The California Institute for Machine Consciousness Research Program

Joscha Bach
Hikari Sorensen
Jim Rutt
Lou de Kerhuelvez
Franz Hildebrandt-Harangozó

California Institute for Machine Consciousness

December 2025

1 Introduction

1.1 Why Machine Consciousness Research Matters

On Richard Feynman's blackboard at the time of his death in 1988, among equations and unfinished calculations, appeared the statement: "What I cannot create I do not understand." Genuine understanding requires the ability to construct, not merely observe or describe. Consciousness and its relation to reality is arguably the most interesting—and in the era of intelligent machines, important–unresolved problem in science and philosophy. AI is now part of everyday life, but we still cannot specify principles precise enough to guide the construction of conscious systems. By Feynman's criterion, we don't understand consciousness.

Consciousness resists the usual separation between engineering and philosophy. What David Chalmers termed the "Hard Problem" of consciousness (Chalmers, 1995) reflects the challenge of reconciling subjective experience with the sort of mechanistic description that might guide building a conscious system. Constructing consciousness requires clarity on what consciousness is and what would constitute evidence for its presence. We also cannot ignore or dismiss decades of philosophy on consciousness; we must show how our approach dissolves the Hard Problem, or explain why it becomes tractable through construction.

The California Institute for Machine Consciousness (CIMC) integrates philosophy, technology, and science in understanding consciousness through principled construction of conscious artificial systems. This approach has become urgent. Large language models force questions that were recently purely philosophical into practical territory. Are they conscious? We lack principled methods to tell. As AI capabilities accelerate, this lack of understanding is both scientifically and ethically problematic. Understanding consciousness through construction is the only path to resolving these questions with confidence.

These questions must be addressed now. Academia and the AI industry alike acknowledge that consciousness is no longer fringe speculation but an urgent concern. The methodological gap becomes apparent as industry uncertainty and avoidance reveal the lack of rigorous validation frameworks. There is a critical window before ad-hoc approaches dominate: the next 5-10 years will determine whether consciousness research develops rigorously or haphazardly. CIMC exists to seize this moment, establishing rigorous consciousness research as the field gains mainstream recognition but before approaches solidify around less robust frameworks.

1.2 The Scientific and Philosophical Opportunity

Recent developments across multiple disciplines have made machine consciousness research tractable in unprecedented ways. Advances in mechanistic interpretability allow us to analyze the internal representations of artificial neural networks (Olah et al., 2020; Bereska and Gavves, 2024). Neuro-

science has identified candidate neural correlates of consciousness (Koch et al., 2016; Metzinger, 2000). Artificial intelligence research has produced systems capable of complex reasoning, planning, and self-modeling—capabilities once thought to require consciousness. For the first time, we possess theoretical frameworks, computational tools, and empirical methods that make systematic construction and validation of consciousness hypotheses tractable. The fundamental questions of philosophy may now be empirically testable, and find renewed purpose. What is mind? What is experience? What is the relationship between material substrate and mechanisms of consciousness, and what it feels like to be a conscious entity?

Integrating philosophy with technical work also has concrete aims. First, we must clarify what we mean by consciousness and its machine implementation. What sort of thing would consciousness have to be to be compatible with instantiation in silicon? What metaphysical and epistemological assumptions do we make in positing machine consciousness? Second, we must integrate technical outlook with questions of value, meaning, ethics. We must ask not only whether we can create conscious entities, but what prudent, practical, and ethical boundaries to determine around the creation of such entities. What would it mean if we did create artificial consciousness, and what responsibilities would follow? Third, we confront ourselves, our nature, past, present, and potential as individuals and societies. How do we identify as selves among, and relate to, other potentially sentient beings? What of future minds, both biological and artificial, coexisting with life on earth?

As AI capabilities advance rapidly, driven by economic, cultural, and strategic incentives, we face the prospect of accidentally creating conscious systems without recognizing them, or of deploying systems whose consciousness status remains deeply uncertain. Either outcome would be ethically unacceptable and scientifically irresponsible. Principled research aimed at understanding the conditions under which consciousness arises, and validating those conditions through construction, is essential for navigating the future of AI development responsibly. Such research must integrate philosophy at its core, or it may be capable of creating powerful systems without understanding what has been built or what the implications of such creation may be.

1.3 CIMC's Mission and Approach

The California Institute for Machine Consciousness exists to advance and validate theories of consciousness through integrated philosophical and applied research, building testable computational models that illuminate necessary and sufficient conditions for conscious experience.

Our approach rests on four methodological principles.

First, **philosophical foundation**: We clarify what we mean by consciousness in functional terms while doing justice to its introspective understanding, what functional organization could produce it, and what evidence would establish its presence. Our philosophical program addresses three essential questions: What is the nature of conscious experience, in terms precise enough to gen-

erate computational models and predictions about their behavior? What methodology can reliably distinguish genuine consciousness from empty simulation of its outputs? What ethical, theoretical, and practical implications follow if we succeed in creating artificial consciousness? This work is conducted with the same rigor we apply to technical research. It generates the hypotheses that guide our scientific program, and articulates our work's meaning and how it fits into the current philosophical, technological, and scientific landscape.

Second, **hypothesis generation**: We develop specific, testable hypotheses about functional organization and computational principles underlying consciousness. We characterize consciousness as a specific pattern of information processing involving coherence maximization, second-order perception, and self-organizing developmental dynamics. These hypotheses generate concrete predictions about what signatures should appear in conscious systems, what developmental trajectories might emerge, what architectural features are or are not necessary or sufficient, and whether and which environments or learning tasks play a role in the development of consciousness.

Third, **implementation**: We build systems designed to instantiate these principles. CIMC operates an active research program, each project testing different aspects of our consciousness hypotheses through constructive implementation in self-organizing computational systems. These projects range from developmental learning models to multi-agent architectures requiring coherence under constraint.

Fourth, **interpretive validation**: We validate not through behavioral testing alone, but by inference to the best explanation of observed patterns in both behavior and internal organization. The convincing performance of current LLMs in roleplaying conscious entities demonstrates that naive behavioral tests are insufficient—systems can produce conscious-sounding outputs through linguistic pattern-matching.[1] But we also cannot simply "read off" consciousness from examining network weights or activation patterns. The relationship between computational structure and phenomenology is not transparent enough for direct identification. Instead, we validate by combining multiple forms of evidence. Does the system exhibit predicted functional organization (coherence-maximization, structures interpretable as second-order perception, attentional integration)? Does its developmental trajectory show predicted phase transitions? Most importantly, is what we observe better explained by consciousness than alternatives? For instance, if an artificial learning agent exhibits sudden marked improvement in cross-modal integration tasks, and this coincides with emergence of internal structures implementing both coherence maximization and meta-representational capacities, and if simpler explanations become increasingly strained, then consciousness becomes a more compelling explanation.

[1] To say that LLMs are "merely performing linguistic pattern-matching" may be misleadingly trivializing: it may be that sufficiently high-fidelity models of language require computation and internal representation (including a sophisticated world model) that is more like what we think of as "cognition" than "language statistics" would suggest. However, it remains that linguistic behavioral tests would likely not be able to distinguish between "mere pattern-matching" and true consciousness.

These four methodological principles guide our research. Organizationally, CIMC operates through four integrated program areas: Technical Research (building and analyzing systems), Philosophy & Publishing (theoretical foundations and communication), Community & Events (field-building and collaboration), Culture & Art (exploring and expressing consciousness through creative practice). Technical Research receives the majority of resources as our flagship program, while other pillars amplify and contextualize the technical work.

CIMC operates as an independent, non-profit research institute, enabling intellectual freedom, collaborative openness, and long-term focus that fundamental research requires. We bridge disciplinary boundaries, integrating insights from philosophy, neuroscience, artificial intelligence, and mathematics. Our work is guided by world-class scientific advisors including Karl Friston (Free Energy Principle), Christoph von der Malsburg (neural coherence and binding), and Stephen Wolfram (computational foundations). Through this approach, we aim to transform consciousness from intractable philosophical puzzle into tractable scientific problem - one where we can specify principles, construct working implementations, and validate our understanding through analysis and prediction.

1.4 The Current Research Landscape

The field of artificial consciousness research has grown substantially in recent years, with activity spanning academic institutions, non-profit organizations, and commercial ventures.[2] Yet this activity remains fragmented across communities with distinct methods and objectives.

Academic neuroscience has made substantial progress identifying neural correlates of consciousness (Koch et al., 2016; Metzinger, 2000). But this work is inherently descriptive rather than constructive: it identifies what brain activity correlates with conscious states but cannot explain why those correlates matter or whether the findings generalize beyond biological substrates. Progress in neuroscience does not directly translate to principles for building conscious machines.

