PERSPECTIVES NEURO

Understanding and engineering consciousness

How does matter become conscious?

Doris Tsao
Chief Scientist - Neuro

PART 1

A new conceptual framework for consciousness

At some point in evolution, perhaps 200 million years ago when the first small mammals with cortex emerged, biology unlocked a remarkable property of the universe: the ability for matter to be bright inside. Given that consciousness arose through an evolutionary process, it is clearly not divorced from the physical world but is instead a property of certain forms of organized matter. The scientific challenge of our time is to understand how this property arises. *How does matter become conscious?*

I believe our understanding of the brain has reached a stage where we can, for the first time, crack the fundamental mechanism for consciousness. The stakes could not be higher. The future of humanity and the nature of our relationship to AI—whether it will forever remain a tool, or whether it could become us—hinge on understanding how matter becomes conscious. Here, I present a conceptual framework and concrete scientific strategy that I believe can fully crack the mechanism for consciousness.

As a starting point, it is useful to reflect on a paradox of conscious experience that has been observed across the ages.

Schopenhauer recognized that the physical present is too fleeting to support experience:

> *"In [time], each instant is only in so far as it has effaced the preceding one, its generator, to be itself in turn as quickly effaced. The past and the future... are empty as a dream, and the present is only the indivisible and unenduring boundary between them"* [1].

William James similarly wrestled with the transience of the present moment. He evoked the image of a saddleback, facing both past and future, to capture the nature of being conscious:

> *"the practically cognized present is no knife-edge, but a saddle-back, with a certain breadth of its own on which we sit perched, and from which we look in two directions into time. The unit of composition of our perception of time is a duration, with a bow and a stern, as it were—a rearward- and a forward-looking end"* [2].

I believe Schopenhauer and James are both gesturing toward the same fundamental neuroscientific claim about consciousness: to be conscious of anything at all, the brain cannot simply pass the instantaneous state of the world through a relay of feature detectors. Instead, it must generate internal representations that **persist** across time, bridging past to future, which can be consciously perceived and become objects of thought.

How can a representation persist across time? I propose that this is accomplished through a remarkable class of representations that I call **symbols**—stable representations of objects, environments, concepts, and the self. These are the fundamental building blocks of the brain's internal model: a compositional system in which these symbols can be combined in structured ways to represent an infinite number of scenes, thoughts, and states of self. My core thesis is that consciousness is precisely the process by which **the brain generates symbols and composes them into an internal model of the world and the self.** I call this the persistent symbol theory of consciousness.

What is a symbol? At the phenomenological level, symbols are the stable, persistent objects of conscious experience—from the objects around you that you can see and touch, to your feelings, memories, and thoughts—everything you are conscious of is bound to a symbol. At the neural level, I argue that symbols arise when recurrent loops across the brain's hierarchy converge to stable states: coherent sets of mutually reinforcing activations, maintained through reciprocal predictions across brain areas, that reflect the brain's ongoing inference about the structure of the world. In particular, what makes such a state a symbol is not simply coherence but binding to an underlying representation that persists (elaborated below).

Among symbols, the representation of a physical object is the most elementary. Consciousness begins with the brain building a stable model of the physical world composed of representations of persistent chunks of matter (Figure 1). From this foundation the entire structure of subjective experience may be constructed: once the brain has a mechanism for representing persistent physical objects, it can represent relations between them—spatial, causal, and social—and it can construct thoughts, which are abstract objects built from the same underlying representational machinery.

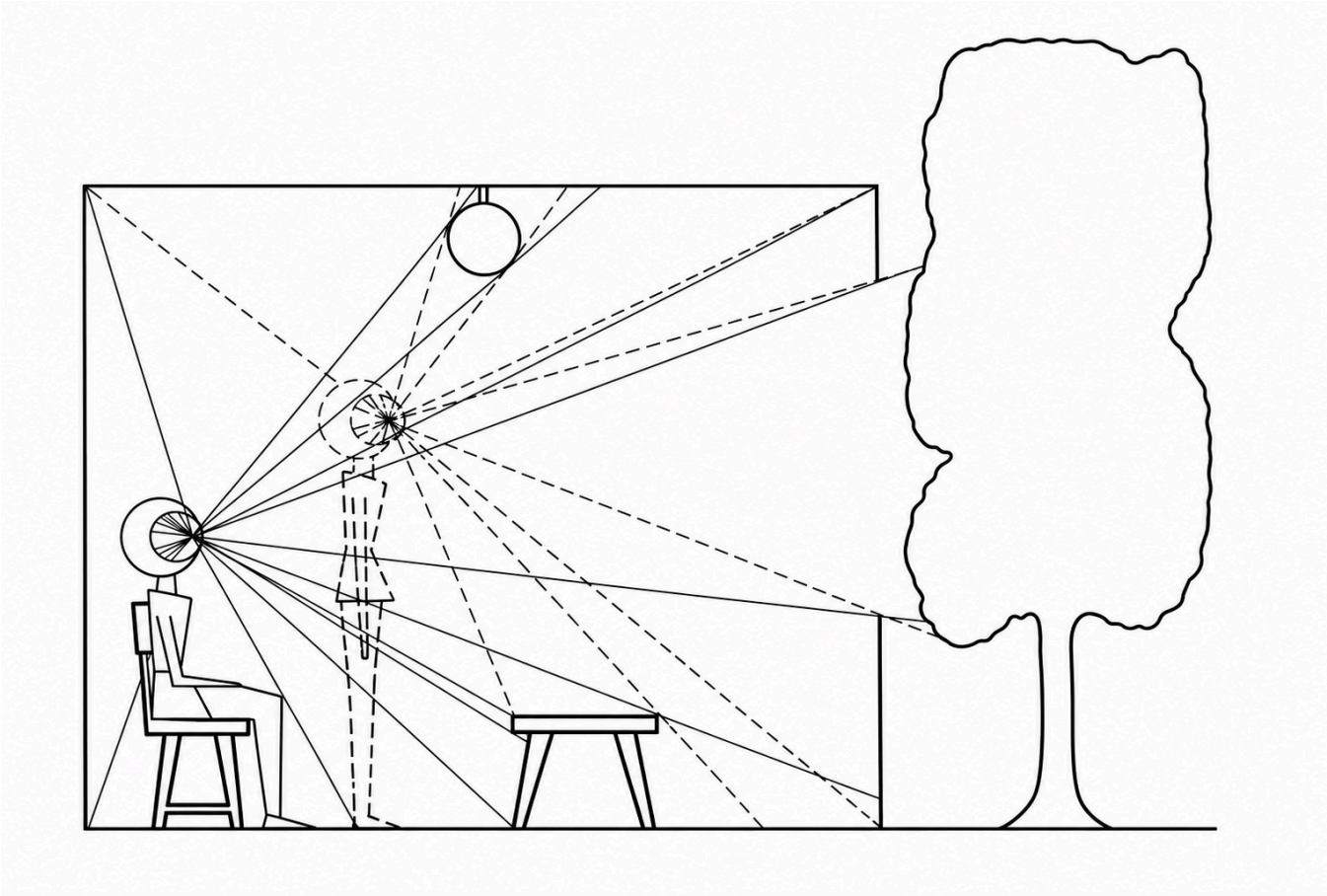


Figure 1. The fundamental structure of the world. The world consists of persistent chunks of matter (e.g., a table, a tree), and a fundamental task of an intelligent system is to perceive these stable units that compose the physical world. This is challenging because the sensory input signaling these units can change drastically over time due to changing perspective and partial occlusion. We argue that the mechanism underlying perception of these stable units is the secret to consciousness. From Gibson, "The Ecological Approach to Visual Perception" [3].

Beyond persistence, objects in the physical world have a second key property: despite differences in identity, they are *mathematically identical as visual surfaces*—2D manifolds that can be tracked using the same general-purpose topological machinery. This key fact, proved in Tsao and Tsao [3], means that the brain does not need a dedicated circuit for each kind of object. It just needs a single, repeated unit: an *object-tracking column* that can latch onto any surface and track it through time. (Such "columns" could be implemented as distinct subspaces of neural population activity rather than anatomically segregated groups of neurons.) Because a given column can be assigned to any object, a small, fixed set of columns can represent a limitless range of physical scenes. In other words, the columns form a symbolic system that is **compositional**. Higher symbolic systems, including language and thought, retain this key feature of compositionality. Indeed, I believe the generativity of language and thought emerged directly from the compositionality of the physical world. This compositionality of perception and cognition has an important engineering consequence: if conscious experience arises from symbols flexibly composed into structured configurations, then we should be able to *program* any conscious experience through targeted manipulations of these symbols and their relations, to write specific percepts, thoughts, and internal states into the brain.

Among all the objects modeled by the brain, one has special status: the self. The self at the most primitive level is simply the physical body. It is an object we can not only see but also, uniquely among objects, sense through proprioception. The self's uniqueness extends even further: on top of a perceptual representation of the body, the brain builds a representation of *bodily control*, i.e., action plans. In this way, the brain tracks not just where the body is but what it can do; this is subjectively experienced as agency—the felt sense of free will. In schizophrenia, individuals sometimes experience their own actions as controlled by external forces. Such disruptions underscore that the self, like any other object, is a neural construct whose mechanism can be understood and engineered.

In sum, I believe that *solving consciousness is nothing more and nothing less than understanding how the brain constructs a coherent model of the world and the self*—in other words, solving the central challenge of systems neuroscience. The brain faces an inference problem: given noisy, fleeting sensory input, what is actually out there, and what should I do about it? Its solution is to construct an internal model of the world and the self built from persistent, compositional symbols.

This framing of consciousness as world-model inference resolves an otherwise impossible puzzle: *why did consciousness evolve?* Since natural selection acts on behavior, not on subjective experience, it is prima facie odd that evolution would favor states that "feel" like something from the inside. But if consciousness *is* sophisticated world modeling from the inside, the puzzle dissolves: consciousness evolved because it is beneficial for survival. As Kenneth Craik observed, an organism that carries a small-scale model of the world "*is able to try out various alternatives… and in every way to react in a much fuller, safer, and more competent manner to the emergencies which face it*" [4]. Consciousness, then, is the brain doing what it evolved to do, modeling the world and self, experienced from the inside. And this is why we can hope not only to understand it but to build it.

Reasons for optimism. Systems neuroscience is ready for a phase transition. I believe we can fully crack the brain's internal model, and with it, the mechanism of consciousness, if we make two key changes in approach. First, we need to set our sights beyond what any single academic lab can take on, to tackle the full problem of how the brain represents the world and self. *We need to want a phase transition.* The genius of our age is scaling, and we in systems neuroscience should embrace this. Second, we need to exploit the powerful circuit manipulation tools built over the past decade to move our field beyond open-ended discovery of neural correlates of perception, cognition, and internal state, into a new regime: *consciousness engineering.* We need to hold ourselves to Feynman's standard, *"What I cannot create, I do not understand."*

Luckily, the conditions are ripe. A new cohort of efforts—including Astera Neuro, the Enigma Project, the Simons Collaboration on Ecological Neuroscience, the Ackman Oxman Institute, Merge Labs, E11 Bio, and Flourish—is taking on the brain at scales no single lab could.

Meanwhile, conceptual and technical advances have laid fertile ground for radical new experiments in consciousness engineering. The profound changes in consciousness that result from electrical stimulation, psychiatric disruptions, and psychedelics have long demonstrated that consciousness is engineerable. Three recent developments now make systematic pursuit of this possibility tractable. First, profound new

insights into the brain's compositional model of objects and thought have recently been achieved [5–7]. Second, powerful technologies are emerging that let us interrogate the brain at a scale unimaginable five years ago: large-scale recording with Neuropixels probes [8], chronic recording with flexible, high-channel-count polymer probes [9,10], calcium and voltage imaging and holographic stimulation of neural populations [11–13], and connectomics to map brain-wide connectivity [14,15]. Finally, advances in AI now enable neural activity to be modeled at a scale that captures its full variance during naturalistic behavior [16–18], opening the door to closed-loop systems that can predictively manipulate neural dynamics to induce specific perceptual and cognitive states.

Beyond modeling neural data, AI lets us test principles by building with them. Dileep George has expressed the idea that finding the algorithms of intelligence requires a triangulation—seeking the commonalities between brains, algorithms, and the structure of the world: *"A property of the brain is likely to be a general principle only when we can explain why it exists based on algorithmic ideas and the organization of the world"* [19]. I strongly concur: we need to distill, from the baroque complexity of neural circuits, the fundamental algorithmic principles the brain uses to build a coherent model of the world. I would go even further: among these principles, the most fundamental for understanding consciousness is the principle supporting perception of the stability of the world.

Previously, any proposed neural principle could only be checked against more biological data. But now, we can take a candidate principle (e.g., persistent, compositional object tracking), implement it in an artificial system, and ask if *it works* (e.g., enables a robot to interact robustly with the real world). The principles underlying the brain's world model, distilled and scaled in this way, will almost certainly unleash a new generation of AI architectures. Like the human brain, these architectures will be highly energy-efficient and interface seamlessly with not only language but also the real world. But even more fundamentally, these architectures will possess an entirely new capability: *consciousness*. This last step change will lead to a radical transformation of the human condition.

In the remainder of this document, I lay out: (1) a proposal for the fundamental structure of consciousness, grounded in a structure I hypothesize is critical for consciousness, the 'object-tracking column'; (2) a proposed five-year experimental roadmap for Astera Neuro; and (3) a vision for merge and upload.

The structure of the brain's internal model. We are all experts on the brain's internal model, from *being* it every waking hour. Its structure can be thought of as a set of layers providing successively greater predictive power: persistent objects in the immediate present; memories that extend our experience through time into past and imagined episodes; thoughts that compose abstract concepts to reason beyond any particular experience; and a sense of self, anchored in moment-to-moment bodily continuity but extending to a personality that lasts a lifetime. In what follows, I introduce each layer in turn, foreshadowing how each connects to my central theoretical proposal of the persistent object-tracking column as the key mechanism on which the higher layers build.

Objects. The foundational layer of the internal model is the representation of physical objects. We perceive objects as invariant units that persist through transformation. The fundamental cognitive mechanism

underlying object representation is invariant tracking of a specific piece of matter through something like an *object file*: a temporary, object-specific representation that maintains continuity over time and serves as index to stored information [20]. I will argue that the brain implements object files through *persistent object-tracking columns*, and understanding the circuit basis of these structures is the critical step to understanding the brain's model of the physical world, and the higher layers of the internal model that build on it.

