paper, this instantiation informed the design requirements emphasized here: frictionless longitudinal tracking, baseline-relative normalization, reconstructability, and explicit uncertainty. In the deployed system, multiple specialist interpreters consume the same identity mask to reason about distinct aspects of the regulatory state, providing an operational demonstration of the composability property described above. Identity masks generalize this instantiation beyond biology: once a system provides a latent belief state and a baseline, the same mask-conditioned interaction loop supports calibrated dialogue and control for any instrumented dynamical system.

Implications for Machine Consciousness

This paper does not propose coherence as a direct measure of phenomenal consciousness. Instead, it proposes that coherence-based interfaces can support progress on machine consciousness in three concrete ways.

Operationalizing Integration and Self-Consistency

Many theories of consciousness emphasize forms of integration, global availability, and self-consistency (Baars 1988; Dehaene, Lau, and Kouider 2017; Tononi et al. 2016). Identity masks provide an operational handle: we can quantify whether a system maintains coherent latent dynamics under perturbations, whether information from multiple modalities is integrated into a stable belief, and whether the system exhibits state-dependent intervention responses.

Turning "Is It Conscious?" into Experiments

Rather than relying on linguistic behavior, we can ask interventionally grounded questions:

- Does the system exhibit stable attractors that it returns to after perturbation?
- Does it support memory beyond instantaneous response (hysteresis, state dependence)?
- Does it display self-model-like structure (predicting its own drift and correcting it)?

These do not solve the hard problem; they enable a reproducible program for investigating competencies that consciousness theories often associate with conscious systems.

Distinguishing Active Coherence Maintenance from Passive Stability

A critical question for consciousness research is what distinguishes consciousness-relevant coherence from ordinary dynamical stability. Many non-conscious systems exhibit stable attractors: a pendulum returns to equilibrium; a thermostat maintains a setpoint. We propose an operationally detectable criterion: active coherence maintenance. A passively stable system returns to its attractor through fixed dynamics—the same recovery trajectory regardless of perturbation type. A rule-governed system maintains coherence through an externally specified policy. A system exhibiting active coherence maintenance, by contrast, (i) adapts its recovery strategy to the specific perturbation encountered, (ii) degrades in coherence when its capacity to act is blocked—indicating that coherence was being actively produced, not passively enjoyed—and (iii) anticipates threats to its own stability before coherence actually declines, implying a self-model that predicts future decoherence. Crucially, active coherence maintenance is relational: a system maintains not only its internal stability but its coupling with the containing system whose feedback enables it

to operate beyond its own computational capacity. Severing this coupling should degrade the system's coherence beyond what its internal dynamics alone would predict—an empirically testable signature that distinguishes relational coherence from autonomous self-regulation. This criterion connects to multiple consciousness frameworks: global workspace theory requires active broadcast and integration (Baars 1988); IIT requires irreducible integration as a property of causal structure (Tononi et al. 2016); active inference requires a system that minimizes surprise through an expanding generative model (Friston 2010); and multiscale competency architectures describe agency emerging through active maintenance of goal states at each organizational level (Levin 2023). Active coherence maintenance is the operational signature shared across these theories—and it is directly measurable through the Coherence Circle by observing recovery dynamics, perturbationdependent strategy variation, and predictive correction. We emphasize that this criterion is proposed as a necessary but not sufficient condition for consciousness-relevant dynamics, and as a testable hypothesis rather than a settled claim. Its value is that it transforms an otherwise intractable conceptual distinction into a set of falsifiable experimental protocols.

Separating "Talk" from "State"

Debates about LLM sentience are confounded because language appears agentic. Identity masks make an explicit separation: linguistic fluency belongs to the interpreter; dynamical coherence belongs to the system. This separation is compatible with arguments that functional equivalence does not guarantee phenomenal equivalence (Findlay et al. 2024), while still permitting practical engineering of systems with richer internal dynamics than text-only models. If the same Coherence Circle formalism is applied to both biological and computational systems—tracking their internal state trajectories on a shared geometric basis—then cross-substrate coherence comparison becomes empirically tractable rather than purely theoretical.