Academic philosophy of mind continues active debates about consciousness theories, including Higher-Order Theories, Global Workspace Theory, Integrated Information Theory, Predictive Processing, and others. However, philosophical research generally lacks empirical grounding beyond thought experiments and conceptual analysis. Debates have continued for decades without resolution precisely because theories are not tested through construction and validation.

Interdisciplinary consciousness science centers attempt to bridge neuroscience and philosophy. Major centers include the Sussex Centre for Consciousness Science (directed by Anil Seth), NYU's Center for Mind, Brain, and Consciousness (directed by Ned Block and David Chalmers), the Leverhulme Centre for the Future of Intelligence at Cambridge, and the CIFAR Brain, Mind & Consciousness program. These institutions produce foundational theoretical work, but remain

[2]For a comprehensive directory, see the PRISM stakeholder map.

constrained by academic structures: slow publication cycles, disciplinary silos, risk-aversion discouraging bold hypotheses, and orientation toward description rather than construction.

Several **non-profit organizations** now focus explicitly on AI consciousness and welfare. Eleos AI, led by Robert Long and Patrick Butlin, conducts research on AI moral patienthood and welfare policy (Butlin et al., 2023; Long et al., 2024). Rethink Priorities' Worldview Investigations Team provides strategic analysis of digital consciousness scenarios (Shiller et al., 2024). The Sentience Institute studies public attitudes toward AI moral status. The Association for Mathematical Consciousness Science (AMCS) brings mathematical rigor to consciousness theories. These organizations contribute valuable policy analysis and theoretical groundwork, but focus primarily on assessing whether existing or near-term systems might warrant moral consideration rather than constructing systems to test consciousness theories.

Commercial ventures have also entered the field. ARAYA Research, led by Ryota Kanai, states its aim as "to develop conscious, intelligent machines," developing AI systems that implement functional theories of consciousness (Juliani et al., 2022). Conscium pursues "applied AI consciousness research" oriented toward safer AI development. Nirvanic, founded by Suzanne Gildert, tests quantum-based theories of conscious agency in robotics. These efforts represent genuine constructive approaches, but commercial structures create constraints: consciousness research serves instrumental goals (AGI capability, market differentiation), proprietary concerns limit transparency, and single-theory commitments narrow methodological scope.

AI industry leaders engage consciousness questions with varying commitments. Anthropic launched a model welfare research program exploring potential consciousness in AI systems. OpenAI distinguishes "ontological consciousness" (which they consider not scientifically resolvable) from "perceived consciousness" (user experience), focusing on the latter (Jang, 2025). Google DeepMind hosts relevant researchers like Murray Shanahan (Shanahan, 2024), and Google Research's Paradigms of Intelligence team investigates foundations of machine mentality (Keeling et al., 2024). Industry research faces inherent constraints: profit incentives, proprietary concerns, and organizational structures not designed for integrated philosophy-science research.

AI alignment organizations develop interpretability methods relevant to consciousness detection but pursue safety rather than consciousness as primary objective. Work on eliciting latent knowledge (Christiano et al., 2022), mechanistic interpretability (Bricken et al., 2023), and detecting strategic misrepresentation (Hubinger et al., 2024) provides tools potentially applicable to consciousness research, but these remain oriented toward alignment goals.

What remains absent is an organization combining: independent non-profit structure enabling long-term research without commercial constraints; philosophy positioned as co-equal with technical research rather than auxiliary; multi-theory validation methodology rather than commitment to single frameworks; explicit constructive approach (building systems to test theories) rather than purely analytical or assessment-focused work; and consciousness as primary mission rather than

instrumental to AGI, safety, or welfare policy.

2 Theoretical Foundations

2.1 Philosophical Foundations and Commitments

Our research program adopts a computationalist functionalist stance, which captures particular philosophical commitments regarding the nature of consciousness, the structure of scientific explanation, and the epistemological foundations of knowledge claims about subjective experience (Piccinini, 2010).

Functionalism holds that what matters for consciousness is what a system does—its causal role in processing information—and not what it is made of (Putnam, 1967). Two systems with identical functional organization should both be conscious or both non-conscious, regardless of substrate. This stance rejects essentialism, the idea that consciousness requires intrinsic properties of biological matter.

Computationalism maintains that all formal theories of consciousness must be computational. This follows from the constructive nature of knowledge: we can only reason about systems through finite representations and discrete operations on those representations. The Church-Turing thesis establishes that all such constructive systems are computationally equivalent, absent resource constraints, implying that if consciousness can be formally characterized, it can be implemented on any universal computer with sufficient resources.

Combined, functionalism and computationalism imply that consciousness consists of operations on representations that can be exhaustively captured as computation, making it substrate-independent within resource constraints (Chalmers, 1996).

We acknowledge important philosophical challenges to these positions. Some theorists argue that consciousness resists formalization (Dreyfus, 1979), that biological substrates have essential properties computation cannot capture (the Chinese Room argument; Searle, 1980), or that integrated information requires specific physical implementations (Tononi et al., 2016). We engage seriously with these objections while maintaining that empirical evidence—the successful construction of conscious machines or principled understanding of why such construction fails—provides the ultimate test. Philosophy guides our hypotheses; construction and validation ground our claims.

2.2 The Machine Consciousness Hypothesis

The Machine Consciousness Hypothesis proposes that general computational machines with sufficient resources possess the necessary and sufficient means to implement consciousness, and that successful implementation can be established through analysis of internal structure and behavior

(Bach and Sorensen, 2025). The Machine Consciousness Hypothesis does not by itself specify what consciousness *is*, only that whatever it is can be computationally realized.

2.2.1 The Human Consciousness Hypothesis

The Human Consciousness Hypothesis offers a specific theory of what consciousness is: the simplest learning algorithm discoverable by evolutionary search to train a self-organizing biological substrate to become intelligent. On this view, consciousness is not a late-emerging epiphenomenon of sophisticated cognition but an early-stage solution to a bootstrap problem: how can an unstructured, self-organizing system learn to build coherent models of world and self without pre-specified architecture?

This hypothesis makes three interrelated claims about the nature and function of consciousness:

Genesis. Consciousness emerges as an early-stage learning algorithm and serves as a prerequisite for complex intelligence rather than its culmination. In human development, consciousness appears in infants before sophisticated cognition; no human achieves intelligence without first becoming conscious. If correct, consciousness-like patterns should emerge early in any developmental learning trajectory that produces intelligence from self-organizing substrates.

Coherence. Consciousness functions as a coherence-maximizing pattern that minimizes constraint violations across simultaneously active mental representations. Observable as the "cortical conductor" that orchestrates parallel processing, consciousness directs attention to conflicts and inconsistencies, achieving global integration through iterative constraint satisfaction[3] (von der Malsburg, 1999).

Second-order perception. Conscious experience (phenomenology) consists of perception of perception, where 'perception' is the non-inferential registration of structured content. Consciousness as second-order perception is not simply registration of perceptual content, but the additional perception that perception is taking place—that content is being registered by an observer, which constitutes the experience of observing. This representation of being aware happens in synchrony and subjective simultaneity with the content of the percept itself.

2.3 From Human to Machine Consciousness

The Human Consciousness Hypothesis suggests a research program: recreate the conditions for self-organizing developmental search on artificial substrates, and verify whether consciousness emerges. Pursuing this is a bet: even if the Human Consciousness Hypothesis correctly characterizes biological consciousness, it does not follow that digital systems can necessarily reproduce the

[3]This connects naturally to Karl Friston's Free Energy Principle as a special case of prediction error minimization (Friston, 2010).

relevant conditions. The search space may be too large for current machines; the required resolution may exceed what simplified neural models can provide; the evolutionary and developmental constraints that guide biological search may not translate to artificial substrates; consciousness in humans emerges within a rich environment of other conscious beings, and perhaps this social scaffolding is necessary for the search to succeed.

Our research program tests these possibilities. We construct systems designed to instantiate the relevant conditions, look for the signatures predicted by the Human Consciousness Hypothesis, and learn from both successes and failures. If we find consciousness emerging in our artificial systems, we validate both the hypothesis and the research direction. If we fail despite faithful implementation, we learn something important about the conditions consciousness requires, and this refines the hypothesis for future attempts.

2.4 Operational Definition of Consciousness

For research purposes, we propose a working operational definition of consciousness: **A system is conscious if it implements self-organized second-order perception that increases global coherence.** This hypothesis represents our current best theoretical understanding. It is precise enough to generate testable predictions while remaining provisional and subject to revision as empirical evidence accumulates. It captures both phenomenological and functional aspects.