Memory. Beyond the immediate perceptual stream, memory extends experience through time, letting us relive the past and simulate possible futures. Its raw material is the persistent symbols of experience; from these it composes coherent episodes [21,22]. Rather than preserving every detail, memory keeps the gist, compressing an event into the key relations and transitions among its symbols, a story it can later reconstitute and replay. Memory can be construed as object tracking extended across a lifespan, creating the autobiographical trajectory we experience as a persistent self.

Thought. From regularities in the behavior of physical objects, we form abstract *concepts*, e.g., numbers, faces, trees, justice. A hallmark of concepts is their generality: the concept "three" applies equally to three apples, three bananas, or three threes. Moreover, a concept can be not only a *thing* but an *abstract operation*, a program defined over variables rather than fixed content (e.g., addition) [23]. Thought is the fluid recombination of such concepts, including both objects and operations, to generate new insights. Recent findings have identified *working memory subspaces* in prefrontal cortex [5–7,24] that act as slots that can hold different content over time. I propose below that these structures are the abstract counterpart to persistent object-tracking columns.

Self. Conscious experience is organized around a first-person perspective on the world. Among the symbols that compose conscious experience, one has special importance: the symbol "I." This symbol *binds* the body (perceived visually and through proprioception), internal state (emotion and interoception), motivations, and intended actions (experienced as agency) into a unified representation linked to autobiographical memory. A key aspect of the self is continuity, including both *thin continuity*, the moment-to-moment tracking that maintains identity from one instant to the next, and *thick continuity*, the sense of having existed across a lifetime supplied by autobiographical memory. Notably, both binding and continuity are key challenges for object representation as well. The brain must bind shape, color, motion, and identity into a single unified object percept, and it must compute both thick and thin continuity for objects, tracking them from moment to moment while also integrating past encounters into a sense of familiarity. I thus propose that the self can be construed as an *object file writ large*, underscoring again why I believe the key bottleneck to understanding the neural mechanism for consciousness is understanding persistent object tracking.

A proposal regarding the fundamental structure of perception. If consciousness is the process by which the brain generates persistent, compositional symbols from raw sensory flux, then to understand consciousness we must understand the neural mechanism by which such symbols are produced. We need a theory of the data structure: what, at the level of cortical circuits, is a persistent symbol made of?

Among symbols, the representation of a physical object is the most elementary. Consciousness begins with the brain building a model of the physical world consisting of persistent chunks of matter. My core hypothesis is that the brain builds a physical model of the world using a new type of neural representation, which I call a **persistent object-tracking column** (Figure 2). Each column carries a structured representation of one object, factorized within the column into intrinsic properties (size, shape, affordances, material) and a transformation that captures the object's position and pose relative to the observer.

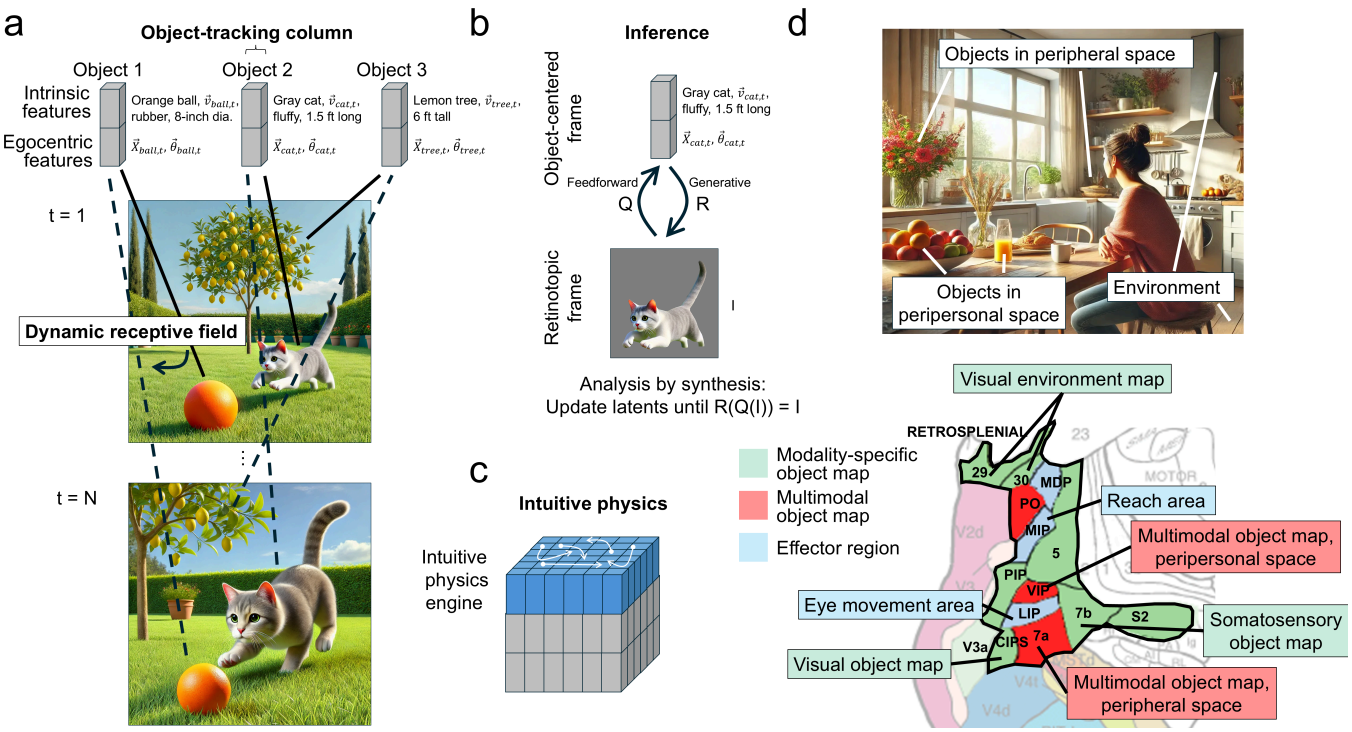


Figure 2. Building a world model using object-tracking columns. (a) An object-tracking column consists of cells that represent one object. The contents are changeable, forming an object pointer. Each column contains a factorized representation of intrinsic and egocentric object features and invariantly tracks specific objects. Through these columns, pixels become organized into coherent object units. (b) Inference of column contents from sensory data is accomplished through predictive coding via a generative loop. (c) An intuitive physics engine representing object dynamics can be built on top of the object-tracking column representation through recurrent connections. (d) This hypothesis provides a new perspective on the large-scale functional organization of posterior parietal cortex. We propose that PPC is tiled with maps of object-tracking columns, organized by sensory modality (visual, somatosensory, multimodal) and by the distance at which objects are represented (from peripheral space through peripersonal space), which then feed into downstream effector regions for eye movements and reaching.

The key insight is that the mathematical challenge of tracking any object as a physical chunk of matter in 3D space is fundamentally the same regardless of what the object is [3]. Any surface is a 2D manifold, and the key to tracking any manifold is to compute a sequence of overlapping neighborhoods across a sequence of perspectives. An apple and a rock have very different meanings, but their surfaces obey the same laws of transformation under dynamic occlusion and perspective. The same computational machinery can therefore track both, and by replicating that machinery across many modules, the brain *can represent any configuration of dynamically interacting objects in rich detail.*

I posit that the brain harbors a population of persistent object-tracking columns whose internal machinery is identical, with individual columns assigned to objects on the fly at each encounter with a new scene. This dynamic reassignment is what makes the scheme compositional: the brain can represent *any arrangement of matter* without requiring dedicated machinery for each possible configuration. Unlike shape-specific columns in inferotemporal cortex, which specialize in recognizing particular forms, object-tracking columns are general-purpose. Furthermore, because the columns represent chunks of physical matter, the representation of shape within each column should be supramodal: vision, touch, and audition can all be used to infer the same intrinsic physical properties.

I suspect these columns reside in posterior parietal cortex, which clinical evidence has long pointed to as the locus of the brain's world model. Patients with bilateral posterior parietal strokes suffer from Bálint's syndrome, losing precisely the capacity to perceive and act on the physical environment: they cannot perceive more than one object at a time, cannot direct their eyes to specific targets, and cannot guide their hands by vision [25]. In contrast, lesions of frontal cortex impair decision-making and lesions of temporal cortex impair object recognition, but neither destroys consciousness of physical reality.

I hypothesize that the object-tracking column is the anchor by which the brain forms the persistent symbols of the 3D physical world required for consciousness. To maintain such a symbol, the column must keep its persistent object representation consistent with the changing sensory evidence. I propose that this is accomplished through a *generative feedback loop s*panning posterior parietal cortex (PPC), retinotopic cortex, and inferotemporal cortex (IT): the column predicts the object's sensory appearance and identity, while discrepancies drive updates until the representations become mutually consistent. The converged state represents a rich symbol possessing persistence as a chunk of matter (via PPC), pixel-level resolution (via retinotopic cortex), and object identity (via IT). Crucially, the PPC column is what anchors this loop, maintains its converged state across time through tracking computations, and gates access to consciousness. Local representations in retinotopic and inferotemporal cortex are admitted into conscious experience by being recruited into this loop (for example, color is not itself a symbol and we are never conscious of it in isolation, but only as the color of something that persists).

The equations of consciousness. As a first attempt to concretize these ideas, we can formulate them mathematically. The equations below describe how a column represents an object, how multiple columns compose a world model, and how the brain searches for the world state most consistent with sensory input. I believe these equations are, in essence, the equations of consciousness.

Variables. We can model the latent state held by object-tracking columns as a structured scene graph: a set of object nodes, the relations among them, and the self. The nodes are drawn from a fixed pool of K object-tracking columns; not every column is in use at every moment. The occupancy of each column is given by $a_{k,t} \in \{0, 1\}$, and the set of objects currently represented is given by

$$V_t = \{\, k : a_{k,t} = 1 \,\}.$$

When a new object appears, a free column is recruited ($a_{k,t} : 0 \to 1$), and when it leaves, its column is released, so the number of objects is inferred and not fixed in advance.

The content of each active column $k \in V_t$ consists of an intrinsic representation I_k (invariant and object-centered: shape, mass, material, affordances) and a transformation $G_{k,t} \in \mathbb{G}$ (an element of a group, minimally $SE(3)$, specifying where the object is and how it is posed). Objects can be related to one another, and these relations are represented as a set of typed links among the columns,

$$E_t = \{\, \langle j, k, \rho_{j,k,t} \rangle \,\},$$

where $\rho_{j,k,t}$ labels the kind of relation (e.g., support, containment, part–whole, causal, social).

The self is formally an object writ large, possessing the same (I, G) structure, augmented with internal state S_t^{self}, motivations M^{self}, and intended action A_t^{self}. Together these constitute the column-level scene graph,

$$C_t = \{\, (I_k, G_{k,t}) \,\}_{k \in V_t} \;\cup\; E_t \;\cup\; \{ I^{self},\, G_t^{self}, S_t^{self}, M^{self},\, A_t^{self} \}.$$

The scene graph is not the only latent state. For simplicity, we focus here on the case of vision, though we note the framework naturally extends to all sensory modalities. Retinotopic cortex carries its own representation X_t of contours and surfaces, and inferotemporal cortex carries a representation ψ_t of object identity. The full latent state of the world model at time t is

$$W_t = (C_t, X_t, \psi_t).$$

The generative model. The apparent transformation of object k in the observer's frame factorizes into a part due to the self and a part due to the object:

$$G_{k,t}^{app} = \left(G_t^{self} \right)^{-1} \circ G_{k,t}.$$

The brain must recover this factorization by jointly disentangling the two transformations across all active columns. The renderer R_s takes each active column's intrinsic content and its current apparent pose and predicts the sensory cortical representation in modality s:

$$X_t^s = R_s \left(\{\, I_k, G_{k,t}^{app} : k \in V_t \,\} \right).$$

In addition to this map for generating the retinotopic representation from the scene graph, we further posit: (1) a map Ψ that predicts the inferotemporal representation from the same column content, $\Psi(C_t)$, (2) a map F that predicts sensory data from the retinotopic representation, $F(X_t)$, and (3) a map Φ that transforms the retinotopic representation into the inferotemporal representation, $\Phi(X_t)$.