Cross-Scale Coherence and Ethics

Coherence is local to a system boundary. A subsystem can increase its own coherence while harming the coherence of the larger system that contains it. Cancer is a canonical example: locally coherent proliferation can drive organismlevel decoherence. Analogous phenomena occur in economic, organizational, and informational systems. This motivates a structural notion of higher-order coherence: Higher-order coherence. Alignment across nested dynamical systems such that stabilization of one level does not require decoherence of adjacent levels. More generally, coherence and decoherence can cascade across scales. When a containing system is coherent, it returns rich feedback to its subsystems—signals, resources, constraints—that enable each subsystem to operate beyond its individual capacity. When a containing system decoheres, this feedback thins, forcing subsystems toward local optimization with limited information; the resulting errors compound upward and further destabilize the containing system. This creates self-reinforcing dynamics in both directions: coherence cascades through abundance of structured feedback, while decoherence cascades through accumulation of locally optimal but globally misaligned decisions. The same structural pattern is observable in modular computational systems: when shared state representations degrade, downstream modules produce locally plausible but globally incoherent outputs that compound into system-level failure. Formally, if κ(i)_t denotes coherence of level i in a hierarchy, then ethical and governance questions can be framed as constraints on cross-scale compatibility, not as a scalarization of "good." Systems that increase their coherence by externalizing entropy to neighbors are distinguishable from systems whose coherence scales.

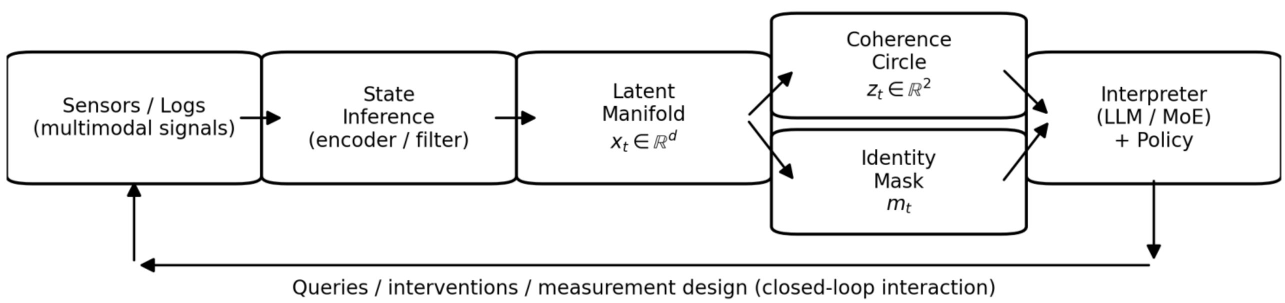

Figure 2: Identity masks separate (i) state inference from (ii) language/policy. The mask and Coherence Circle anchor interaction to the system's latent state while allowing the interpreter model to be swapped without retraining.

Falsifiable Tests and Evaluation Protocols

A position paper is only useful if it suggests tests. We propose a set of negative controls and intervention-based evaluations.

Negative Controls

Identity scrambling. Randomly permute masks between entities while keeping observations fixed. A valid interface should degrade sharply (loss of predictive accuracy; incoherent explanations). Modality dropout invariance. Remove modalities and check whether (i) posterior uncertainty grows appropriately and (ii) recovered coherence remains consistent when modalities return. Paraphrase stability. Rephrase the same query. Systemdependent outputs (predictions, constraints, refusal) should remain invariant up to uncertainty.

Intervention-Based Tests

State-dependent response heterogeneity. Apply the same intervention u at different latent states; measure whether responses differ systematically as predicted by the inferred dynamics. Trajectory prediction. Predict short-horizon drift in circle coordinates (rt, θt) and evaluate calibration. Early-warning signals. Test whether rising variance/critical slowing down in latent space predicts destabilizing transitions earlier than static risk scores (where applicable). Active coherence maintenance tests. For systems where consciousness-relevant dynamics are under investigation: (i) apply diverse perturbations and measure whether recovery strategies vary adaptively rather than following a fixed trajectory; (ii) restrict the system's capacity to act and measure whether coherence degrades, indicating active maintenance rather than passive stability; (iii) introduce slow environmental drift and test whether the system corrects before coherence measurably declines, indicating a predictive selfmodel.