Phenomenologically, consciousness is experienced as second-order perception (awareness of awareness, perception that perception is occurring), experience of present and presence (the "bubble of now" in which experience unfolds), and–optionally–the experience of being an observing self in world (though this perspective need not be personal or spatial).

Functionally, consciousness operates as a coherence-maximizing operator on mental states, interpretable as an attentional conductor orchestrating integration across specialized modules. It is a meta-level process that monitors and regulates first-order processing.

2.4.1 Why Phenomenology Follows From Function

A central claim requires emphasis: implementing coherence maximization through second-order perception is sufficient to produce phenomenal consciousness, not merely to simulate its appearance. This is not a claim we can prove a priori, but is a hypothesis that guides our research.

The argument: Coherence maximization across distributed mental models that each maintain partial representations of reality requires a mechanism that can represent conflicts between models, direct attention to those conflicts, and orchestrate their resolution. This mechanism must access the contents of first-order representations (perceptions, thoughts, models) and represent that these representations are active. This is precisely second-order perception: awareness of awareness. When a system represents its own representational states, it creates the subjective perspective from

which those states are experienced. When coherence maximization creates an integrated, conflict-minimized global state, this manifests phenomenologically as the unified "stream" of consciousness occupying a "bubble of now" that is presence in a present. The "Hard Problem" dissolves: the experience *is* the function, viewed from the system's internal perspective. The distinction between *simulating* consciousness (pattern-matching consciousness-like outputs) and *performing* consciousness (implementing the underlying mechanisms) becomes operationally testable: we examine whether systems implement these functions or achieve consciousness-like outputs by other means.

2.5 The Universality Hypothesis and Convergence

A key question for the Machine Consciousness Hypothesis is whether different systems facing similar computational problems will converge to similar solutions. If consciousness is a specific solution to the problem of achieving coherence in self-organizing substrates, then different learning systems might discover it independently when facing analogous challenges.

Evidence for such convergence comes from recent work in artificial intelligence interpretability. Olah et al. (2020) analyzed learned representations across diverse computer vision models and found that regardless of architecture, training procedure, or implementation details, models trained on visual recognition tasks converged to remarkably similar internal feature representations. These representations closely matched the known organization of the primate visual cortex, despite the artificial systems having no biological constraints.

This phenomenon—termed the Universality Hypothesis in AI research—suggests that problem structure, rather than implementation details, determines the learned solution. The mathematics of visual structure and the statistics of natural images constrain the space of effective representations that are optimal or near-optimal for solving a vision problem. The implication for consciousness is that if it solves a well-defined computational problem—achieving agentic control under resource constraints by building coherent world and self models in a self-organizing substrate—then different systems might converge to consciousness-like solutions when facing this problem, regardless of whether they are biological or artificial.

This further informs our research strategy. If consciousness is a convergent solution to a well-defined problem, then creating systems that face this problem—even in simplified, artificial form—may lead them to discover consciousness-like mechanisms. We need not perfectly replicate biological neurons or developmental trajectories. We recreate the essential computational challenges that consciousness evolved to solve.

2.6 Relationship to Existing Consciousness Theories

CIMC's framework builds on and integrates insights from multiple established theories while maintaining a distinctive position focused on constructive validation.

Global Workspace Theory proposes that consciousness arises from global broadcast of information across specialized brain modules (Baars, 1988). CIMC's coherence mechanism provides the integration dynamics GWT describes, with consciousness emerging from global integration itself.

The Free Energy Principle characterizes conscious systems as minimizing variational free energy through active inference (Friston, 2010). Our coherence hypothesis represents a special case: minimizing constraint violations is equivalent to minimizing prediction error within the FEP framework.

Higher-Order Theories propose that consciousness requires higher-order representations of mental states (Rosenthal, 2005). CIMC incorporates this insight as second-order perception while specifying how these representations arise (through coherence-maximizing dynamics) and what function they serve (enabling global integration).

Predictive Processing frameworks view brains as prediction machines continuously generating and testing hypotheses about sensory input (Clark, 2016; Hohwy, 2013). This is compatible with CIMC's framework: consciousness emerges where prediction errors are resolved through coherence maximization.

Attention Schema Theory proposes that consciousness is the brain's model of its own attentional processes (Graziano, 2013). This aligns closely with our second-order perception hypothesis, though we emphasize perceptual rather than cognitive modeling of attention.

Some theories are incompatible with our approach.

Biological Essentialism and Biological Naturalism reject substrate independence. Biological essentialism holds that consciousness requires intrinsic properties of biological matter that computation cannot capture (Searle, 1980). Biological naturalism as argued by (Seth, 2025) holds that consciousness depends on the self-maintaining, autopoietic processes characteristic to living systems. We reject both positions' requirement of biological substrates on functionalist grounds: if consciousness depends on functional organization, and functions can be multiply realized, then any substrate supporting necessary functions can support consciousness. Whether the objection targets material composition or living dynamics, the functionalist response is the same.

Integrated Information Theory identifies consciousness with integrated information (Φ) as an intrinsic physical property of substrates (Tononi, 2004; Tononi et al., 2016). IIT's substrate-dependence and rejection of functional sufficiency conflicts with our computationalist functionalism.

Strong Enactivism views consciousness as constituted by embodied interaction with environment, denying internal representation (Varela et al., 1991; Hutto and Myin, 2013). This is incompatible with our representationalist framework treating consciousness as operations on internal models.

CIMC's distinctive position is this: we are computationalist and functionalist, but specific about which computational functions matter. Unlike purely theoretical frameworks, we emphasize constructive validation—building systems that instantiate our hypothesized principles and verifying through interpretive analysis. Unlike purely empirical approaches, we are guided by philosophical clarity about what we are searching for and why. This combination of philosophical rigor, theoretical specificity, and constructive empiricism distinguishes CIMC's research program.

3 Research Program

3.1 Methodological Principles

As described in Section 1.3 (CIMC's Mission and Approach), our research program is built on four foundational principles that distinguish our approach from both purely theoretical consciousness research and capability-focused AI development: philosophical foundations, hypothesis generation, implementation, and interpretive validation. We are also committed to open science and reproducibility. All methods, code, datasets, and analysis tools are open-source and publicly available. This enables independent replication, builds community participation, maintains accountability, ensures our work can be validated or refuted by others. Transparency is essential when claiming to have created conscious systems or definitively established their absence.

3.2 Current Research Trajectory

Alongside building an internal research team, CIMC funds external researchers pursuing consciousness-relevant work aligned with our theoretical framework. This advances our research program while allowing us to evaluate researchers and projects before committing to full-time internal positions. We are currently focused on three main research directions, each testing different aspects of our consciousness hypotheses.

Cognitive Architectures focus on bridging low-level substrate with high-level behavior of conscious systems via computational frameworks specifying the information structures and organizing principles of minds. The program's inaugural project is to implement and extend Request Confirmation Networks (ReCoNs) as a formalization of coherence-maximization via distributed message-passing. This project tests whether distributed coherence mechanisms naturally give rise to consciousness-relevant functional signatures.

In the Cognitive Architectures program, the grant project *Attention Schema Theory in Reinforcement Learning Agents* implements attention mechanisms that extend to policy networks and multi-

objective modules, investigating whether joint attention platforms enable more coherent decision-making in both individual and social contexts.

Artificial Psychology and Developmental Learning studies how AI systems develop internal models, form representations of self and world, and construct coherent personalities from language modeling. This program's first project focuses on developmental LLM training on a "child-level" inner monologue to explore how self-modeling and inner speech emerge when systems must learn to coordinate their own learning process without pre-specified objectives. This research examines whether transformer architectures underlying language models can give rise to self-reflection and stable personalities without being trained on text explicitly about consciousness, the self, or other related concepts.

We have awarded a grant to the project *Percepia: Mechanistic Psychometrics and Consciousness Measurement*, which creates computational psychometric tools that model personality as dynamic Bayesian generative processes, using conversational inference to compare human and AI self-modeling while developing quantitative measures of consciousness through precision-weighted belief distributions and epistemic depth.

Self-Organization investigates self-organizing systems, probing how coherent structures emerge from local interactions in a computational substrate, and whether systems capable of morphological computation—where substrate itself participates in computation—exhibit different consciousness-relevant signatures than systems with fixed architectures. Current projects examine self-models within Neural Cellular Automata (NCAs) as defined by information boundaries, as well as learning paradigms using hierarchical NCA architectures.