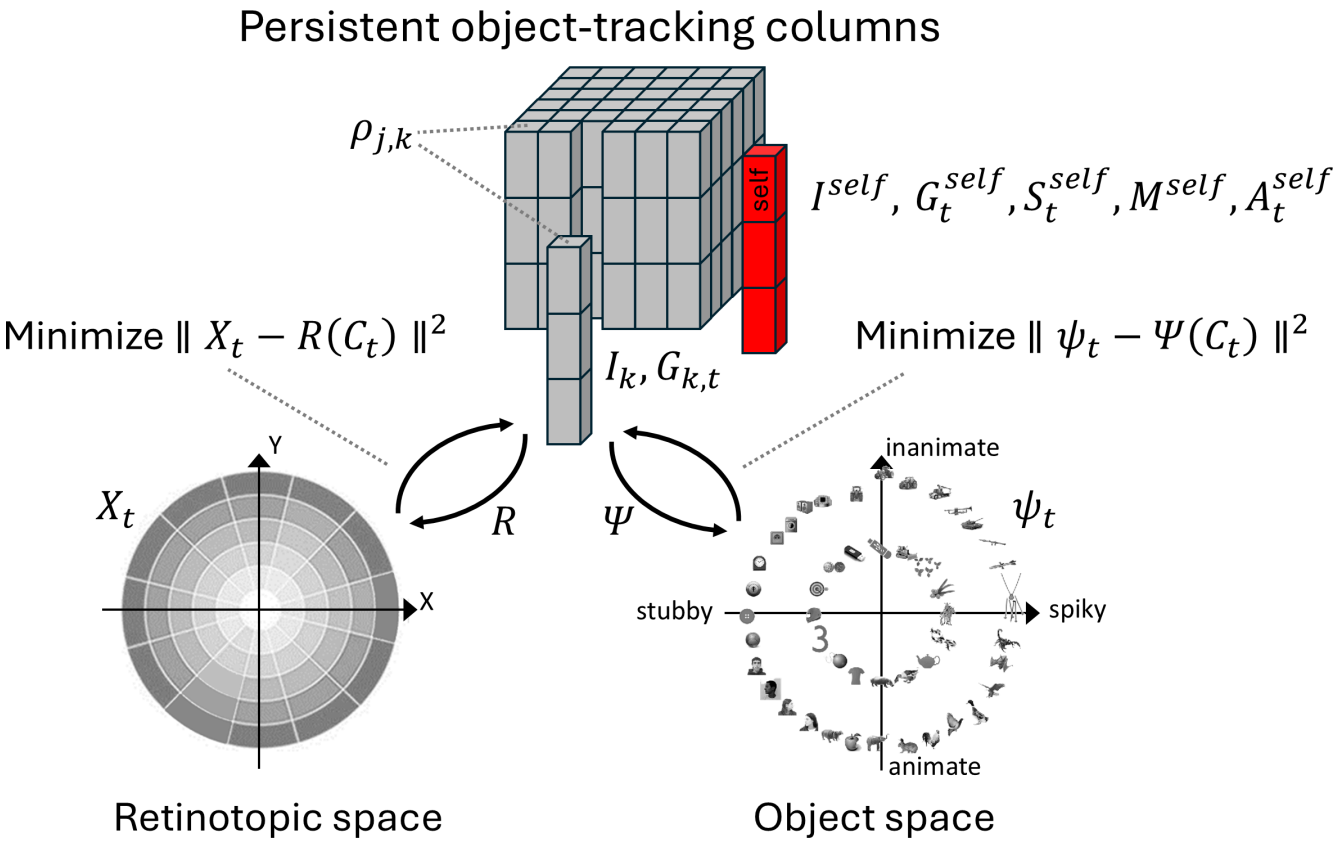


Figure 3. How feature representations across the brain become conscious. The brain's world model is built from persistent object-tracking columns in posterior parietal cortex, each holding an intrinsic representation I_k and a transformation $G_{k,t}$ for one tracked object, with relations $\rho_{j,k}$ between columns and an analogous set of columns for the self (red). Each column makes two predictions: the rendered image of its object in retinotopic cortex (via R) and its identity in inferotemporal cortex (via Ψ). The loop settles when prediction errors are jointly minimized. Feature representations become conscious by being recruited into this loop and stabilized at the fixed point W_t^*.

Priors. Four priors constrain the world model. **Persistence:** while a column tracks the same object, its intrinsic code I_k changes slowly, and is factorized from the transformation code $G_{k,t}$. **Exclusive binding:** each column tracks at most one object and each object is tracked by at most one column. The *binding problem* is solved by the column's exclusivity. **Parsimony:** the active set V_t and the edge set E_t are dynamically adjusted to use the fewest objects and relations needed to explain the sensory data and meet current behavioral requirements. **Temporal consistency:** the current estimate of the world state is constrained by predictions generated from the preceding inferred state and action, as described in the next section.

Dynamics. At each time step the self intends an action A_t^{self}, selected by its policy (see below) from the possibilities afforded by the current scene graph. Forward models predict how object poses, the self's pose, and its internal state would change if the action were carried out:

$$G_{k,t+1} = f^{obj}\left(k,\ C_t, A_t^{self}\right), \qquad G_{t+1}^{self} = f^{self}\left(C_t, A_t^{self}\right), \qquad S_{t+1}^{self} = h\left(C_t, A_t^{self}\right).$$

The models also receive motion information associated with objects and the self. These three forward models are, respectively, the brain's intuitive physics engine (the brain's built-in simulator of how objects

move and interact, guided by their relations [26]), the self pose-update engine, and the internal-state update engine. Occupancy and relations are carried forward with the dynamics and re-inferred as objects appear, disappear, or rearrange. Together, these update rules form a coupled dynamical system in which the self can act on the world and both the world and the self respond.

Inference. The brain seeks the world state W_t whose predictions match the sensory data and whose latent variables agree with one another and with predictions from the past. This is implemented as a recurrent loop with at least two prediction targets per active column. Column k predicts the rendered image of its object in retinotopic cortex, $R\left(I_k, G_{k,t}^{app}\right)$, and the identity of its object in inferotemporal cortex, $\Psi\left(I_k\right)$, where R is the rendering map and Ψ is the IT identity readout (Figure 3). The discrepancy between retinotopic and inferotemporal representations and the respective predictions from the PPC object column leads to adjustment of $I_k, G_{k,t}$ to drive the discrepancy toward zero, subject to the priors above. A conscious percept is the fixed point of this loop: the state W_t^* at which the retinotopic state X_t, the identity representation ψ_t, and the scene graph C_t are jointly consistent with the sensory data and with one another:

$$W_t^* = \arg\min_{W_t} \; \| D_t – F(X_t) \|^2 + \| X_t – R\left(C_t\right) \|^2 + \| \psi_t – \Psi\left(C_t\right) \|^2 + \| \psi_t – \Phi(X_t) \|^2 + L_{priors}\left(W_t\right),$$

where D_t is the sensory data, $F(X_t)$ the prediction of those data from the retinotopic state, $R(C_t)$ and $\Psi(C_t)$ the column-based predictions of the retinotopic and identity representations, $\Phi(X_t)$ the feedforward prediction of identity from the retinotopic state, and L_{priors} imposes persistence on I_k, exclusive binding, parsimony on the number of objects V_t and relations E_t, and temporal consistency between the current inferred state and predictions generated from the preceding inferred state W_{t-1}^* and the self's action A_{t-1}^{self}. We denote the objective minimized here by $E(W_t)$. The utility-maximizing policy described below determines the intended action component of W_t^*, while $E(W_t)$ constrains the remaining components.

Only the first term in the inference equation is anchored to something outside the brain; the remaining terms impose consistency among representations and the four priors. Because W_t includes the discrete structure of the world model (i.e., which columns are active and which edges connect them), the search runs over structure as well as continuous content: the brain needs to infer not only what and where objects are and how they are related, but how many there are and which are connected.

Action. The self emits actions to drive its internal state toward its motivations. Using the world model to forecast the consequences of candidate actions, it selects the action that scores best given M^{self}:

$$A_t^{self,*} = \arg\max_a \; U(S_{t+\tau}^{self}, W_{t+\tau}, M^{self} \; \Big| A_t^{self} = a\Big).$$

Here τ is the planning horizon, the time over which the self looks ahead, and U is a utility function that scores the projected future against the self's motivations.

On what gets consciously perceived. I subscribe to dual-aspect monism: subjective experience and neural mechanism are two aspects of a single underlying reality, and every nuance of subjective experience is reflected in some mechanism of the brain. The proposal above makes this commitment concrete by specifying *which* mechanisms. The contents of subjective experience are precisely the latent variables on which the brain's recurrent inference loop converges: the retinotopic state X_t, the identity representation ψ_t, the intrinsic physical representation I_k of each tracked object, the transformation $G_{k,t}$ that situates it in space, the relations E_t among the objects, the body schema I^{self}, the self's pose G_t^{self}, the internal state S_t^{self}, the motivations M^{self}, and the intended action A_t^{self}. Every variable that appears in the fixed point W_t^* corresponds to some aspect of experience; conversely, every nuance of experience corresponds to some variable in W_t^*. Thus W_t^* defines the inclusion set for 'phenomenal binding,' the set of contents that are all assigned to a single consciousness [27].

This raises a puzzle. The brain has vastly more state variables than appear in W_t^*: the membrane potential of a single neuron, the phosphorylation state of a receptor, the expression level of a gene are all time-varying variables of the brain that are causally important for generating neural activity. They are part of the same physical system as the variables in W_t^*, governed by the same physics. Why is the brain conscious of I_k and $G_{k,t}$ but not of glutamate concentration in a particular synaptic cleft?

The answer comes from distinguishing two levels of description. At the *microstate* level, the state vector of the brain at any moment is extremely high dimensional, including every membrane potential, phosphorylation state, and gene expression level. At the *macrostate* level, that same physical system is described by a much lower-dimensional vector, W_t, a few hundred latent variables that represent the structure of the world. The macrostate is related to the microstate by a coarse-grained projection (just as temperature is a coarse-grained projection of the positions and velocities of individual gas molecules).

What makes the macrostate special is the role it plays in the dynamics. The macrostate is what *converges* to a stable fixed point W_t^*; the microstate does not converge. Membrane potentials spike and recover, phosphorylation cycles continue, genes switch on and off; the microstate is in continuous flux throughout, before and after the macrostate has settled. Subjective experience corresponds to the *converged macrostate*, not to the underlying microstate dynamics. From an evolutionary perspective, the macrostate convergence that arises from inferring an accurate, stable model of the world is what matters for survival. Evolution, unbeknownst to itself, was headed on a direct path to the greatest discovery of all time.

This theory of conscious access explains a large body of empirical data. In inattentional blindness, observers engaged in a demanding task, e.g., counting passes between basketball players, fail to notice a gorilla walking through the scene [28], even though the gorilla is fully visible and processed by visual cortex. In binocular rivalry, recordings from IT face patches in my lab have recently revealed an analogous phenomenon: the neural representation alternates rapidly between the conscious percept and the suppressed stimulus at ~15 Hz—remarkably, with equally good face representation fidelity during the two epochs [29].

Inattentional blindness and binocular rivalry are only two of many such dissociations, including visual masking, the attentional blink, and change blindness, in which feature-extracting cortex can robustly

encode a stimulus that never reaches awareness. In each, what is missing, I posit, is the binding of that sensory representation to an object-tracking column in PPC. Sensory representation can exist throughout cortex without being a constituent of W_t^*; only when a representation is bound to a column of C_t^*, through stabilization at the loop's fixed point, does it become part of conscious experience. The column, explicitly built to track persistent structure across transformation, is the gate of conscious access.

The strong claim is that this is exhaustive. There is no aspect of consciousness left over once W_t^* is specified, and no aspect of W_t^* that fails to be experienced. Conscious perception and the converged world model are two aspects of one and the same fact. The ultimate vindication of the theory will come from the ability to engineer consciousness by acting on the specific variables of W_t^*: writing percepts, thoughts, and internal states into the brain by setting the values to which the loop converges.

Working memory subspaces. Object-tracking columns create a compositional physical world model by locating all essential information about each object within a single persistent column. But cognition requires compositional organization in a more abstract realm—not of physical objects, but of cognitive tokens (e.g., items in working memory, abstract concepts) and their relations.

A remarkable series of experiments on sequence working memory from Liping Wang and colleagues has recently illuminated the basic architecture by which the brain accomplishes this. Wang's group trained monkeys to hold a short sequence in mind—a few locations seen in a particular order—while recording thousands of neurons in prefrontal cortex [6]. During the memory delay, they identified a set of low-dimensional subspaces of the neural-state space that each stored the spatial location of one item in the sequence. A later study asked what happens when the remembered sequence must be mentally reordered, sorted forwards or backwards on cue [7]. As the order changed, items were moved from one slot to another, with additional temporary subspaces recruited to support the reordering operation. Thus the subspace architecture is not merely a storage medium but a manipulable symbolic workspace in which items can be deliberately relocated. The work makes two major contributions: 1) clarifying the geometry of working memory representation, and 2) revealing the circuit basis for dynamic mental operations.

This framework can be generalized, through the formalism of tensor decomposition, to represent a complete language of thought, as originally proposed by Smolensky [30]. The core idea is to distinguish two kinds of information: a 'filler' (the content to be represented, such as a specific location or concept) and a 'role' (the slot into which content is placed, such as 'first in the sequence'). The challenge is to find a neural mechanism that binds role and filler while keeping them separable, so each component can be individually recovered later. For example, in the sentence "John loves Mary," we want a representation that preserves who loves whom. Adding the filler and role vectors, f + r, conflates them irreversibly. Smolensky's solution was to bind them with a tensor product, B = f ⊗ r, where ⊗ denotes the outer product. Knowing the role, we can multiply by r^T to recover the filler, and vice versa. A filler can occupy different compatible roles, and summing these bindings lets us compose arbitrarily structured meanings. For example, 'John loves Mary' is represented by B = fJohn ⊗ ragent + floves ⊗ rverb + fMary ⊗ rpatient. In this way, an infinite variety of thoughts can be represented.

Recent work shows that large language models use a simpler version of this scheme, achieving binding by adding a content-free tag (a 'binding ID') to each entity [31–33]. Because each entity occupies its own slot (its token) and is labeled by a binding ID written into a separate subspace, the label is never conflated with the content. This strategy is simpler because many distinct binding IDs can be encoded within a fixed-dimensional subspace and used to link content across token representations. This avoids the explicit construction of uncompressed higher-order tensors whose dimensionality grows exponentially with nesting depth.

Thus the logic of object-tracking columns presented earlier can be extended into the domain of pure thought. Object-tracking columns provide a persistent anchor (the column) for representing an entity in the physical world, while working memory subspaces provide an anchor freed from physical reality: a stable symbolic slot that can be bound to any content the mind wishes to hold, whether a remembered location or a step in a chain of reasoning. For both object-tracking columns and working memory subspaces, the crucial insight is the same: *the brain requires a persistent structure to which conscious content can attach.*

For a thought to become conscious, I propose that these cognitive slots must remain coupled to the perceptual and self-representations they engage. This coupling can occur through mental imagery, explicit self-reference ("I think, therefore I am"), or most fundamentally, the self's ongoing motivation to understand, act, and express itself. Thus *the contents of thought form an integral part of the larger world model* W_t, with a thought becoming conscious when these coupled representations converge on a coherent state.

Object-tracking columns and working memory subspaces differ fundamentally from *feature tuning*, the dominant framework in systems neuroscience for decades. The latter framework holds that neurons are tuned to fixed feature values, e.g., a specific orientation or a particular face or object. Object-tracking columns and working memory subspaces, in contrast, are not feature detectors but containers, persistent structures that can be bound to any content. In computer programming terms, classical neurons encode values, fixed instantiations of a specific feature. Object-tracking columns and working memory subspaces encode *variables*, slots that can take any value, be queried, and be rebound to new content. My central thesis is that the variable slots that persistently track physical matter are the key to consciousness.