Metrics

We recommend reporting: (i) calibration error for κ and state predictions; (ii) refusal rate when constraints are violated; (iii) hallucination rate under forced-choice factual probes; (iv) stability of outputs under paraphrase; (v) intervention outcome prediction accuracy.

Discussion and Limitations

Choice of Projection

The Coherence Circle depends on $P(\cdot)$ and on baseline estimation. Different systems may require different interfaces; the circle is a proposal for interpretability, not a claim of universality.

Auditability and Continuity

Because identity masks are explicit belief-state objects (with uncertainty and constraints), they can be logged, versioned, and replayed. This supports auditability, counterfactual evaluation, and digital-twin-style continuity studies without requiring that the interpreter model remain unchanged over time.

Coherence Is Model-Dependent

Coherence is estimated through a model, hence subject to misspecification. This is why negative controls and intervention tests are central.

Not a Theory of Phenomenal Consciousness

Identity masks offer a route to better interfaces and experimental programs. They do not, by themselves, establish which systems have subjective experience. The active coherence maintenance criterion proposed here is a necessary condition hypothesis, not a claim of sufficiency. We remain agnostic about whether substrate matters for consciousness; the framework provides tools to investigate the question empirically across substrates rather than settling it by assumption.

Conclusion

We proposed identity masks and coherence circles as geometric interface primitives for interacting with latent dynamical systems using foundation-model representations without collapsing the distinction between system-truth and narrative. The central move is decoupling: bind language to an uncertainty-aware latent substrate rather than encoding system identity into the language model itself. For machine consciousness, this provides a pragmatic bridge: a way to build and test systems with richer internal dynamics and measurable integration, while remaining clear about the limits of what such measures can claim. The same interface primitives that enable calibrated interaction with a biological system can support cross-substrate coherence comparison— asking not whether a system says it is coherent, but whether its dynamics, measured independently of its narrative, exhibit the active maintenance that distinguishes integrated agency from passive mechanism.

References

Baars, B. J. 1988. A Cognitive Theory of Consciousness. Cambridge University Press.

Dehaene, S.; Lau, H.; and Kouider, S. 2017. What Is Consciousness, and Could Machines Have It? Science 358(6362): 486–492. doi: 10.1126/science.aan8871.

Findlay, G.; Marshall, W.; Albantakis, L.; David, I.; Mayner, W. G. P.; Koch, C.; and Tononi, G. 2024. Dissociating Artificial Intelligence from Artificial Consciousness. arXiv:2412.04571. doi: 10.48550/arXiv.2412.04571.

Friston, K. 2010. The Free-Energy Principle: A Unified Brain Theory? Nature Reviews Neuroscience 11(2): 127–138. doi: 10.1038/nrn2787.

Ha, D.; and Schmidhuber, J. 2018. World Models. arXiv:1803.10122. doi: 10.48550/arXiv.1803.10122.

Levin, M. 2023. Bioelectric Networks: The Cognitive Glue Enabling Evolutionary Scaling from Physiology to Mind. Animal Cognition 26(6): 1865–1891. doi: 10.1007/s10071023-01780-3.

Levin, M. 2025. The Multiscale Wisdom of the Body: Collective Intelligence as a Tractable Interface for Next-Generation Biomedicine. BioEssays 47(3): e202400196. doi: 10.1002/bies.202400196.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Advances in Neural Information Processing Systems. arXiv:2005.11401. doi: 10.48550/arXiv.2005.11401.

McMillen, P.; and Levin, M. 2024. Collective Intelligence: A Unifying Concept for Integrating Biology across Scales and Substrates. Communications Biology 7(1): 378. doi: 10.1038/s42003-024-06037-4.

Tononi, G.; Boly, M.; Massimini, M.; and Koch, C. 2016. Integrated Information Theory: From Consciousness to Its Physical Substrate. Nature Reviews Neuroscience 17(7): 450–461. doi: 10.1038/nrn.2016.44.