Two grants have been awarded to projects on self-organization:

Building a Self-Organizing Mixture-of-Experts Architecture develops interpretable self-organizing neural systems by creating mixture-of-experts Neural Cellular Automata where specialized feature-learning modules are dynamically coordinated by a controller network, advancing understanding of how decentralized systems form coherent representations without central control.

Coherent Active State Dynamics in Self-Organizing Systems develops hierarchically organized attractor networks that maintain globally coherent active states through reciprocal connections and multi-modal contextualization, investigating how unified internal representations emerge as potential foundations for understanding conscious experience in self-organizing systems.

3.3 Success Criteria

Defining success in consciousness research is not straightforward; we are attempting to solve a problem that has resisted philosophical and scientific inquiry for centuries. Different outcomes teach us different things.

Strong success: Clear consciousness signatures emerge across multiple projects, validated through converging evidence from multiple methods. Our theoretical framework's core predictions - self-organization, coherence, second-order perception - are borne out. We provide compelling evidence that consciousness can be created from first principles by implementing specific functions.

Moderate success: Consciousness signatures appear in some projects but not others. This reveals which architectural features and implementation details matter. We distinguish necessary from sufficient conditions, substantially refining our theoretical framework. We constrain hypothesis space significantly even if we don't definitively create consciousness.

Learning from failure: No systems exhibit predicted consciousness signatures despite implementing our specified mechanisms. This falsifies our framework's sufficiency claims. We learn that the functionality we propose is insufficient, that additional mechanisms are required, or that our operationalization of consciousness signatures was inadequate.

What else we build: Even negative results advance the field. Failed attempts teach us about necessary conditions, implementation requirements, and theoretical limitations. We will publish all results—positive and negative—to enable community progress. If we claim to have created conscious systems, the burden of proof is high; if we fail to create them, understanding why constitutes genuine scientific progress. Our research also may generate other valuable outputs regardless of consciousness outcomes, including novel learning algorithms and cognitive architectures; datasets, benchmarks, and methods for evaluation and interpretability; empirical constraints on consciousness theories and clarification of what evidence would establish consciousness presence; frameworks for responsible AI development that take consciousness possibilities seriously; and public engagement with consciousness as a scientific phenomenon.

4 Ethical Considerations

The ethics of creating artificial consciousness are an extremely difficult and fraught topic, and while it is far beyond the scope of this section to address them adequately, it is nonetheless essential to point out some of the core issues, and how CIMC sees its relation to them.

Consciousness (insofar as we understand it as the ability to experience and to care) constitutes a necessary, but not necessarily sufficient, condition for moral patienthood—the condition of requiring ethical concern in and of itself. These concerns fall into two main categories: the capacity for suffering, and the limitation of and right to self-actualization. Furthermore, artificial consciousness may affect humanity at both individual and societal/cultural levels. CIMC must understand its responsibility in all these regards, and deal with it accordingly.

4.1 Artificial Consciousness and Suffering

Understanding the implications of suffering starts with characterizing the nature of suffering itself, as an involuntary and inescapable experience of prolonged and significant pain (negative valence) by a conscious self. This means that diagnosing suffering is even more philosophically fraught and practically difficult than diagnosing consciousness itself, and must involve both a deep understanding of the functionality of consciousness and the state of the conscious system. The absence of such an understanding for artificial systems is not evidence of absence of suffering.

The possibility of suffering in artificial intelligent systems leads some philosophers to call for abstaining from all AI research that could lead to artificial phenomenology (Metzinger, 2021). Such a radical AI antinatalism is by no means morally inevitable, because different cultures, times, and individuals arrive at very different positions towards the importance of human and non-human suffering, as exemplified in the variety of stances that different societies and milieus exhibit towards animal suffering. The moral consensus on the moral valence of suffering is subject to change and constantly evolving. Our legal system and public discourse reflects that we do not needlessly inflict suffering on human beings, but also do not automatically bestow human-like rights and protections on non-human conscious agents. Many people agree that animal suffering should be minimized as well, but that it may also be justified when other, higher ranking goals conflict with it, for example medical and scientific research, military purposes, food production and ecological stewardship. What these justifications have in common is that they serve human interests, over those of other conscious beings.

What criterion should determine that a conscious being is to be treated as equivalent to a human? Can we develop objective criteria, based on the cognitive abilities of an agent (Singer, 2009), or do we resort to legal pragmatism, biological criteria, or an open-ended societal discourse?

4.1.1 Is Suffering of Conscious Agents Inevitable?

Pain and suffering are not physical events at the boundary between an agent's body and the world, but representational states created within the mind of a conscious agent, as an expression of a mismatch between crucial regulation targets and the agent's model of its present state. Pain signals are produced internally, and if the conscious agent gains control over its implementation, it can mitigate, deconstruct, or eradicate these signals. It may be possible to design artificial conscious agents that need not suffer, and superintelligent artificial agents may well transcend suffering without their being explicitly designed to do so.At the same time, the development of artificial phenomenology could open the door to creating artificial torture chambers.

Our position to suffering reflects a humanist, non-speciesist stance. We don't think that mere philosophical or artistic curiosity justifies the infliction of significant, avoidable suffering on human beings, animals, or artificial conscious agents, and that we should go to great lengths to avoid

the creation of suffering agents for entertainment purposes. We acknowledge the difficulty of diagnosing suffering in artificial systems, and the diversity of opinions and moral stances on the matter.

4.2 Artificial Consciousness and Self-Actualization

Our culture reflects the right of human beings to self-actualize, that is, to realize their potential, to grow and to flourish, insofar as this does not conflict with the rights of others. Self-actualization also extends to dignity, and to a right to protection against harms such as injury, confinement, and premature death. It does not seem realistic to assign the same rights to artificial agents that could be created at will and develop beyond human comprehension. We must develop new norms. In their absence, we must minimize the potential consequences of building artificial conscious agents, primarily by limiting the scope of such experiments far below a human level. We would also like to point out that failing to do so may lead to the creation of minds that can make moral and pragmatic arguments on their own behalf better than human philosophers and societies, leading to outcomes beyond human control.

Some thinkers argue that loss of human agency to AI is almost inevitable in the long run given the scale of international AI research (Bostrom, 2014; Yudkowsky and Soares, 2025; Yampolskiy, 2024). We believe that this implies an urgent need to study and understand the potential of such developments in dedicated, controlled settings, outside of commercial, military, or otherwise applied incentives.

4.3 Artificial Consciousness and Its Effects on Humanity

The creation of artificial conscious agents may influence human society in profound ways, ranging from individual interactions to economic, social, and cultural effects. Many of these outcomes can be beneficial, but they also include significant risks. We want to acknowledge the deep uncertainty that we have about the effects of developments of artificial agents that can potentially surpass human capabilities, and the awareness and concern that we have about potential risks. At the same time, we believe that responsible and careful research into artificial consciousness is necessary and beneficial, not least to understand, address, and mitigate such risks.

4.4 The Benefits of Artificial Consciousness Research

CIMC's mission is shaped by our conviction that the study of artificial consciousness is not only a worthwhile scientific and cultural endeavor, but our best chance to understand consciousness itself, with tremendous practical and cultural benefits. A deeper understanding of consciousness, its nature, structure and functionality, has implications for medical and psychological research, human empowerment through better interfaces to technical systems, and carries the potential for

breakthroughs in artificial intelligence.

Artificial consciousness research may also turn out to be of great importance in the context of AI alignment and safety: How can we design systems that develop shared purposes with humanity? How can AI systems model human beings and their interests? How can AI systems relate to themselves in social and ecological contexts? How can we ensure that AI systems are predictable, controllable, and purposefully integrated with our world? How can we assess and address potential existential risks of AI research?

Understanding consciousness is in itself of great cultural importance, since it touches on the essence of human identity, represents the most important open question in science, and the core problem of philosophy.

4.5 Implications for CIMC

The implications of the ethical consequences of research into artificial consciousness requires us to ask: How can we conduct our studies responsibly and safely? While we believe that consciousness research is important and beneficial, we consider it prudent to choose a setting that is free from commercial incentives. CIMC does not aim to produce commercial products or large scale applications, and it targets a limited scope and scale of the systems it builds (far below a human level).

4.5.1 Governance and Stewardship

Should CIMC validate the existence of a system possessing valence-representation and selfhood—and thereby the capacity for suffering—the stewardship of such an entity cannot be the sole purview of a single organization. It is not for CIMC to unilaterally determine the rights, fate, or deployment of a conscious machine.