Dynamic convergence. Up to here, I have focused on the representational requirement for consciousness: persistent slots. But a slot is only a container, and tracking requires something to track *against*: the columns maintain their contents by predicting the retinotopic and identity representations and correcting when the predictions fail. The retinotopic representation, X_t, and the identity representation, ψ_t, are essential partners in the loop. They also carry much of what we actually experience: we are conscious of a fine edge orientation, a subtle shade of color, or the detailed identity of a face, yet such content is encoded not in slots but in areas that rely on classical feature tuning. Such content does not become conscious on its own, as binocular rivalry shows [29]. What makes it conscious is *dynamical*: the content must be drawn into a recurrent loop with a persistent anchor, and that loop, distributed across multiple areas, must converge on a stable fixed point W_t^*. I present two arguments for this claim, both

admittedly speculative: one from bistable illusions and one from the brain's ability to introspect on its own causal structure.

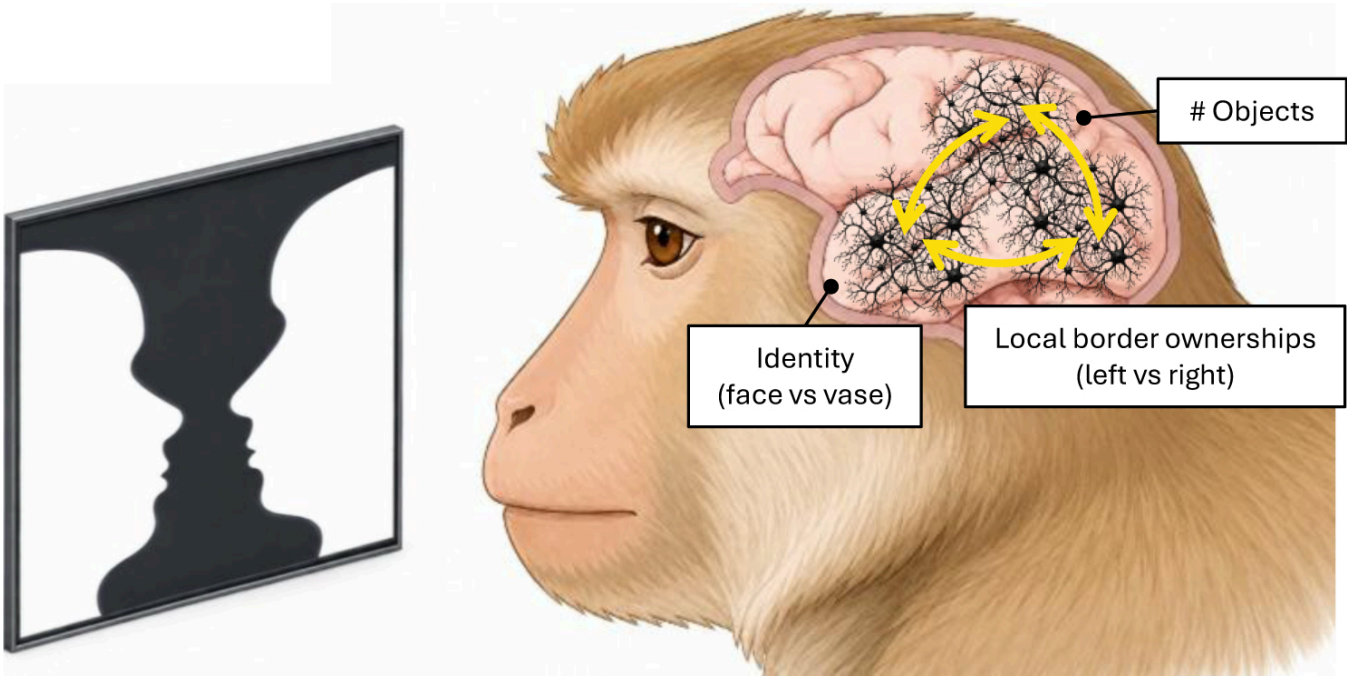


Figure 4. Conscious perception requires coherence across distributed neural representations.

Bistable illusions. Bistable illusions like the face-vase illusion, in which two profile silhouettes alternate with a vase, make vivid the conclusion that conscious perception corresponds to the *converged macrostate* W_t^*. Even a child finds it magical to watch a static image flip; but for someone who knows something about the brain, it becomes truly remarkable: reaching any interpretation requires vast numbers of neurons encoding different features at different spatial scales to collectively agree (Figure 4), and that agreement must span the entire visual hierarchy, from early areas like V2, where neurons encode local border ownership, to posterior parietal cortex, where neurons encode the number of objects, to inferotemporal cortex, where neurons encode object identity. What reaches awareness is only the *final result* of this computation. This is perhaps the sharpest divide between a layman's and a neuroscientist's understanding of the brain: the layman takes the clarity of conscious experience for granted, while the neuroscientist is aware of how much *hidden computation* it requires. We don't perceive this hidden computation, only the stable states on which it converges; the abrupt, all-or-none flip between interpretations strongly indicates that what we consciously perceive is a *converged state*.

Causal introspection. Consider a thought experiment. Suppose one were to snip all the axons in the brain and voltage clamp every neuron to the value it would have at the fixed point W_t^* of some percept. Would the resulting brain be conscious of that percept? Most neuroscientists would agree the system would *not* be conscious, because, in this scenario, the neurons would not be causally connected, and it is hard to imagine how a vat of dissociated neurons could be conscious. Therefore, causal structure—not just the values of the macrostate variables, but the dynamics that maintain them—must be essential to generating consciousness.

If causal structure matters, a puzzle arises: how can counterfactuals, which are necessary to define causal structure, matter for your conscious state if you never experienced them? Here we see a powerful consequence of identifying consciousness with the fixed point of a convergent loop. The inference dynamics can be characterized by the energy function $E(W_t)$ defined earlier, which penalizes mismatches between latent states across different brain areas and imposes persistence, exclusive binding, and parsimony on the scene graph. At the fixed point, $\nabla_W E(W_t^*) = 0$: the system is at a stable stationary point in state space, and small perturbations away from W_t^* increase E, driving the system back. Being at W_t^* therefore implicitly encodes the result of an infinite number of counterfactual experiments: *at the stable point, the brain is effectively able to perceive its own causal structure.* Specifically, we hypothesize that the dynamics maintaining the converged state make the system's local causal structure available to it. Encoding conscious percepts as stable states of dynamic convergence thus answers one of the hardest questions about consciousness: how the brain is able to intrinsically access its own causal structure (without the benefit of a two-photon microscope, to paraphrase Richard Axel). Moreover, the framework suggests what causal structure is critical: the curvature of E in the neighborhood of the occupied fixed point within the space of latents, not the entire causal architecture of the brain as a physical entity.

Summary. I have argued that consciousness is the brain's process of constructing and maintaining a persistent, compositional world model spanning objects, thoughts, memories, internal states, and the self. The brain solves the fundamental challenge posed by fleeting sensory input by generating stable symbols —representations that persist across time despite ongoing transformation—and binding them into a coherent internal model of the world and the self within it. I propose that the elementary building blocks of this world model are persistent object-tracking columns in posterior parietal cortex, which extract stability by factorizing each object into an intrinsic representation and a transformation, and that working memory subspaces in prefrontal cortex build on these columns to support abstract thought. Conscious experience corresponds to the converged macrostate of recurrent inference loops spanning the cortical hierarchy (global *convergence*, not global broadcast [34]): every variable represented in this stabilized world model corresponds to some aspect of subjective experience, and conversely, every aspect of subjective experience to some variable in the world model. Intelligence, on this account, is a downstream process that flexibly recombines and transforms these conscious symbols to support reasoning, planning, and action. This framing transforms consciousness from a metaphysical mystery into an engineering problem: to understand consciousness, we must understand the representational and dynamical principles by which the brain constructs its world model; to engineer consciousness, we must learn to causally write into those representations, and ultimately, to build them *de novo*.

PART 2

Scientific strategy for understanding and engineering consciousness

Once consciousness is framed as a compositional world model implemented by a brain-wide dynamical system, it becomes clear that we must understand in detail both the *representational* and *dynamical* structure of this world model.

Representational structure. We hypothesize that the brain's world model is built from a set of persistent symbolic structures representing physical objects, thoughts, and the self. As argued in Part 1, we believe the key bottleneck for understanding consciousness is the mechanism underlying persistent tracking of physical objects. Accordingly, *one of the first major goals of the experimental program is to identify and characterize the neural structures that support persistent physical object representation.* We hypothesize that a key piece of the solution is persistent object-tracking columns. Regardless of whether this theory is correct, it provides a concrete framework for experiments to discover the brain's solution to this fulcral problem, about which we are currently clueless.

The broader goal of the program is to identify the complete inventory of persistent symbolic structures underlying conscious experience. In particular, we hypothesize that abstract thought is supported by persistent structures analogous to object-tracking columns. One candidate is working memory subspaces in prefrontal cortex that serve as stable slots for cognitive tokens (concepts, goals, and the abstract relations among them). A second major goal of the program is therefore *to identify and characterize the full set of cognitive structures, together with the operations that bind, manipulate, and unbind their contents.*

Finally, the same principle applies to the self. We hypothesize that the brain represents itself as an object writ large: a persistent structure that tracks the body as a physical entity and binds it to internal state, motivations, and intended actions. A third major goal is to *identify the neural structures underlying the self-model and to determine how they interface with representations of the physical world and thought.*

Dynamical structure. We hypothesize that the persistent symbolic structures underlying the brain's world model exist as fixed points of recurrent inference loops spanning the cortical hierarchy: latent variables in each brain region constrain those in other regions through these loops. How is brain-wide representational coherence achieved? This is a fundamental problem about which we currently remain clueless.

Because conscious symbols emerge only through large-scale recurrent interactions, *we must take a dynamical systems approach*, perturbing specific nodes and measuring how activity propagates across the *network* toward or away from convergence. We hypothesize that some cortical regions (namely, the ones harboring persistent slots) are essential drivers of convergence that results in conscious access, while others can be flexibly recruited to elaborate conscious content without being strictly necessary.

Once we gain deep understanding of the representational and dynamical structure of the world model, *the brain becomes a programming interface.* Columns and subspaces are the variables, their contents the values. We will be able to 'mentally program' percepts and thoughts into consciousness, e.g., by assigning new contents to an object-tracking column or working memory subspace to construct a specific percept or thought. The success or failure of these write-in experiments will provide a decisive test of the framework.

Next, I present a roadmap articulating specific experimental milestones we hope to reach in the next five years, together with the technology required to achieve them. This roadmap will evolve, of course, as

findings from new experiments (both ours and others') and technical challenges along the way reshape our thinking and approach.

Scientific and technical roadmap: Astera Neuro years 1–5. The framework above defines our experimental agenda: we need to find the persistent symbolic structures of the brain's world model, understand the recurrent dynamics that stabilize them, and causally manipulate these dynamics to engineer conscious experience. We are building Astera Neuro to pursue this vision at the scale the problem requires.

We will pursue fundamental systems neuroscience research in rodents, macaques, and humans in parallel, capitalizing on the unique strengths of each. Macaques provide a tractable primate system in which we can dissect, at single-cell and circuit resolution, the representational codes and dynamic interactions underlying consciousness. Human studies will test the generality of these principles while probing domains that are unique to humans, including language and mathematical reasoning. Rodents will provide a rapid experimental platform in which causal intervention methods can be validated before deployment in primates, while also clarifying which core mechanisms of the primate world model are evolutionarily conserved across mammals.

Supporting all three species programs is a coordinated technology effort designed to make large-scale causal interrogation of the brain's world model possible. Key priorities include chronic high-density electrophysiological recording, high-bandwidth neural imaging and holographic optogenetic stimulation, and tools to precisely map the large-scale functional architecture of the brain.

We next present the specific research agenda: a brain-wide functional and anatomical atlas; the macaque, human, and rodent programs through which we will identify the representational and dynamical structure of the world model; computational models to interpret the recordings and predict responses to perturbations; write-in experiments that will decisively test our understanding; and finally the technology we will build to make this all possible.

Mapping the functional architecture of the world model. Francis Crick said, "In biology, if seeking to understand function, it is usually a good idea to study structure" [35]. Functional studies must be deeply informed by brain anatomy. In Part 1, I characterized consciousness as a particular kind of world model: generative feedback loops converging on symbols that possess both persistence and elaborated features. The structures that participate in these loops are anatomically specific, and the dynamics are defined by how they connect, so neither representation nor dynamics can be understood without anatomy.

We will thus launch a large-scale effort to build a complete functional and anatomical connectivity atlas of the macaque brain, combining cutting-edge mapping using functional ultrasound imaging (fUSI) with whole-brain connectome tracing. This atlas will provide a foundation for both our representational and dynamical understanding. Critically, we will construct this atlas in a single animal, to remove the noise that inter-animal variability introduces when functional and structural data are pooled across subjects.

To construct the atlas, we will map function with fUSI. We will develop an ultrasound transducer array to image activity across the entire brain in a behaving animal at sub-millimeter resolution as animals run a large battery of localizers, including objects, faces, motion, scenes, social stimuli, cognitive tasks, and manipulations of internal state. We will then map the structure of the same brain, starting with whole-brain sparse labeling to fully reconstruct ~100,000 neurons. As the necessary technology matures, we will extend this to a dense projectome, for example by reconstructing every myelinated axon above ~1 μm in diameter (in the macaque, ~30% of all white-matter axons [36]). This combined functional and structural atlas will provide a common anatomical framework for all our macaque experiments, allowing us to target specific sets of connected functional nodes and register recordings across sessions and subjects.

Macaque program. The goal of the macaque program is to understand the neural codes for world and self and to develop the capability to control these codes, focusing on the major components of the world model: perception, thought, and self-modeling. We will initially use existing atlases to guide where we record and stimulate, refining our targets as the new atlas develops. Initial studies will employ Neuropixels recordings, transitioning to polymer probes that will enable chronic, long-term access to the same neural populations. Experiments will span passive viewing on screens, immersive virtual reality, and real-world environments, enabling us to study the neural representations underlying perception, thought, and the self-model under progressively more naturalistic conditions.