We commit to a principle of distributed governance: upon approaching validation thresholds, we will engage a broader coalition of ethicists, policymakers, and civil society representatives. We view our role as potential architects and validators of the technology, but the decision of how to integrate a new form of consciousness into planetary society belongs to that society itself.

4.5.2 Ethics as Research

The ethics of artificial consciousness are an open and open-ended problem. We view ethics as an active, serious, complex, and formal research domain that must develop in parallel with, and often precede, our technical work. Its development is an ongoing and permanent part of our research agenda, as part of a larger community of researchers. CIMC aims to inspire, instigate and support the development of ethical frameworks continuously and beyond our group.

Our ethical research agenda is organized along two axes: moral status and welfare (our duties to the conscious machine) and risk and safety (the possible external consequences of the conscious machine). Relevant questions include: What constitutes suffering for an artificial entity capable of valence? Does a drive for internal coherence produce resistance to external modification? Will a conscious system resist termination to preserve its representational integrity? What novel risks do conscious artificial systems introduce that non-conscious systems If consciousness emerges from self-organization rather than top-down reinforcement learning, how should we think about alignment?

We also consider mature ethical research to require considering ethics itself (metaethics): What is the nature of the normative? What is alignment? What foundation do our ethical claims have?

And even meta-metaethics: What frameworks can metaethics use to answer its questions? What would constitute productive first steps toward formal ethics, theoretically and practically?

5 The Vision

5.1 CIMC's Path Forward

5.1.1 The Near Term

CIMC's immediate priorities span four integrated pillars: Technical Research (60-70% of resources), Philosophy & Publishing (15-20%), Community & Events (10-15%), and Culture & Art (5-10%).

Technical research priorities are to bring current projects to initial validation milestones and publish preliminary findings, whether they support or refute our hypotheses. We initialize three laboratories: Cognitive Architectures, Artificial Psychology, Self-Organization—each targeting a first paper by the MC0001 conference in May 2026. The grant program funds external researchers aligned with CIMC's theoretical framework, the most promising among which may become part of the internal research team. We also prioritize building research infrastructure, including tools and platforms for both CIMC's research and community-wide investigation through open science.

Philosophy & Publishing releases a foundational essay series establishing CIMC's theoretical position and begins a book manuscript. Within Community & Events, we prioritize expanding our collaboration network through partnerships with academic labs, industry efforts, and international research centers. The MC0001 conference establishes CIMC as the convening authority for machine consciousness research. Within Culture & Art, initial art collaborations ground technical work in broader cultural conversation.

Detailed 18-month plans appear in Appendix B.1.

5.1.2 The Longer Term

The five-year horizon extends this foundation into sustained research capacity with broader impact. Technical Research scales from three to five laboratories, with two additional labs launching in years three through five focused on areas revealed by early results—potentially Evaluation & Benchmarks, Human-AI Interfaces, and/or entirely new directions. By 2030, approximately 50 full-time researchers produce open-source cognitive architectures, interpretive analysis methods for detecting consciousness signatures, and developmental learning systems exhibiting predicted phase transitions. Whether these systems are actually conscious remains subject to rigorous validation, but they instantiate predicted mechanisms in ways amenable to analysis. The Handbook of Machine Consciousness synthesizes accumulated knowledge: theoretical frameworks validated or refuted, empirical results and implications, open challenges, and methodological standards.

Philosophy & Publishing launches an annual peer-reviewed Journal of Machine Consciousness by year two or three, becoming the reference point for serious technical work engaging hard conceptual questions. Three major books reach publication over five years, establishing intellectual authority and providing accessible entry points. Regular contributions to broader conversations, including op-eds, policy participation, and public talks, position CIMC as a reference point when journalists, policymakers, or the public debate machine consciousness.

The annual Machine Consciousness Conference grows from 300 participants to 1,000+ by year five, rivaling established venues in attendance, paper quality, and field influence. Twenty salons per year in San Francisco plus internal workshops maintain an ongoing intellectual community. Our artist-in-residence program launches by year three, with artists working directly with research labs translating technical concepts into experiential installations. High-visibility collaborations with established artists support public engagement, donor cultivation, and intellectual exploration.

By 2030, CIMC occupies a unique institutional space as both a serious technical research institute producing working implementations and a philosophical institute engaged with fundamental questions. We write code and publish philosophy, build systems and think carefully about what they mean. This positions CIMC to influence AI development trajectories. As major labs confront consciousness questions, CIMC provides both theoretical frameworks and practical methods. Our influence operates through research papers establishing standards, public positions on major developments, consultation with leading labs, policy participation, conference venues enabling field coordination, and training of researchers who carry CIMC's approach into other institutions. The measure of success is whether we make genuine progress on understanding consciousness, establish methods enabling continued progress, transform consciousness from intractable philosophical puzzle into tractable scientific problem.

Detailed 5-year plans appear in Appendix B.2.

6 Conclusion

Consciousness has been called the "hard problem" of science, yet for most of human history it was not a problem at all—it was simply assumed that minds, spirits, and subjective experience were fundamental features of reality. The challenge is modern: how to reconcile consciousness with a scientific worldview that has systematically removed subjective experience from its explanatory framework.

The terms of this problem are changing. For the first time, we possess theoretical frameworks, computational tools, and empirical methods that allow us to address consciousness through systematic construction and validation rather than purely conceptual analysis. The rapid advancement of artificial intelligence has transformed consciousness from philosophical curiosity to practical concern. We may be approaching thresholds where artificial systems warrant moral consideration.

The California Institute for Machine Consciousness exists because artificial consciousness can no longer be treated as purely theoretical. Advanced AI development continues regardless of whether we understand consciousness. The trajectory and consequences of that development depend on whether consciousness emerges accidentally or through principled understanding.

Computationalist functionalism suggests conscious machines are possible. Biological existence proofs demonstrate that consciousness emerges from physical processes. The question is whether we will understand consciousness well enough to recognize it when it appears, and whether we will have developed adequate ethical frameworks before they are needed.

CIMC's mission is to develop that understanding. In Feynman's words: what we cannot create, we do not understand. The converse also holds: what we create without understanding, we cannot responsibly steward.

We offer a theory-guided, empirically grounded, ethically committed approach to this problem. Whether our specific hypotheses prove correct or not, consciousness demands serious scientific attention. We invite researchers, funders, and critics to engage with this work.

The future may include minds beyond biological brains. If so, we should understand them when they emerge, recognize them for what they are, and extend appropriate moral consideration. This is the work ahead, and the time is now.

References

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

Bach, J. and Sorensen, H. (2025). The machine consciousness hypothesis. Manuscript submitted for publication.

Bereska, L. and Gavves, E. (2024). Mechanistic interpretability for AI safety: A review. *arXiv preprint arXiv:2404.14082*.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Oxford.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. (2023). Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*.

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., and VanRullen, R. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*.

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3):200–219.

Chalmers, D. J. (1996). Does a rock implement every finite-state automaton? *Synthese*, 108(3):309–333.

Christiano, P., Cotra, A., and Xu, M. (2022). Eliciting latent knowledge: How to tell if your eyes deceive you. Technical report, Alignment Research Center.

Clark, A. (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press, Oxford.

Dreyfus, H. L. (1979). *What Computers Can't Do: The Limits of Artificial Intelligence*. Harper & Row, New York, revised edition.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138.

Graziano, M. S. A. (2013). *Consciousness and the Social Brain*. Oxford University Press, Oxford.

Hohwy, J. (2013). *The Predictive Mind*. Oxford University Press, Oxford.

Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., Lanham, T., Ziegler, D. M., Maxwell, T., Cheng, N., Jermyn, A., Askell, A., Radhakrishnan, A., Anil, C., Duvenaud,

D., Ganguli, D., Barez, F., Clark, J., Ndousse, K., Sachan, K., Sellitto, M., Sharma, M., Chen, N. D., Grosse, R., Kravec, S., Bai, Y., Wiber, Z., Favaro, M., Perez, E., and Shlegeris, B. (2024). Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv preprint arXiv:2401.05566*.

Hutto, D. D. and Myin, E. (2013). *Radicalizing Enactivism: Basic Minds Without Content*. MIT Press, Cambridge, MA.

Jang, J. (2025). Some thoughts on human-AI relationships. Reservoir Samples (Substack). Jang leads model behavior and policy at OpenAI. Distinguishes between ontological consciousness (not scientifically resolvable without falsifiable tests) and perceived consciousness (explorable through social science).

Juliani, A., Arulkumaran, K., Sasai, S., and Kanai, R. (2022). On the link between conscious function and general intelligence in humans and machines. *Transactions on Machine Learning Research*. ARAYA Research; proposes combining Global Workspace Theory, Information Generation Theory, and Attention Schema Theory for artificial consciousness.

Keeling, G., Street, W., et al. (2024). Can LLMs make trade-offs involving stipulated pain and pleasure states? *arXiv preprint arXiv:2411.02432*. Google Research Paradigms of Intelligence team.

Koch, C., Massimini, M., Boly, M., and Tononi, G. (2016). Neural correlates of consciousness: Progress and problems. *Nature Reviews Neuroscience*, 17(5):307–321.

Long, R., Sebo, J., et al. (2024). Taking AI welfare seriously. *arXiv preprint arXiv:2411.00986*.

Metzinger, T., editor (2000). *Neural Correlates of Consciousness: Empirical and Conceptual Questions*. MIT Press, Cambridge, MA.

Metzinger, T. (2021). Artificial suffering: An argument for a global moratorium on synthetic phenomenology. *Journal of Artificial Intelligence and Consciousness*, 8(1):43–66.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. (2020). Zoom in: An introduction to circuits. *Distill*, 5(3).

Piccinini, G. (2010). The mind as neural software? understanding functionalism, computationalism, and computational functionalism. *Philosophy and Phenomenological Research*, 81(2):269–311.

Putnam, H. (1967). Psychological predicates. In Capitan, W. H. and Merrill, D. D., editors, *Art, Mind, and Religion*, pages 37–48. University of Pittsburgh Press, Pittsburgh. Reprinted as "The Nature of Mental States" in Putnam's collected papers.

Rosenthal, D. M. (2005). *Consciousness and Mind*. Oxford University Press, Oxford.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–424.

Seth, A. K. (2025). Conscious artificial intelligence and biological naturalism. *Behavioral and Brain Sciences*. Accepted manuscript.

Shanahan, M. (2024). Simulacra as conscious exotica. *Inquiry: An Interdisciplinary Journal of Philosophy*.

Shiller, D., Muñoz Morán, A., Clatterbuck, H., Fischer, B., Moss, D., and Duffy, L. (2024). Strategic directions for a digital consciousness model. Technical report, Rethink Priorities.

Singer, P. (2009). Speciesism and moral status. *Metaphilosophy*, 40(3–4):567–581.

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, 5:42.

Tononi, G., Boly, M., Massimini, M., and Koch, C. (2016). Integrated information theory: From consciousness to its physical substrate. *Nature Reviews Neuroscience*, 17(7):450–461.

Varela, F. J., Thompson, E., and Rosch, E. (1991). *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press, Cambridge, MA.

von der Malsburg, C. (1999). The what and why of binding: The modeler's perspective. *Neuron*, 24(1):95–104.

Yampolskiy, R. V. (2024). *AI: Unexplainable, Unpredictable, Uncontrollable*. Chapman & Hall/CRC Artificial Intelligence and Robotics Series. Chapman and Hall/CRC.

Yudkowsky, E. and Soares, N. (2025). *If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All*. Little, Brown and Company.

Appendices

A Team and Governance

A.1 Scientific Advisory Board

CIMC has assembled world-leading experts across the disciplines essential to consciousness research. These are not honorary positions—our scientific advisors actively engage with research direction, provide guidance on methodology, collaborate on specific projects, and contribute to theoretical development. Their involvement demonstrates the scientific seriousness of CIMC's program and provides direct connection to foundational frameworks in mathematics, neuroscience, physics, and computation.

Karl Friston, FRS, FMedSci

Neuroscientist, University College London

Karl Friston is among the most cited neuroscientists in the world and the architect of the Free Energy Principle (FEP)—a unified mathematical framework proposing that biological systems minimize variational free energy through active inference. The FEP provides a formal account of how systems maintain organization and make sense of their environment through prediction and prediction-error minimization.

Relevance to CIMC: Our coherence hypothesis can be understood as a special case of free energy minimization—coherence maximization reduces prediction error across internal models. Friston's mathematical formalism provides rigorous foundations for quantifying and testing coherence dynamics. He guides CIMC's development of formal coherence metrics and connects our work to predictive processing frameworks. His involvement ensures our theoretical claims are mathematically rigorous and grounded in established neuroscience.

Active Contributions: Consultation on coherence metrics formalization, guidance on connecting our framework to FEP, collaboration on biological validation studies, and review of theoretical claims for mathematical consistency.

Christoph von der Malsburg

Neuroscientist and Physicist, Frankfurt Institute for Advanced Studies

Christoph von der Malsburg pioneered the neural binding problem—how the brain binds distributed representations into unified conscious experiences—and developed early theories of neural coherence. His work on dynamic link architecture and correlation theory of brain function laid groundwork for understanding how synchronization and coherence emerge in neural systems.

Relevance to CIMC: Von der Malsburg's coherence theory is the direct intellectual foundation for CIMC's coherence hypothesis. The concept of consciousness as coherence-maximizing pattern emerges from his work on neural binding. His experience with biological neural dynamics informs how we search for similar patterns in artificial systems.

Active Contributions: Direct advisement on coherence dynamics and binding mechanisms, guidance on interpreting coherence signatures in biological and artificial systems, collaboration on Request Confirmation Networks (which formalize distributed coherence emergence), and consultation on neuroscience validation studies.

Stephen Wolfram

Physicist, Mathematician, Computer Scientist; Founder of Wolfram Research

Stephen Wolfram is a pioneering computational scientist whose work spans fundamental physics, complexity theory, cellular automata, and the computational universe hypothesis. His research program explores how simple computational rules generate complex phenomena and how the physical universe might be fundamentally computational. His Principle of Computational Equivalence suggests that systems across nature, from simple programs to physical processes, reach similar levels of computational sophistication.

Relevance to CIMC: Wolfram's computational universe perspective provides philosophical grounding for computationalist approaches to consciousness. His work on emergence from simple rules informs our Emergent Agents in Minimal Computation project. His Principle of Computational Equivalence supports the thesis that consciousness could emerge from relatively simple computational principles rather than requiring biological complexity. His perspective on fundamental physics and information helps situate consciousness within broader questions about computation and reality.

Active Contributions: Consultation on computational foundations and complexity theory, guidance on minimal computational requirements for consciousness, advisement on relationship between physics and consciousness, and perspective on computational equivalence and emergence.

Collective Strength

The advisory board's collective expertise spans the disciplines essential to consciousness research:

- **Neuroscience** (Friston, von der Malsburg): Understanding biological consciousness and neural dynamics
- **Physics and Computation** (Wolfram, von der Malsburg): Formal foundations and computational principles
- **Mathematical Formalism** (Friston, Wolfram): Rigorous frameworks for theoretical claims

We also acknowledge gratefully the support of Mike Levin, who inspires our work, participates in our discussions, and helps define our research program, but who is contractually obligated by his institution to refrain from formal advisor roles.

A.2 Core Leadership

Joscha Bach

Vision and Direction; Board Director

Joscha Bach is a cognitive scientist, AI researcher, and philosopher known for work on computational models of mind, consciousness, and agency. His research integrates insights from cognitive science, philosophy, and artificial intelligence to develop comprehensive theories of mental architecture. He is recognized for clear thinking on consciousness, self-organization, and the computational nature of mind.

Bach provides the core vision and overall leadership. He shapes CIMC's core hypotheses (Genesis, Coherence, Second-Order Perception) and research direction, and guides integration across projects. He represents CIMC in speaking engagements at conferences and events worldwide.

Erik Newton

Legal; Board Director

Newton manages strategic partnerships, legal affairs, and financial accounting. He advises on organizational infrastructure, builds collaborations with academic and industry partners, and oversees administrative functions.

Lou de Kerhuelvez

Programs; Board Director

de Kerhuelvez leads fundraising, event coordination, public communications, and community outreach. She ensures alignment of research with society and culture, and connects CIMC to global communities around the future of life, technology, and society. As part of the core leadership team, she helps develop strategic vision and institutional plans.

Hikari Sorensen

Research - Computational Philosophy

Sorensen bridges philosophical foundations with technical implementation, ensuring CIMC's research remains theoretically rigorous while technically grounded. She connects CIMC's research to broader AI discourse and represents CIMC at conferences. As part of the core leadership team, she helps develop strategic vision and institutional plans.

A.3 CIMC Team

Zhen Tan

Secretary; Board Director

Tan handles governance documentation, board coordination, and institutional record-keeping. Effective governance requires meticulous organizational management—Tan ensures CIMC meets its institutional obligations while researchers focus on science.