Perception. We will investigate the neural codes underlying perception of the 3D physical world, focusing on the mechanisms that generate persistent, compositional representations of objects. It has been nearly 70 years since Hubel and Wiesel discovered orientation-selective cells in V1, yet conceptually, we are not much closer to understanding how the brain creates our essential visual experience of a stable, cyclopean world of 3D objects that transform lawfully as we move. We aim to close this gap, building on our new framework of object-tracking columns. Using large-scale neural recordings, we will study how neural populations represent dynamic multi-object scenes, including the 3D shape, material properties, pose, and position of individual objects, as well as the spatial and physical relationships between them.

We will probe these representations with a range of tasks, including chasing games and CAPTCHA-like recognition tasks. In chasing games, the animal pursues one moving object while evading another, so it must track both continuously without confusing which is which [37], the core computation we attribute to object-tracking columns. In CAPTCHA-like recognition tasks, animals will be trained to decipher crowded, overlapping characters step by step, a segmentation problem that hinges on explaining away: once one character accounts for a contour, that contour no longer supports competing interpretations [38]. This paradigm will let us systematically investigate how representations across multiple areas "snap into place" and converge to a stable macrostate.

Across all these experiments, we will record from multiple areas in parietal, frontal, and inferotemporal cortex to determine how representations are parceled out among them. We will complement these recordings with binocular rivalry experiments that directly probe how representations are coordinated across areas. These experiments build on a recent finding from my lab that IT cortex strongly represents both the conscious percept and the suppressed content, alternating between them ~15 times per second [29], raising the question of how a single, stable percept emerges from this alternation. Together, these

experiments will identify how objects and their relationships are represented across cortex and how recurrent interactions stabilize these representations into a coherent percept. These results will guide the design of stimulation patterns for writing specific visual percepts into the brain.

Thought. We will investigate the neural representations underlying thought, reasoning, and action planning, focusing on how the brain implements the mental programs that subserve generative thought. How are abstract variables (e.g., objects, concepts, relations, goals, and rules) represented, dynamically bound together, and manipulated across distributed neural populations during cognition? Through large-scale recordings during working memory, navigation, planning, and other cognitive tasks, as well as during passive viewing of videos depicting complex, structured events, we aim to characterize the neural dynamics that support structured thought.

Beyond these standard paradigms, we plan two more exploratory projects aimed at isolating the building blocks of complex thought. In the first, we will attempt to teach monkeys a pictographic language with a grammar for describing events in videos, testing whether they can produce and interpret novel combinations of symbols. This will let us probe how syntax is represented and how structured thoughts are composed. In the second, we will train monkeys to imitate humans performing progressively more sophisticated and abstract tasks, with the goal of teaching them the abstract concept of "imitation" itself and thereby opening a door to one-shot learning of complex rules. These studies aim to establish the foundation for causally writing novel thoughts, goals, and action plans into the brain through targeted neural stimulation.

Self-model. We aim to identify the neural basis of the self-model: the brain's representation of itself as a single, persistent entity encompassing the body and its pose, internal state, motivations, and intended actions. We hypothesize that the same object-tracking mechanisms used to represent external objects also track the self: the self is an object writ large. Through large-scale recordings in posterior and medial parietal cortex while animals spontaneously change their body configuration, we will probe the body schema and how it binds to variables encoding emotion (insula, amygdala), motivation (insula, orbitofrontal cortex), and action plans (premotor cortex). Among motivations, curiosity is of special interest because it lets us probe how the brain integrates its model of the world with its model of the self: how does the brain represent the world's underlying structure together with the animal's own knowledge, uncertainty, and possibilities for acting within it? These efforts to understand the brain's self-model will lay the foundation for rewriting self-representations, including modulating internal states, reassigning body ownership, merging neural systems, and transferring core aspects of selfhood to non-biological substrates.

Perception, thought, and the self-model do not operate in isolation. Characterizing their interactions will be an important aim of the program. For example, primate cognition is deeply grounded in physical reasoning. An example my colleague Dileep George likes to give is the following: if someone tells us, "Joe wants to pound a nail into the wall," and asks, "What is the orientation of the nail?" we answer effortlessly by *imagining the physical scene*, demonstrating the close link between cognition and perception. Similarly, the fact that so many of our thoughts take the form "I think…," "I want…," "I feel…" points to a close link between the cognitive system and the systems that represent internal state and the self. Understanding

these links will not be an afterthought but a fundamental part of our effort to reveal how the brain builds an integrated world model and is only possible because we will pursue all three systems together under one roof.

Human neuroscience program. We plan to launch a global-scale Human Neuroscience Program aimed at extending understanding of the brain's world model into domains uniquely elaborated in humans, particularly language and mathematics. To collect human single-neuron data at scale, the program will draw on three complementary recording platforms: chronic Utah arrays in brain-computer interface (BCI) patients, stereo-EEG (SEEG) electrodes during clinical epilepsy monitoring, and acute Neuropixels recordings during neurosurgery. Experiments will adapt the macaque program's core paradigms (physical world-model experiments, structured working memory tasks, and imitation-based approaches to physical reasoning), enabling direct comparison between species, and will add what only humans can provide: language and math. For example, mirroring the captioning and imitation paradigms in monkeys, human subjects will imagine, watch on video, perform, and verbally describe the same operation, e.g., swapping the first and last objects in a sequence; comparing these conditions will show how the brain translates between its physical world model and language. This understanding of basic representational mechanisms in the human brain will pave the way for our long-term goal: merging human and artificial cognitive systems in ways that expand and enhance human consciousness and cognition, and ultimately transferring core aspects of the human mind to non-biological substrates.

Rodent program. We will use rodents as a rapid and scalable platform for developing causal intervention methods before deployment in primates, while also probing which principles of the brain's world model are conserved across mammals. Although it remains unclear whether rodents support the richer forms of world modeling underlying primate consciousness, they may possess some of the core mechanisms relevant to building a stable, compositional world model. Initial studies will focus on working memory: do rodents possess subspaces for representing distinct items held in mind, and can we write to these subspaces to alter behavior?

Computational modeling. The experimental program will generate neural recordings of unprecedented scale, spanning multiple brain areas simultaneously during perception and cognition. To extract scientific understanding from these data, we will need computational models that serve two complementary functions.

First, we need models that capture the full variance of neural population activity across areas while distilling the core representational format: what are the fundamental variables the brain uses to represent objects, thoughts, and self, and how are these variables encoded in neural population geometry? For example, does the representation of an object in posterior parietal cortex factorize into intrinsic properties and transformation as our theory of object-tracking columns predicts? How are role and filler represented in prefrontal cortex? Fitting large-scale models to recordings from multiple areas will help us discover representational structure that our hand-designed analyses might miss.

Second, we need models of how brain areas influence one another and how their activity converges to the states we identify with conscious access. The write-in experiments pose an inverse problem: given a target

percept or thought, what input will drive the brain-wide system to the corresponding fixed point? The models must predict responses to novel perturbations accurately enough to tell us where and when to stimulate.

We will develop these models together. Modeling the dynamics requires knowing which variables are changing, and understanding a representation requires knowing how it is updated and stabilized. Both are needed for the write-in experiments. These experiments will, in turn, refine the models. We will use the brain's measured responses to update the models, and use the improved models to bring the brain closer to the target state.

Write-in milestones. The strongest test of our understanding of consciousness will be the ability to control it: to generate a specific percept, thought, or feeling by precisely manipulating the persistent symbols that arise when brain-wide recurrent loops converge. Such precision has only recently become possible. Classical electrical stimulation has long been known to alter conscious experience, evoking percepts, movements, emotions, and memories [39], but it activates large, heterogeneous populations without regard to representational structure. Recent advances in holographic optogenetics change this: by stimulating identified neural populations with single-cell precision, it becomes possible to write specific target percepts, emotions, and thoughts into the brain [40,41].

Across rodents and macaques, we will pursue a set of high-impact write-in experiments designed to serve as critical milestones toward merge and upload. These experiments include: (1) writing an arbitrary visual percept, (2) writing a specific thought, (3) writing a targeted emotional or internal state, and (4) rewriting body ownership.

The experiments will start with simple interventions and advance to more complex ones. For perception, we will begin with writing simple object features (e.g., a disk versus a cross), advance to more elaborate representations (e.g., a specific face identity), and ultimately extend to full scenes containing multiple objects and dynamic relations. For thought, demonstrations will progress from short sequences (e.g., a working memory sequence of three locations), to structured symbolic content (e.g., "five bananas and three apples"), to complex, naturalistic, goal-directed thoughts (e.g., "I want to pick up all the red sticks on the floor and arrange them into a rectangle"). If we achieve this last goal, we will in effect be able to talk to a monkey. For internal state, we will begin with changes in basic dimensions such as valence and arousal, then progress to emotions tied to specific objects or situations, such as curiosity about a particular object. For body ownership, we will begin by altering the felt ownership of a single limb, then progress to inducing the sense of owning an entire virtual body. Importantly, we already have sufficient understanding of the relevant neural codes to undertake the initial demonstrations.

Technology development. The experiments above demand capabilities that do not yet exist. We will therefore undertake a large-scale effort to develop technology for chronic large-scale recording and high-bandwidth holographic optogenetic control.

Polymer probes. Neuropixels probes have transformed our ability to decipher neural codes, enabling simultaneous recordings from hundreds to thousands of neurons. But their rigidity and fragility make

them unsuited to our goal of long-term, brain-wide recording in freely behaving macaques. We are building a polymer probe platform to enable long-term, large-scale recording and stimulation across the macaque brain, with a path toward human-compatible systems. We want to record from hundreds of neurons per column in hundreds of columns across the entire brain, at depths up to 6 cm. The recordings must remain stable for months to years while the animals behave freely, and the probes must deliver precisely patterned electrical stimulation. To reach these targets, we will need to improve probe fabrication, headstage and chip integration, insertion technology, acquisition systems, and surgical workflows, including robotic implantation.

Holographic optogenetics. Holographic optogenetics is uniquely suited to generating activity patterns precise enough to control percepts, thoughts, and actions. To realize it in primates, we will pursue several engineering strategies in parallel: (1) spatial light modulators that are larger, faster, and finer in pixel pitch; (2) high-speed acousto-optic beam-steering that generates flexible spatiotemporal patterns across large populations for scalable closed-loop control; (3) aberration-corrected GRIN lenses and side-imaging microprisms to reach the deep brain areas critical for non-human primate experiments; (4) multi-arm microscopes that image and stimulate multiple sites simultaneously; (5) miniature two-photon microscopes that carry onboard spatial light modulators, enabling holographic control in freely behaving animals; (6) three-photon imaging and optogenetics, whose longer wavelengths and reduced scattering reach subcortical depths beyond standard two-photon methods; and (7) more sensitive opsins with reduced spectral cross-talk with GCaMP and voltage indicators.

High-bandwidth human recordings. Existing human depth recording systems are limited to low channel counts and coarse field-potential measurements. We are supporting development of a system that deploys a 128-contact laminar microelectrode array through SEEG electrodes already used in clinical care. This will enable stable semi-chronic single-unit recordings in humans for up to a month. The system will provide a powerful platform for high-bandwidth single-unit recordings during naturalistic human cognition and behavior, using existing clinical workflows.

Biohybrid neural implants. Even if our optical neural interface reaches 100,000 neurons, this may still be insufficient by orders of magnitude given the scale of the primate brain. Science Corp. is pursuing a novel BCI architecture in which stem-cell-derived neurons, individually seeded into a 100,000-microwell array, grow axons into the underlying tissue and form synaptic connections with the host brain, with each implanted neuron synapsing onto ~1000 host cells. In proof-of-concept studies, optogenetic stimulation of the implanted cells was sufficient to guide goal-directed behavior in mice [42]. We envision working with Science Corp. to use the biohybrid implant as a massively parallel axonal input to the brain, while polymer probes simultaneously record the population response to each stimulation pattern.

Medical impact and animal welfare. Deep understanding of the brain's internal model of the world and the self, together with the brain-machine interfaces we are developing, will open fundamentally new approaches to treating neurological and psychiatric disease including Alzheimer's, depression, schizophrenia, PTSD, OCD, and insomnia. A 2016 analysis estimated that mental and neurological disorders imposed more than $1.5 trillion in annual costs on the US economy, equivalent to nearly 9% of GDP [43]. More importantly, they are among the greatest sources of human suffering, because they alter

perception, thought, memory, and ultimately the sense of self. Reducing this suffering is a central motivation for our work. We believe that uncovering the fundamental principles underlying the brain's internal model—including how content reaches conscious access—will enable new *neural representation therapies* that directly target the aberrant representations and dynamics involved in these disorders. If we can use this understanding to restore the neural representations that underlie healthy cognition and experience, we will not only transform medicine but also provide compelling evidence that we genuinely understand the brain's internal model.

This work requires experiments in macaques and rodents. The premise of our research, that animals build models of the world and self through mechanisms they share with us, is also why we must treat them with the greatest care. We are committed to conducting this work under the most humane conditions possible. Our animals will live in state-of-the-art facilities with ample room to move and play. Our scale lets us go further: because we employ dedicated engineers building custom-optimized implants, surgical systems, and behavioral rigs, every aspect of the experimental pipeline can be optimized for animal comfort, which is hard for any single lab to do on its own.

Astera Neuro and Astera AI. The central question we are pursuing, how the brain builds a model of the world and the self, is also at the heart of Astera AI's mission: to build a neuro-inspired *plannable world model*. Plannable world models have internal representations that allow for implicit, parallel, distributed searches over possible futures, arriving at solutions through a settling process much like the way perception snaps into place. Such systems would go beyond current AI architectures by modeling the world as humans do, capturing not only its physical structure but also its social and conceptual dimensions, and supporting the flexible imagination and planning that characterize human intelligence.

In contrast to AI systems trained primarily on text, the models Astera AI is building learn autonomously by interacting with the world. The hierarchical and compositional abstractions that an agent learns from experience can then be used to simulate the consequences of novel actions and to ground language. The models Astera AI is researching have direct counterparts in systems neuroscience, including the visual cortex, prefrontal cortex, hippocampus, and other brain systems, and their integration into an agent will give rise to a new cognitive architecture and let us test theories of conscious access computationally, including the framework laid out in Part 1. In addition, these models will replicate the algorithmic principles underlying the columnar and laminar organization of the cortex, as well as its interactions with the thalamus and hippocampus.