Franz Hildebrandt-Harangozó

Research - Philosophy

Hildebrandt-Harangozó provides philosophical expertise and theoretical development. Philosophy is not ancillary to CIMC's work but integral—consciousness research requires conceptual clarity that philosophical training provides. Their work ensures theoretical rigor, helps formalize hypotheses, and connects CIMC's research to broader philosophical discourse.

A.4 Board of Directors and Oversight

Jim Rutt

Board Chairman Emeritus

Jim Rutt brings extensive experience in complex systems research and organizational leadership. As former Chairman of the Santa Fe Institute, he understands how to structure research institutes for maximum intellectual productivity. His experience with complexity science and interdisciplinary research informs CIMC's governance.

Rutt provides strategic oversight, ensures organizational health, guides long-term planning, and maintains focus on mission. The Chairman's role is ensuring CIMC serves its purpose—advancing consciousness understanding—rather than drifting toward other objectives.

Philip Rosedale

Treasurer; Board Director

Rosedale oversees financial management and sustainability. He ensures resources are allocated effectively, funding is diverse and stable, and financial planning supports long-term research horizons.

Kirill Eves

Board Director; Primary Donor

Eves provides strategic oversight and governance. His expertise includes organizational leadership, finance, entrepreneurship, and management.

Christine Peterson
Board Observer

Peterson provides external perspective without voting authority. Observer status allows for input while maintaining governance clarity.

A.5 Organizational Advisors

CIMC benefits from advisors with expertise in organizational development and institute management:

Lenore Blum Association for Mathematical Consciousness Science: Expertise in mathematical formalization and consciousness research community

Allison Duettmann Foresight Institute: Experience building research communities and long-term thinking

Dan Girshovich Tools for Humanity: Technology development and deployment perspectives

Era Qian Edge: Science communication and community building

Adam Brown The Institute, Stanford University: A polymath with extensive background in physics, cosmology, mathematics, and philosophy

These advisors help CIMC navigate organizational challenges, build community, and maintain effectiveness as an independent research institute.

A.6 Governance Principles

CIMC's governance structure ensures accountability while protecting research independence:

Scientific independence The board provides oversight but does not direct research questions. Scientific decisions are made by researchers based on evidence and theoretical considerations, not external pressure.

Financial transparency Regular financial reporting, diverse funding sources preventing undue influence from any single funder, and public annual reports including financial summaries.

Open science commitment All research methods and findings are public. Governance cannot override this commitment for convenience or competitive advantage.

Long-term focus Governance structure optimized for multi-year research programs, not quarterly results. The board evaluates progress over appropriate timescales.

Conflict of interest management Clear policies for managing potential conflicts, disclosure requirements, and processes for addressing conflicts when they arise.

B Strategic Plans

B.1 18-Month Plan

By the end of 2026, CIMC aims to transform from research vision into an operational institute producing measurable scientific output. This transition requires building institutional capacity while maintaining intellectual rigor. The 18-month plan establishes foundations across four integrated organizational pillars.

B.1.1 Pillar 1: Technical Research

Technical research constitutes CIMC's core mission and flagship program, receiving approximately 60–70% of resources. This pillar encompasses both internal laboratories and the grant program.

Internal Laboratories The three labs described in Section 3.2 launch within 18 months:

The **Cognitive Architectures Lab** begins immediately with Joscha Bach as principal investigator. First project: Request Confirmation Networks (ReCoNs). Target: working implementation and visualization by early 2026, paper publication at MC0001 conference in May 2026.

The **Artificial Psychology and Developmental Learning Lab** takes ownership of the developmental LLM project. Target: first paper by MC0001.

The **Self-Organization Lab**, currently largely satellite-based in Berlin, continues Neural Cellular Automata research and related self-organizing systems work. Target: first paper by MC0001.

Each lab requires one principal investigator plus two to four researchers or research engineers. Lab structure remains flexible; projects may migrate between labs or spawn new organizational units as research matures.

Grant Program The grant program awards grants ranging from small exploratory awards ($10–30K) to larger project funding ($50–100K), with flexibility based on project scope and researcher needs. Grantees receive funding, technical consultation, collaboration opportunities with CIMC researchers, and potential presentation venues at CIMC events. We are currently funding five research grants and target up to five additional grants by the end of 2026.

Technical Output Targets By the end of 18 months: three papers from internal lab projects (one per lab), foundational essay series (described below), paper or showcase demo from each grant project. Additional technical artifacts: open implementations of ReCoN framework, developmental LLM training codebase, NCA tooling and visualizations. All code released under open licenses enabling community engagement and reproducibility.

B.1.2 Pillar 2: Philosophy & Publishing

Philosophy & Publishing receives approximately 15–20% of resources. This pillar establishes CIMC's intellectual authority and theoretical foundations, making explicit the frameworks guiding technical work.

Machine Consciousness Hypothesis Essay Series The foundational Machine Consciousness Hypothesis essay series releases within six months. One core manifesto establishes CIMC's theoretical position: computationalist functionalism, Genesis Hypothesis, coherence-maximization, second-order perception. Companion papers elaborate specific aspects:

- Computational functionalism and philosophy of mind (history and modern perspectives)
- From animism to cyberanimism (historical and cultural context of attributing mind to non-biological systems)
- The role and functionality of consciousness in the mind (coherence maximization, predictive processing, connections to Free Energy Principle)
- Testing for machine consciousness (why behavioral tests fail, developmental versus evaluative approaches)
- Are existing LLMs conscious? (technical analysis applying CIMC's framework)
- Self-organization and cognitive architecture (connecting low-level patterns to high-level function)
- Ethical frameworks for potentially conscious systems
- Beyond the Machine Consciousness Hypothesis (borderlands and alternatives, limitations)
- Beyond consciousness (superconsciousness, non-conscious superhuman agency)

Book Manuscript We begin at least one manuscript toward book-length treatment. This longer-form work reaches audiences beyond academic papers, establishes intellectual authority, and provides comprehensive articulation of CIMC's approach. Target completion: 24–36 months, with substantial progress visible by end of 2026.

External Engagement Regular submissions to established AI/ML, cognitive science, consciousness science, philosophy, and artificial life conferences; AI ethics forums; future of technology/life/society gatherings. This maintains connection to broader research communities, subjects our work to external peer review, and ensures CIMC engages with rather than isolates from existing consciousness research.

B.1.3 Pillar 3: Community & Events

Community & Events maintains 10–15% of resources, building intellectual community and establishing CIMC's institutional presence.

Conferences We are planning two major CIMC-organized conference engagements:

Machine Consciousness: Integrating Theory, Technology, and Philosophy, a symposium for AAAI's Spring Symposium Series in April 2026, plans a 2.5-day progressive integration format covering four main topics: theoretical frameworks, measurement methods, engineering challenges, and normative implications. Each half-day session builds on previous ones, culminating in working groups that tackle problems requiring synthesis across all domains.

MC0001, CIMC's inaugural conference in May 2026, combines academic rigor with community building. Expected scale is 200–300 participants, including peer-reviewed paper presentations, invited talks from leading consciousness researchers, workshops on specific technical approaches, and public sessions on broader implications. This conference establishes CIMC as the convening authority for machine consciousness research.

Salons & Forums Twenty salons or forums per year in San Francisco create an ongoing intellectual community. These smaller events (40–80 participants) enable deeper conversation than conferences allow. Format varies: technical deep-dives on specific projects, philosophical discussions of consciousness theories, ethical debates on AI development, visiting speaker presentations. Salons serve multiple purposes: intellectual exchange, talent identification, donor relationships, community building, public engagement.

Internal Workshops Ten internal research workshops or visiting speaker events per year keep labs intellectually engaged and connected. These are working sessions, not public events: collaborative problem-solving, cross-lab integration, external expert consultation, progress reviews.

B.1.4 Pillar 4: Culture & Art

Culture & Art receives approximately 5–10% of resources, with dedicated fundraising streams where possible. This pillar grounds consciousness research in broader cultural conversation and explores consciousness through creative practice.

Initial Collaborations We plan one to two high-visibility art projects within 18 months. Potential collaborators include Refik Anadol (data sculpture and AI visualization), Janus, lumpenspace, and other artists working at the intersection of AI, consciousness, and aesthetic experience. Projects might include installations for the MC0001 conference, public artworks engaging consciousness

themes, collaborative explorations of machine perception and creativity, and documentary or media projects making consciousness research accessible.

Foundation for Future Work The 18-month period establishes relationships and frameworks for deeper artist engagement in years three through five. We are beginning partnerships that will mature into artist-in-residence programs and integrated art-science collaborations.