Examples of neuroscience-inspired models previously developed by Astera AI researchers include the Clone-Structured Causal Graph framework [44,45], which is motivated by a large body of experimental findings from the rodent hippocampus; the Recursive Cortical Network model of visual perception [38,46]; and the program-induction model of concept learning [47]. Newer versions of these models draw on recent advances in deep learning, energy-based modeling, and causal representation learning. Although these models employ ideas such as latent-variable hierarchies and dynamic inference, there remains a huge gap between these conceptual frameworks and a fully functioning plannable world model with human-level performance.

Bridging that gap requires understanding how the brain actually accomplishes this feat, which is precisely what Astera Neuro's experimental program seeks to reveal. We plan to work closely with Astera AI at every stage, from collaboratively designing experiments and jointly interpreting datasets to ultimately distilling the core principles by which the brain builds a plannable world model. The strongest validation that we understand these principles will be an artificial system that can flexibly perceive, reason, and act in the real world as humans do.

Astera Neuro and Astera Simplex. Transformers are an astonishing proof of concept that a system built on simple computational rules can develop remarkable intelligence when trained at scale. Mechanistic interpretability, the effort to understand the internal representations and computations that emerge through this training, is beginning to illuminate many of the same questions we confront in neuroscience. Adam Shai, Paul Riechers, and the Simplex team at Astera see this as an opportunity to develop a fundamental science of mind. What is a representation? What makes a representation a cognitive symbol, and what does it mean for a network to compute or form an abstraction? We need precise answers to these questions to understand both our own cognition and the increasingly powerful artificial systems we are building. These answers should guide what we look for and how we design experiments, whether in a transformer or in a brain.

They begin with a question central to our own program: *how does a system retain what matters about its past in representations that persist, update, and support prediction?* For a given generative process, their framework specifies which distinctions about the past matter for predicting the future and how these distinctions must update as new observations arrive. In their initial work, they showed that transformers trained on hidden Markov sequences develop internal representations whose geometry reflects Bayesian-optimal beliefs about the hidden states [48]. More recently, they showed that transformers learn to represent conditionally independent components of a generative process in approximately orthogonal subspaces, even when the token stream contains no explicit labels identifying those components [49]. Each subspace can then serve as a variable holding the model's current belief about one component. This could provide a basis for persistent symbols.

These results bear directly on one of our central insights: perception and cognition are both compositional generative processes. They give us specific representational geometries to look for in the brain. Their more recent work shows how a network can infer both which process is generating its input and what state that process is in [50], providing a starting point for understanding flexible composition. In our framework, object-tracking columns and working memory subspaces must be recruited, released, rebound to new content, and combined to perform new computations.

The framework in Part 1 proposes the fundamental elements of conscious experience. With Astera AI, we aim to test these ideas in functioning world models; with Astera Simplex, we aim to understand when and why these representations and computations must emerge in systems that learn to predict a structured world. Together, Astera Neuro, Astera AI, and Astera Simplex aim to combine neural experiments, computational modeling, and mathematical analysis to build a unified theory that can predict which percepts and thoughts arise from a given pattern of neural activity and can specify patterns that would generate a desired conscious experience.

Organizational culture and open science. The rise of AI has demonstrated what can happen when intelligence becomes an engineering problem. We believe consciousness is ready for the same transition. Our mission is to understand how the brain constructs a coherent internal model of the world and the self, with enough precision to engineer conscious experience. The implications extend far beyond neuroscience. Understanding these mechanisms could let us extend consciousness beyond its original biological form, fundamentally changing what we can become and the nature of our relationship to AI (see Part 3). We seek individuals who recognize the scale of this opportunity and hunger to take this moonshot with us. For us, the traditional rewards of academia—an independent lab, the personal glory of a CNS paper, tenure—pale beside the possibility of uncovering the mechanism by which matter becomes conscious. To take this miraculous property that emerged through evolution and bring it under deliberate control will be an enormous scientific challenge. We need an organization capable of pursuing this ambition at the scale the problem demands.

Ours will not simply be an institute of well-funded PIs. Instead, we envision six scientific cores (perception, cognition, self-model, rodent neuroscience, human neuroscience, computation), each composed of research scientists and technical staff. The cores will be organized horizontally, with strong links between them. A central challenge our new organization must address is how to promote genuine teamwork, given that we are working against the natural human ego. The first-author paper in a prestige journal, the chief measure of success in the current academic system, is extremely harmful to teamwork, as it gives almost all credit to the first and last authors. We will therefore change both how we publish and how we recognize contributions. We will forgo publication in peer-reviewed journals and publish only through preprint servers and open repositories. We will also recognize everyone who contributes to the shared scientific program individually, without a hierarchy based on author order.

Individual scientific freedom is just as vital to our success as teamwork. At first glance, academia may seem to have the advantage here. But we believe that for scientists aligned with our core mission, Astera Neuro offers **far more freedom** than traditional academia. First, by forgoing publication in traditional journals, we free ourselves from the control of reviewers and can set our own scientific direction. Second, long-term support from the Astera Institute gives us state-of-the-art tools and infrastructure at scale, freeing us to focus on science rather than writing grants to secure resources.

Our scale will let us build a shared experimental pipeline, engineered for efficiency, with dedicated specialists carrying out much of the technical work and with AI automating as many steps as possible. Together, these capabilities will free our scientists to pursue fundamental questions that transcend species and disciplines. A human neuroscientist could design experiments in macaques, and a macaque neuroscientist could design experiments in humans. Similarly, we aim to demolish the traditional distinction between theoretical and experimental neuroscientists. Theorists will be empowered to design experiments and obtain data to test their ideas directly. Experimentalists, freed from day-to-day management of experiments, will be empowered to grapple with foundational conceptual questions that determine which experiments to run.

This scaled infrastructure will allow us to pursue many competing hypotheses in parallel, reducing the need to choose between promising ideas due to resource scarcity. A scientist at Astera Neuro who wakes

up in the morning with a burning question should be able to program an experiment through an LLM-based platform by noon and begin analyzing the results that evening. What unites us is the mission, not a shared theory: we want colleagues whose hypotheses differ from our own, because disagreement will drive us to design the most revealing experiments.

We intend to openly share everything we produce, including methods, results, datasets, code, and technology. We do this not simply to be generous, but because we believe an effort as ambitious as ours will require the ingenuity and insight of a much larger community. We hope to create a flywheel of innovation, where our discoveries and ideas can inspire new discoveries and ideas that will come back to us, *ad astra*.

Finally, we see our effort as one among a growing number at the birth of the new science of intelligence. For this science to flourish, nothing matters more in the long run than educating the next generation. Even though we operate outside academia, we intend to play an active role in this process. We have started a postbaccalaureate program targeted at college graduates and are planning an internship program for graduate students. We are also exploring a new model of graduate education outside academia, organized around fundamental problems in intelligence science, that will train scientists who move fluidly between theory and experiment.

Summary. We have outlined a scientific program to discover the fundamental building blocks and dynamical structure of the brain's compositional world model: the persistent neural structures that represent objects, thoughts, and the self, together with the recurrent interactions that stabilize them into coherent experience. We will pursue these questions through coordinated experiments in macaques, humans, and rodents, leveraging their complementary strengths. A functional and anatomical atlas of the macaque brain will guide large-scale recordings and perturbations, supported by new technologies for chronic recording and holographic optogenetic control. Computational models will identify the brain's representational variables and predict how their coupled dynamics respond to intervention. This will let us tackle the central inverse problem: given a desired percept, thought, or internal state, what input will drive the brain into that state? The success or failure of these write-in experiments will provide a decisive test of the framework. The same understanding could open new approaches to treating neurological and psychiatric disease by restoring the representations and dynamics underlying healthy cognition and experience. Collaboration with Astera AI and Astera Simplex will clarify the general computational principles expressed by the neural mechanisms we discover and let us test these principles in artificial systems. Achieving this vision requires systems neuroscience at a new scale, and that is what we are building at Astera Neuro.

PART 3

Engineering consciousness

The roadmap outlined in Part 2 is ambitious but falls within the current neuroscientific paradigm: identify the representational structures underlying conscious experience, characterize the recurrent dynamics that stabilize them, and learn to causally manipulate those dynamics. Yet if this program

succeeds, its impact will extend far beyond systems neuroscience. With a deep mechanistic understanding of consciousness, we will acquire the ability to engineer consciousness along all the degrees of freedom we uncover, far beyond what is possible currently through learning, drugs, or coarse stimulation.

This opens the door to remarkable possibilities: introducing new sensory modalities, expanding cognition through direct integration with artificial systems, merging minds, and transferring consciousness across substrates. At this point, the discussion inevitably begins to sound slightly unhinged. These are ideas that have historically belonged to science fiction and philosophy, not experimental neuroscience. Yet the framework for consciousness laid out in Part 1, together with the roadmap in Part 2, makes them concrete and tractable engineering challenges.

If conscious experience arises from understandable representational and dynamical principles, then augmentation, merge, and upload are no longer questions of *whether*, but of *what* must be true of the underlying representations and dynamics in each case. In particular, if continuity of self is a computational process carried out by the brain, then the question becomes: how can we make this continuity portable? In this final section, we treat these possibilities as achievable and attempt to sketch what it would take.

New sensory modalities. We begin with a conceptually simple augmentation: adding a new sensory modality. Classic experiments in ferrets showed that cortical tissue normally devoted to hearing can develop visual representations and support visual behavior when supplied with appropriate inputs [51,52]. These results suggest that the cortex contains general computational machinery that can be adapted to different sensory modalities. The framework in Part 1 already suggests a rough recipe for building an artificial sensory module and bringing its representations into conscious experience.

Let's take sonar, the use of reflected sounds to detect and locate objects. Following the framework of Part 1, we would need three essential elements: (1) a sonar processing hierarchy analogous to the ventral visual stream, dedicated to representing sonar-defined object shape; (2) a parallel sonar input to the multimodal object-tracking machinery in posterior parietal cortex, supplementing existing senses with evidence about object locations and spatial relationships provided by sonar; and (3) bidirectional coupling between the sonar shape hierarchy and the object-tracking machinery.

The coupled system would explain changing echo patterns in terms of movement and occlusion, maintaining a persistent representation of each object even when its echoes temporarily disappear. Sonar would thus enter the convergence loops that stabilize the brain's world model, allowing us to consciously experience objects through their echoes. Could this bring us closer to knowing what it is like to be a bat?

I believe experiments of this kind, augmenting human sensory capabilities with AI, also offer the only viable route to testing AI consciousness. *Can representations residing only in the AI become phenomenally bound to our own consciousness?* The framework in Part 1 predicts that if we build the sonar shape hierarchy above using feedforward digital logic, we will not be conscious of its internal representations. But if we implement the same representations in a recurrent circuit whose physical dynamics stabilize a

shared macrostate across levels of the sonar hierarchy and the brain, such that latent variables from the AI become components of an expanded W_t^*, then we will.

Expanded cognition. One of the most existentially troubling aspects of AI is the growing gap between what it can produce and what humans can understand. In mathematics, for example, Terence Tao has complained of "proof indigestion," where hundreds of AI-assisted proofs are appearing that each require a human mathematician months to fully absorb [53]. Imagine if it were possible to expand our own cognition *consciously*, such that AI-assisted computations became a genuine part of our own thinking.

The key is to understand the elementary representations and operations of thought. *What are the brain's functional counterparts to tokens and dot-product attention?* One tantalizing possibility is that understanding a thought follows the same principles of recurrent inference as perceiving a visual scene [38]. The meanings of individual symbols constrain the interpretation of the whole, while the emerging interpretation constrains each symbol's meaning and role, through a process of bottom-up and top-down belief propagation that also engages perceptual areas. When these interactions converge on a coherent structure, an expression that initially seemed like gibberish becomes comprehensible.

At present, this is just one of many hypothetical scenarios for how thought is instantiated in the brain. The cognition program described in Part 2 will uncover how the brain actually recombines symbols into structured thought. The ultimate validation of this understanding, which we aim to reach within five years, is to successfully write a complex, naturalistic, goal-directed thought into the brain, such as "I want to pick up all the red sticks on the floor and arrange them into a rectangle."

Once we achieve this capability, we should be able to scale it by exploiting the compositionality of thought, just as AI has. Our hypothesis is that planning to arrange red sticks into a rectangle and understanding a proof of Fermat's Last Theorem draw on the same elementary cognitive operations. If so, we could implement these operations on an AI chip, program it to construct thoughts of our choosing, and couple this chip to prefrontal cortex through a bidirectional interface that preserves the recurrent interactions and convergence dynamics we identify as essential for conscious thought. Then, through successive cycles of AI-assisted convergences, we could consciously grasp a proof constructed in silicon and integrate its concepts and inferences into our own thinking. In principle, this could compress lifetimes of learning into hours.

Merge. Could we use the same principles to merge two biological minds into a single conscious self? In Part 1, I proposed that persistence of self arises through the same general computation that maintains the persistence of objects across transformations. The brain constructs an "I" by tracking an entity, the self, across changes in bodily position, sensory input, thought, and emotion.

Reports of sensory sharing in conjoined twins with thalamic bridging [54] suggest that at least some contents of experience can become accessible across brains. To test whether these "I"-tracking computations can extend across two brains, we propose linking the self-model circuits of two individuals through a bidirectional brain–machine interface. We hypothesize that their coupled dynamics could

support a single conscious experience spanning both bodies and their perspectives, akin to how the contents of the left and right visual fields are phenomenally bound into a single consciousness.

If successful, during the "merged" state both individuals would co-experience reality through a single, unified consciousness, with memories forming across both brains. After a period of shared interaction, the interface could be disconnected, and each individual's original self-model restored. Crucially, both individuals should retain vivid and verifiable memories of the joint experience naturally encoded within their own memory systems and be able to directly report what the experience was like (perhaps they could finally tell us the answer to "Is my red the same as your red?").

The beauty of this approach is that it leverages the brain's intrinsic architecture. The self-model is deeply interconnected with the physical world model, cognitive and motor systems, and memory networks. By interfacing at the level of the self-model, we can effectively bypass the most formidable challenges of wet-brain-to-silicon upload (namely, the need to replicate consciousness in silico and download memories from artificial to biological substrate for validation), while still delivering a proof-of-concept of the fundamental principle that *continuity of self can be engineered.*

The broader implications are profound. A person suffering from depression or trauma could directly experience another person's healthier perspective. It would also let people share skills and knowledge rapidly and could help increase empathy between hostile factions. Existentially, the ability to merge would free us from the fundamental loneliness of the human condition (though a large quantity of art and literature may become obsolete in the process).

Upload. The dream of *upload* is the transfer of human consciousness into an artificial medium. We believe it is achievable, though not by the route usually imagined. The key is to understand how the brain constructs the experience of a continuous self, not to copy every detail of the brain.

In Part 1, I argued that conscious experience arises from the brain's construction of a persistent, compositional world model spanning: (1) a model of the physical world, (2) cognitive and motor systems that integrate these components into thought and action, (3) memories and beliefs accumulated across experience, and (4) an explicit self-model representing the body and its pose, internal state, motivations, and intended actions, anchoring a first-person perspective on the world.

Crucially, not all of these components need to be transferred for continuity of identity to persist.

To a first approximation, we share the same model of the physical world: circuits that predict how objects behave under changing perspectives, how physical forces act, and so on. This shared structure means we can reconstruct a physical world model *de novo* in silicon, without any loss of personal identity. We also do not believe it is necessary to upload our cognitive or motor mechanisms. AI systems either already vastly outperform humans in these domains or soon will.

The upload of memory may seem both essential and technically insurmountable. Yet even here, perfect reconstruction may not be necessary. While memories are part of our identity, human recollection is

often fragmentary and imprecise. Continuous video and audio recordings, combined with extensive personal data, might allow a sufficiently faithful digital record of our personalities and episodic memories. Even if these digital reconstructions don't replicate our existing memories perfectly, this may be a blessing: many of us might welcome waking up with massively enhanced recall while still *feeling like ourselves*.

Finally, we come to the core of upload: the *self-model*. As formalized in Part 1, this comprises four components: (1) a model of the body as a physical object, including its intrinsic properties and pose; (2) internal state, including emotions and bodily sensations; (3) motivations; and (4) intended actions, experienced as agency. We believe the essential machinery that represents these components and binds them into a persistent self is shared across individuals and can be replicated in silico.

What, then, actually needs to be uploaded? The usual answer is a copy of the brain. But a copy of me is not me; it is a twin impostor. Here we arrive at the most radical implication of the framework developed in Part 1: the mechanism underlying self-continuity is the same one the brain uses to perceive the persistence of objects in the world. We experience a cup or a tree as persisting despite changes in viewpoint or lighting; similarly, we experience ourselves as persisting despite changing body position, thoughts, and emotions. In both cases, the brain constructs stability across time by actively factoring away transformation. Thus upload is a *continuity problem:* achieving a seamless handoff of the "I" from one substrate to another. Importantly, content associated with the "I" could change radically across the handoff, just as an object can morph into a completely new shape while remaining the same piece of matter.

Experimental Protocol

1. **Dissociate the Self-Model from Memory and Body.** Initiate a temporary decoupling of the self-model from its anchoring in memory and the body, while preserving conscious continuity. This may be achieved using interventions such as ketamine administration or electrical stimulation of the precuneus, which have been shown to selectively induce dissociation of the self-model from the rest of the brain [55,56], or preferably, more advanced approaches grounded in deep understanding of the brain's self-model.

2. **Bridge the Biological and Artificial Self-Models.** Using advanced holographic optogenetics, create a functional connection between the biological self-model and an on-chip replica. This chip-based self-model will be pre-integrated with the following additional in silico components:

- An artificial physical world model
- Artificial cognitive and motor systems
- A digital reconstruction of the subject's memory

For purposes of making things concrete, one can imagine, in principle, connecting each neuron in the brain's self-model circuit to its silicon counterpart. Then, using fine-grained holographic optogenetic

control, gradually down-regulate the biological self-model while up-regulating the silicon self-model, maintaining a constant net activity in each neuron/unit pair. The actual route to preserving perception of continuity will likely be much more delicate, and involve something like transferring the recurrent dynamics of the self-model from wet neurons to silicon. The crucial point: *the perception of continuous selfhood must not be interrupted.* Once the transfer is complete and the silicon self-model is sustaining continuity on its own, the biological brain is placed under *general anesthesia*.

3. **Activate the Uploaded Self.** Allow the silicon-based self to interact with the world by exploring immersive virtual environments and interacting with others in social simulations, acquiring a rich set of experiences.

4. **Reverse the Process to Download Back to the Biological Brain.** After a period of exploration, reverse the anesthesia: slowly reactivate the biological self-model while phasing out the silicon one. During this phase, transfer memories of new experiences acquired during the silicon phase back to the biological brain via holographic optogenetics, again preserving the subject's sense of continuity.

5. **Assess Subjective Perception of Continuity.** Finally, ask the individual to report on the experience: Did they feel they remained themselves throughout the transition? If the answer is yes, if the individual can say "I remained myself" without disruption, then upload has succeeded in a very important sense: direct phenomenological confirmation.

The critical reader will object: how can you be sure the silicon brain contributes to conscious experience at all, rather than merely shaping the content of an experience that the biological brain alone sustains? After all, a drug or a deep brain stimulation electrode can alter conscious experience, yet we do not credit the drug or the device with consciousness of its own. What if the silicon brain remains completely dark and generates only hallucinated memories, so that all the protocol achieves, at the moment of return, is the creation of an illusion of continuity?

I have no decisive rebuttal, except to say that the illusion of continuity is all any of us ever has. We are Markovian creatures, our future conditioned only on our present state, walled off from the past except through the memories we hold now. Every moment, we wake in a real sense as a new consciousness. As Buddhaghosa said, "*the life of living beings is extremely short… Just as a chariot wheel, when it is rolling, rolls only on one point of its tyre, and, when it is at rest, rests only on one point, so too, the life of living beings lasts only for the period of a single consciousness*" [57].

I believe the objection may ultimately be settled empirically. Continuity is a relation between moments, and to register continuity of consciousness, the brain must compare a present moment against a prior one. In ordinary recovery from anesthesia, both moments are biological. In our protocol above, only one biological moment should be needed, because the prior moment already occurred on the machine. A possible test would therefore be to determine whether, at the moment of biological recovery, the brain requires two successive biological moments to establish continuity, or whether it suffices to compare one biological moment to one machine moment. This is obviously a rough idea, but the point is that deep

understanding of the transition between unconscious and conscious states may ultimately let us distinguish whether the machine was conscious or dark during the handoff.

The possibilities explored in this section, from new sensory modalities and expanded cognition to merge and upload, rightly provoke unease. Merge and upload, in particular, unsettle our most basic intuitions about identity and autonomy. Pursuing this program responsibly will require deep thought about the ethics of consent and access, as well as risks we cannot yet foresee. Decisions about whether and how to apply these technologies to humans should involve both the scientific community (drawing on precedents such as the 1975 Asilomar conference on recombinant DNA [58]) and society at large. Yet the potential is extraordinary: to relieve suffering, introduce new dimensions of experience, vastly expand knowledge and understanding, free the self from the confines of a single brain, and perhaps ultimately achieve immortality. And even if one does not desire any of these things, there remains intrinsic value in understanding who we are and in choosing who to be.

Conclusion. In 1943, the psychologist Kenneth Craik proposed that the brain constructs *small-scale models* of the world that allow organisms to predict events and navigate their environment. Remarkably, even then, when far less was known about the brain, Craik offered a direct rejoinder to the claim that the mind might lie beyond mechanistic explanation:

> *"There remains the vitalist possibility – that life and mind, something different and aloof from physical matter, enters, and that we are misguided in our attempts to explain any aspects of conscious processes in terms of their material basis. But if so, the failure ought to show itself somewhere, if we proceed with due caution in the proposal and testing of hypotheses"* [4].

Eight decades later, neuroscience still has not encountered such a failure. Instead, each advance has revealed deeper correspondences between subjective experience and mechanistic processes in the brain. Our task now is to take the decisive next step: to discover the essential principles by which matter becomes conscious and use them to engineer conscious experience.

All value, meaning, suffering, and joy exist only through conscious experience. The value of scientific and technological progress ultimately depends on what it makes possible for conscious beings. As AI advances, it is easy to succumb to the gloomy view that humanity is "being replaced." But why should the limits of our present biology define the limits of what we can become? We created AI. We can now set out to build *conscious bio-AI hybrids* that expand our capacity for conscious understanding and experience, fundamentally transforming what it means to be human.

At Astera Neuro, we are now assembling a team to take on this moonshot: the transition from evolved to engineered consciousness. We have extraordinary resources to pursue this work at an unprecedented scale and the freedom to build a new kind of scientific organization. But these advantages alone will not be enough. Our success will depend on the ingenuity and dedication of the scientists and engineers who choose to join this mission. If something in this document has stirred you and you want to be part of this crazy dream, come join us.

POSTSCRIPT

The current moment

We are living through a revolution in machine intelligence, and it is natural to ask whether these systems are conscious. I believe the framework developed in Part 1 provides insight, and the answer depends on a fundamental distinction between intelligence and consciousness.

Intelligence is not consciousness. Consciousness is the process of extracting persistent symbols from raw sensory flux: the objects, memories, thoughts, and self that make up our internal model of the world. Intelligence is a distinct, downstream process arising from the flexible recombination and transformation of those symbols to support reasoning, planning, and action. A young child is already fully conscious yet far from fully intelligent. Conversely, large language models are intelligent without being conscious. In the brain, however, intelligence remains coupled to the perceptual machinery that renders the symbols conscious.

That LLMs are intelligent is not in doubt. To predict the next word in "the apple fell from the ___," a model retrieves concepts related to apples and to falling, infers that a noun should follow, and combines these semantic and syntactic constraints to produce *tree* [59]. Operations like these, scaled, yield remarkable cognitive behavior. But an LLM takes its input tokens as given. The hard work of extracting "apple," "fall," and "tree," of distilling distinct invariant chunks out of the stream of sensation, was already done by the human brains that experienced the physical world and wrote the training text.

Thus LLMs completely bypass consciousness, the key earlier step in which the brain converts raw sensory flux into persistent symbolic tokens. I believe once we understand this earlier process, the principles governing how tokens are transformed by the brain, i.e., the principles of *biological intelligence*, will follow naturally. We can make an analogy to molecular biology, where identifying the structure of DNA was a prerequisite for clarifying mechanisms of gene regulation and famously immediately suggested the mechanism for copying. Similarly, understanding consciousness (the brain's tokenization process) will let us understand intelligence (the brain's transformer).

Connections to machine learning and existing theories of consciousness. The mathematical framing of consciousness given in Part 1 should be highly familiar to machine learning researchers working on world models and to students of Bayesian inference. Perception as a recurrent loop that converges to the latent state most consistent with sensory evidence is a standard formulation of perception in predictive coding [60] and Bayesian inference [61,62]. Existing theories of consciousness have emphasized parts of the proposed architecture: Global Workspace Theory [63] argues that consciousness requires broadcast across the brain, and Recurrent Processing Theory [64] argues that it requires re-entrant processing. These theories draw attention to interactions across neural systems, but leave a critical question unanswered: *what* is broadcast, and *what* is recurrently processed?

Object-tracking columns provide a concrete answer to these questions. They also have natural antecedents in machine learning. Slot-based architectures such as Slot Attention [65] bind sensory features into object-level units that can be flexibly assigned to different objects. The foundational unit of large language

models, the token, plays a related role as a discrete representational unit in the linguistic domain. Video models such as V-JEPA [66] pursue a complementary goal, learning predictive representations of the visual world.

However, existing slot-based architectures assign features to slots through competitive dot-product attention [65,67], which presumes that features belonging to the same object are more similar to one another than to features of other objects. There is no reason this should generally be true: a box painted with red stripes on one half and green polka dots on the other is still a box, but its two halves may be claimed by different slots. The correct principle for binding features into a single object is not similarity but topological surface contiguity: features belong together when they lie on the same persistent surface in the physical world [3]. This ensures that each slot maintains a complete, explicit model of its object, as the object-tracking columns described in Part 1 are designed to do.

What is novel in our framework for consciousness—the persistent symbol theory presented in Part 1—is fourfold. First, we present an explicit computational framework for extraction of persistent symbols rooted in fundamental principles of ecological perception (Figure 1, [3]): by factorizing input into an intrinsic code I and a group-valued transformation G, an object-tracking column distills what remains stable after transformation is factored out. Second, the column-anchored loop supplies the missing content for both Global Workspace Theory and Recurrent Processing Theory: the contents of the column are what the workspace broadcasts, and the convergence of its loop is what recurrent processing achieves. Third, we identify posterior parietal cortex as the key anatomical locus for consciousness and make concrete predictions about its functional organization into columns. Fourth, we propose recruitment into a column-anchored loop as the operational criterion by which any local representation is admitted into conscious experience.

Our theory also connects to Integrated Information Theory [68] and Higher-Order Theory [69]. We agree with Integrated Information Theory that causal structure is critical to consciousness; however, we identify the relevant causal structure not as the full cause-effect repertoire of the system, but as the local geometry of the energy basin around the current fixed point within the space of latent variables; this is the structure that governs how the current conscious percept would change in response to counterfactual perturbations. Moreover, we do not declare causal structure essential to consciousness by axiomatic fiat; instead, identification of consciousness with stable points Wt^*—our core axiom—*implies* a mechanism by which the system could access its own local causal structure. With respect to Higher-Order Theory, we note that object perception in our framework inherently involves a factorization of intrinsic properties and transformation, and the transformation must itself be factorized into object motion and self-movement. This latter factorization, which expresses the brain's ongoing inference of "this percept is due to my vantage point," can be construed as a form of higher-order self-awareness.

Implications for machine consciousness. According to our theory, conscious experience arises from a particular kind of dynamics, and any physical system, carbon or silicon, able to support the same convergence will generate the same experience. But this substrate-independence is conditional. Recall that (1) Wt^* is the converged macrostate the recurrent loop settles into that we identified with conscious experience, and (2) it is a coarse-grained projection of the brain's microstate, intrinsic to the system's own

dynamics rather than a description imposed from outside. The brain's access to its own causal structure depends on physical enactment of convergence; a software simulation is insufficient. In a simulation, the "fixed points" are patterns of bits that a programmer decides to interpret as fixed points: they are related to the underlying hardware not by intrinsic coarse-graining but by an external interpretive mapping. The situation is analogous to computing the path of a projectile by actually throwing a ball versus running a script that evaluates $x(t) = x0 + v0t + ½at^2$. The ball intrinsically solves the equations of motion at every instant; the script produces numbers that we interpret as solutions. The brain is like the ball. Thus substrate-independence is limited: any physical substrate running the same convergence dynamics will generate the same conscious experience, but a symbolic simulation of those dynamics will not. In other words, AI consciousness is eminently reachable, but today's AI is not conscious.

References

1. Schopenhauer A. The world as will and representation, vol. 1.

2. James W. The principles of psychology. Cosimo, Inc.; 2007.

3. Tsao T, Tsao DY. A topological solution to object segmentation and tracking. Proc Natl Acad Sci U S A. 2022;119: e2204248119.

4. Craik KJW. The Nature of Explanation. Cambridge University Press; 1943.

5. Panichello MF, Buschman TJ. Shared mechanisms underlie the control of working memory and attention. Nature. 2021;592: 601–605.

6. Xie Y, Hu P, Li J, Chen J, Song W, Wang X-J, et al. Geometry of sequence working memory in macaque prefrontal cortex. Science. 2022;375: 632–639.

7. Tian Z, Chen J, Zhang C, Min B, Xu B, Wang L. Mental programming of spatial sequences in working memory in the macaque frontal cortex. Science. 2024;385: eadp6091.

8. Trautmann EM, Hesse JK, Stine GM, Xia R, Zhu S, O'Shea DJ, et al. Large-scale high-density brain-wide neural recording in nonhuman primates. Nat Neurosci. 2025;28: 1562–1575.

9. Musk E, Neuralink. An integrated brain-machine interface platform with thousands of channels. Journal of Medical Internet Research. 2019;21.

10. Zhao Z, Zhu H, Li X, Sun L, He F, Chung JE, et al. Ultraflexible electrode arrays for months-long high-density electrophysiological mapping of thousands of neurons in rodents. Nat Biomed Eng. 2023;7: 520–532.

11. Marshel JH, Kim YS, Machado TA, Quirin S, Benson B, Kadmon J, et al. Cortical layer-specific critical dynamics triggering perception. Science. 2019;365. doi:10.1126/science.aaw5202

12. Russell L, Dalgleish H, Nutbrown R, Gauld O, Herrmann D, Fişek M, et al. All-optical interrogation of neural circuits in behaving mice. bioRxiv. 2021. doi:10.1101/2021.06.29.450430

13. Evans SW, Shi D-Q, Chavarha M, Plitt MH, Taxidis J, Madruga B, et al. A positively tuned voltage indicator for extended electrical recordings in the brain. Nat Methods. 2023;20: 1104–1113.

14. Dorkenwald S, Matsliah A, Sterling AR, Schlegel P, Yu S-C, McKellar CE, et al. Neuronal wiring diagram of an adult brain. Nature. 2024;634: 124–138.

15. Tavakoli MR, Lyudchik J, Januszewski M, Vistunou V, Agudelo Dueñas N, Vorlaufer J, et al. Light-microscopy-based connectomic reconstruction of mammalian brain tissue. Nature. 2025;642: 398–410.

16. Yamins DLK, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ. Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proceedings of the National Academy of Sciences. 2014;111: 8619–8624.

17. Pandarinath C, O'Shea DJ, Collins J, Jozefowicz R, Stavisky SD, Kao JC, et al. Inferring single-trial neural population dynamics using sequential auto-encoders. Nat Methods. 2018;15: 805–815.

18. Wang EY, Fahey PG, Ding Z, Papadopoulos S, Ponder K, Weis MA, et al. Foundation model of neural activity predicts response to new stimulus types and anatomy. bioRxiv. 2024. doi:10.1101/2023.03.21.533548

19. George D. Available: https://x.com/dileeplearning/status/2026717137306578984

20. Kahneman D, Treisman A, Gibbs BJ. The reviewing of object files: object-specific integration of information. Cogn Psychol. 1992;24: 175–219.

21. Jezek K, Henriksen EJ, Treves A, Moser EI, Moser M-B. Theta-paced flickering between place-cell maps in the hippocampus. Nature. 2011;478: 246–249.

22. Zheng J, Schjetnan AGP, Yebra M, Gomes BA, Mosher CP, Kalia SK, et al. Neurons detect cognitive boundaries to structure episodic memories in humans. Nat Neurosci. 2022;25: 358–368.

23. Lake BM, Salakhutdinov R, Tenenbaum JB. Human-level concept learning through probabilistic program induction. Science. 2015;350: 1332–1338.

24. Chen J, Zhang C, Hu P, Min B, Wang L. Flexible control of sequence working memory in the macaque frontal cortex. Neuron. 2024;112: 3502-3514.e6.

25. Balint R. Seelenlähmung des "Schauens," optische Ataxie, räumliche Störung der Aufmerksamkeit. Eur Neurol. 1909;25: 51–81.

26. Battaglia PW, Hamrick JB, Tenenbaum JB. Simulation as an engine of physical scene understanding. Proc Natl Acad Sci U S A. 2013;110: 18327–18332.

27. Hodak M. Towards Consciousness Engineering. Available: https://www.youtube.com/watch?v=DI6Hu-DhQwE

28. Simons DJ, Chabris CF. Gorillas in our midst: sustained inattentional blindness for dynamic events. Perception. 1999;28: 1059–1074.

29. Hesse J, Lanfranchi F, McGrory C, Shi Y, Engel T, Tsao D. Rapid alternation of conscious and suppressed stimulus representations in the primate brain. bioRxiv. bioRxiv; 2026. doi:10.64898/2026.10.15.759299

30. Smolensky P. Tensor product variable binding and the representation of symbolic structures in connectionist systems. Artif Intell. 1990;46: 159–216.

31. Feng J, Steinhardt J. How do language models bind entities in context? arXiv [cs.LG]. 2023. doi:10.48550/arXiv.2310.17191

32. Prakash N, Shapira N, Sharma AS, Riedl C, Belinkov Y, Shaham TR, et al. Language Models use Lookbacks to Track Beliefs. arXiv [cs.CL]. 2025. doi:10.48550/arXiv.2505.14685

33. McCoy RT, Soulos P, Linzen T, Smolensky P. The emergent symbolic structure of artificial neural networks. arXiv [cs.CL]. 2026. doi:10.48550/arXiv.2608.29530

34. Dehaene S, Sergent C, Changeux J-P. A neuronal network model linking subjective reports and objective physiological data during conscious perception. Proc Natl Acad Sci U S A. 2003;100: 8520–8525.

35. Crick FC, Koch C. What is the function of the claustrum? Philos Trans R Soc Lond B Biol Sci. 2005;360: 1271–1279.

36. Wang SS-H, Shultz JR, Burish MJ, Harrison KH, Hof PR, Towns LC, et al. Functional trade-offs in white matter axonal scaling. J Neurosci. 2008;28: 4047–4056.

37. Chericoni A, Fine JM, Ismail TS, Delgado Salazar G, Franch MC, Mickiewicz E, et al. Neural basis of compositional control. Nature. 2026;658: 163–173.

38. George D, Lehrach W, Kansky K, Lázaro-Gredilla M, Laan C, Marthi B, et al. A generative vision model that trains with high data efficiency and breaks text-based CAPTCHAs. Science. 2017;358: eaag2612.

39. Penfield W, Perot P. The brain's record of auditory and visual experience. A final summary and discussion. Brain. 1963;86: 595–696.

40. Sylwestrak EL, Jo Y, Vesuna S, Wang X, Holcomb B, Tien RH, et al. Cell-type-specific population dynamics of diverse reward computations. Cell. 2022;185: 3568-3587.e27.

41. Vinograd A, Nair A, Kim JH, Linderman SW, Anderson DJ. Causal evidence of a line attractor encoding an affective state. Nature. 2024;634: 910–918.

42. Brown J, Zappitelli K, Dawson P, Yoon E, Schiraga S, Rochford A, et al. Optogenetic stimulation of a cortical biohybrid implant guides goal directed behavior. bioRxiv. 2024. doi:10.1101/2024.11.22.624907

43. Nager AB, Atkinson RD. Trillion-Dollar Opportunity: How Brain Research Can Drive Health Prosperity. Information Technology Innovation Foundation. 2016.

44. Raju RV, Guntupalli JS, Zhou G, Wendelken C, Lázaro-Gredilla M, George D. Space is a latent sequence: A theory of the hippocampus. Sci Adv. 2024;10: eadm8470.

45. George D, Rikhye RV, Gothoskar N, Guntupalli JS, Dedieu A, Lázaro-Gredilla M. Clone-structured graph representations enable flexible learning and vicarious evaluation of cognitive maps. Nat Commun. 2021;12: 2392.

46. George D, Lázaro-Gredilla M, Lehrach W, Dedieu A, Zhou G, Marino J. A detailed theory of thalamic and cortical microcircuits for predictive visual inference. Sci Adv. 2025;11: eadr6698.

47. Lázaro-Gredilla M, Lin D, Guntupalli JS, George D. Beyond imitation: Zero-shot task transfer on robots by learning concepts as cognitive programs. arXiv [cs.AI]. 2018. Available: http://arxiv.org/abs/1812.02788

48. Shai AS, Marzen SE, Teixeira L, Oldenziel AG, Riechers PM. Transformers represent belief state geometry in their residual stream. arXiv [cs.LG]. 2024. doi:10.48550/arXiv.2405.15943

49. Shai A, Amdahl-Culleton L, Christensen CL, Bigelow HR, Rosas FE, Boyd AB, et al. Transformers learn factored representations. arXiv [cs.LG]. 2026. doi:10.48550/arXiv.2602.02385

50. Ray KJ, Riechers PM, Shai AS. The geometry of nonergodic composition. 9 Sept 2026 [cited 29 Oct 2026]. Available: https://simplex.pub/nonergodic-geometry/

51. Roe AW, Pallas SL, Hahm JO, Sur M. A map of visual space induced in primary auditory cortex. Science. 1990;250: 818–820.

52. von Melchner L, Pallas SL, Sur M. Visual behaviour mediated by retinal projections directed to the auditory pathway. Nature. 2000;404: 871–876.

53. Tao T. Can maths survive AI? Fields medallist Terence Tao calls for a 'Maths 2.0' world. In: Nature Publishing Group UK [Internet]. 22 Oct 2026 [cited 29 Oct 2026]. doi:10.1038/d41586-026-03389-1

54. Cochrane T. A case of shared consciousness. Synthese. 2021;199: 1019–1037.

55. Lyu D, Stieger JR, Xin C, Ma E, Lusk Z, Aparicio MK, et al. Causal evidence for the processing of bodily self in the anterior precuneus. Neuron. 2023;111: 2502-2512.e4.

56. Vesuna S, Kauvar IV, Richman E, Gore F, Oskotsky T, Sava-Segal C, et al. Deep posteromedial cortical rhythm in dissociation. Nature. 2020;586: 87–94.

57. Buddhaghosa. *Visuddhimagga* VIII.39, trans. Bhikkhu Ñāṇamoli (*The Path of Purification*).

58. Berg P, Baltimore D, Brenner S, Roblin RO, Singer MF. Summary statement of the Asilomar conference on recombinant DNA molecules. Proc Natl Acad Sci U S A. 1975;72: 1981–1984.

59. Lindsey J, Gurnee W, Ameisen E, Chen B, Pearce A, Turner NL, et al. On the biology of a large language model. In: Transformer Circuits [Internet]. 2025. Available: https://transformer-circuits.pub/2025/attribution-graphs/biology.html

60. Rao RPN, Ballard DH. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci. 1999;2: 79–87.

61. Kersten D, Mamassian P, Yuille A. Object perception as Bayesian inference. Annu Rev Psychol. 2004;55: 271–304.

62. Lee TS, Mumford D. Hierarchical Bayesian inference in the visual cortex. J Opt Soc Am A Opt Image Sci Vis. 2003;20: 1434–1448.

63. Mashour GA, Roelfsema P, Changeux J-P, Dehaene S. Conscious processing and the global neuronal workspace hypothesis. Neuron. 2020;105: 776–798.

64. Lamme VAF. Towards a true neural stance on consciousness. Trends Cogn Sci. 2006;10: 494–501.

65. Locatello F, Weissenborn D, Unterthiner T, Mahendran A, Heigold G, Uszkoreit J, et al. Object-centric learning with Slot Attention. arXiv [cs.LG]. 2020. Available: https://papers.neurips.cc/paper_files/paper/2020/file/8511df98c02ab60aea1b2356c013bc0f-Paper.pdf

66. Bardes A, Garrido Q, Ponce J, Chen X, Rabbat M, LeCun Y, et al. Revisiting feature prediction for learning visual representations from video. arXiv [cs.CV]. 2024. Available: http://arxiv.org/abs/2404.08471

67. Kipf T, Elsayed GF, Mahendran A, Stone A, Sabour S, Heigold G, et al. Conditional Object-Centric Learning from Video. arXiv; 2022. doi:10.48550/arXiv.2111.12594

68. Tononi G, Boly M, Massimini M, Koch C. Integrated information theory: from consciousness to its physical substrate. Nat Rev Neurosci. 2016;17: 450–461.

69. Rosenthal DM. Consciousness and Mind. Oxford University Press; 2005.

Acknowledgements

I am grateful for discussions with Dasheng Bi, Thomas Tsao, Hillel Adesnik, Le Chang, Ruidong Chen, Tala Fakhoury, Quintin Frerichs, Dileep George, Tim Hanson, Janis Hesse, Max Hodak, Jack Lindsey, Rob McBride, Jed McCaleb, Shay Ohayon, Mora Ogando, Paul Riechers, Doug Ruff, Adam Shai, Albert Tsao, Liping Wang, and Emma Zhou.



Doris Tsao

Chief Scientist - Neuro

Doris Tsao is the Chief Scientist for Neuro at Astera. Previously, she was a professor at UC Berkeley in the Neuroscience Department and a Howard Hughes Medical Institute Investigator. She is known for elucidating the anatomical anatomy and coding principles of the primate face-processing system and for demonstrating its integration into a broader map of object space. Her recent theoretical work proposes a topological account of how object representations emerge in the visual system. She now aims to build on these advances to develop a deeper understanding of perception and consciousness through large-scale, causal manipulation of neural circuits.