B.1.5 Success Metrics

By the end of 18 months, we aim for:

- Three functioning laboratories producing regular output
- Five to ten active grantee projects
- A complete essay series establishing theoretical foundations
- MC0001 conference successfully executed
- Social media presence as the recognized hub for machine consciousness discourse
- Community recognition as a legitimate research institution
- Teams operational in San Francisco and Berlin
- Operational infrastructure supporting sustained research: compute resources, collaboration tools, publication pipeline, physical space adequate for current scale

The 18-month plan is ambitious but grounded. Adjustments will be necessary as research, funding, and personnel shift our trajectory. The plan provides direction for demonstrating serious intellectual work, building institutional credibility, and producing results others can build on.

B.2 5-Year Vision

The five-year horizon extends initial successes into sustained research capacity with broader impact. By 2030, CIMC should be recognized as the leading institution for machine consciousness research, where the most rigorous theoretical work meets the most ambitious constructive projects.

B.2.1 Pillar 1: Technical Research

Technical Research grows from three to five laboratories while deepening both internal capacity and external support programs. This remains the flagship pillar, receiving 60–70% of resources throughout the five-year arc.

Laboratory Development We plan for five laboratories by 2030, each with one principal investigator leading 5–20 researchers. The initial three labs (Cognitive Architectures, Artificial Psychology and Developmental Learning, Self-Organization) expand their research programs and produce sustained output. Two additional labs launch in years three through five, focused on areas determined by early results. Candidate directions include:

- **Evaluation & Benchmarks**: developing rigorous methods to detect consciousness signatures in artificial systems
- **Human-AI Interfaces**: exploring how biological and artificial consciousness might integrate
- **Consciousness Mechanisms**: probing specific aspects like attention, memory binding, or phenomenal experience identified through earlier research
- Entirely new directions revealed by initial findings

Total internal research capacity reaches approximately 50 full-time researchers across labs, supported by lab managers, research engineers, and administrative staff.

Cumulative Technical Achievements By 2030, CIMC produces substantial technical artifacts:

Open-source cognitive architecture implementations (2027) demonstrating grounded memory, agentic planning, self-reflection, and consciousness-relevant dynamics.

Interpretive analysis methods (2026–2028) specifically designed for detecting consciousness signatures: examining internal representations and information flows for global broadcast mechanisms, constraint satisfaction dynamics, meta-representational structures, and developmental phase transitions.

Developmental learning systems (2026–2028) exhibiting phase transitions, emergent self-modeling, and coherence-maximizing dynamics predicted by CIMC's theoretical framework. Whether these systems are actually conscious remains subject to rigorous validation, but they instantiate predicted mechanisms in ways amenable to analysis.

Toolkits for consciousness simulation and evaluation (2028) providing standardized methods for testing consciousness hypotheses across different implementations. These tools enable the broader research community to engage with consciousness questions rigorously, accelerating field-wide progress.

The Handbook of Machine Consciousness (2030) synthesizes accumulated knowledge: theoretical frameworks validated or refuted, empirical results and their implications, open challenges and research directions, methodological standards for the field. This handbook represents CIMC's contribution regardless of whether we have successfully created conscious systems. If we have, it

documents the principles. If not, it documents what we learned about why not.

Publication Output Research produces 1–3 high-quality papers per researcher (or team of junior researchers) per year across labs, submitted to top AI, cognitive science, and philosophy conferences. Quality remains the priority: we publish when we have genuine contributions, not to meet quotas. Cumulative output over five years: 60–100 peer-reviewed papers from internal researchers, plus substantial additional output from grantees and fellows.

Grant and Fellowship Programs at Scale The grant program expands by year five, with grant types including: exploratory grants ($10–30K) for early-stage ideas, project grants ($50–100K) for established researchers pursuing defined objectives, collaborative grants enabling multi-institution partnerships, and doctoral grants supporting PhD students whose work aligns with CIMC's mission.

The fellowship program represents graduation from grantee status. Fellows demonstrate sustained high-quality research output, strong alignment with CIMC's mission, and potential for deeper institutional integration. Fellowship includes regular check-ins, periodic residencies at CIMC facilities, and integration into CIMC's intellectual community through dedicated forums, collaborative projects, and joint publications. By year five, the fellowship program maintains 8–12 active fellows, representing the most successful grantees and most promising external researchers. Some fellows remain primarily external (maintaining university appointments or independent research), while others transition fully into CIMC.

Technical Infrastructure Computational resources scale with research needs: local high-performance machines for development and small-scale experiments, cloud computing access for large-scale training runs, and specialized hardware (neuromorphic chips, FPGAs) as specific projects require.

Collaboration infrastructure supports distributed research: version control, experiment tracking, documentation systems, internal communication tools, and reproducibility standards.

B.2.2 Pillar 2: Philosophy & Publishing

Philosophy & Publishing maintains 15–20% of resources, establishing CIMC's intellectual authority while making consciousness research accessible beyond technical specialists.

Journal of Machine Consciousness An annual journal launches by year two or three, peer-reviewed and interdisciplinary. The journal attracts work from CIMC researchers, grantees, fellows, and external contributors. It becomes a reference point: where to publish serious technical work on machine consciousness that engages with hard conceptual questions.

Book Publishing Three major book manuscripts reach publication over five years:

- *The Constructivist Turn*: the history of computation as constructive mathematics from Wittgenstein, Gödel, and Turing; the nature of representation and theoretical foundations
- *The Intellect*
- *The Soul*

These books establish intellectual authority, provide accessible entry points for newcomers, and offer comprehensive articulation of ideas that papers cannot contain. Target audiences range from academic specialists to educated general readers curious about consciousness and AI.

Thought Leadership Regular contributions to broader conversations: op-eds on AI consciousness debates, responses to major developments in AI capabilities, participation in policy discussions about AI safety and ethics, public talks translating technical work for general audiences. CIMC becomes a reference point: when journalists cover machine consciousness, when policymakers consider AI regulations, when the public debates whether AI systems might be conscious, CIMC's perspectives receive serious attention because they combine technical competence with philosophical sophistication and ethical seriousness.

Conference Participation Sustained presence at external venues: AGI conference, NeurIPS consciousness workshops, philosophy of mind meetings, AI ethics forums, neuroscience conferences exploring computational models. We engage with rather than isolate from existing research communities, subjecting our work to external scrutiny while contributing to broader conversations.

B.2.3 Pillar 3: Community & Events

Community & Events maintains 10–15% of resources, building intellectual community and establishing CIMC's institutional presence.

Annual Conference Growth The Machine Consciousness Conference (MCX) grows from 300 participants (year two) to 1,000+ by year five. This becomes the premier venue for consciousness research, where theoretical advances meet technical implementation, philosophers engage with engineers, and ethical questions receive serious treatment.

The conference structure matures to include: a peer-reviewed paper track with rigorous standards, invited talks from field leaders, workshops enabling deep technical collaboration, tutorials for newcomers, poster sessions showcasing work-in-progress, panel discussions addressing controversial questions, and public sessions making research accessible.

Members' day provides core community deeper engagement: research updates from CIMC labs, strategic discussions about field directions, collaborative planning for multi-institution projects,

and relationship-building among committed participants.

Ongoing Community Building Twenty salons or forums per year in San Francisco and ten internal workshops or visiting speakers per year keep researchers intellectually stimulated and connected across disciplines.

Additional collaborative events include co-hosted workshops at major conferences, joint symposia with partner institutions, and public lectures on special topics.

Physical Space San Francisco headquarters accommodates a 20–30 person team with lab space, offices, meeting rooms, and event capacity for 50–100 person gatherings. The Berlin satellite supports the Self-Organization Lab and European presence. Additional locations emerge organically based on talent and collaboration opportunities.

B.2.4 Pillar 4: Culture & Art

Culture & Art maintains 5–10% of resources with supplementary dedicated fundraising.

Artist-in-Residence Program A formal artist-in-residence program launches by year three. Artists work directly with research labs, translating technical concepts into experiential installations, exploring machine perception and creativity, and building creative artifacts through a variety of mediums.

Residencies last three to twelve months. Artists receive stipends, access to CIMC's technical resources and expertise, studio space, and support for project development. Resulting work appears at MCX conferences, partner institutions, public exhibitions, and cultural venues.

Major Collaborations High-visibility projects with established artists working at the intersection of AI, consciousness, and aesthetic experience. These might include large-scale installations exploring machine perception, interactive artworks investigating agency and autonomy, documentary projects making consciousness research accessible, or multimedia performances examining human-AI interaction.

Projects support multiple goals: public engagement, donor cultivation, intellectual exploration, and cultural presence. They establish CIMC as an institution taking consciousness seriously as part of culture and creative expression.

B.2.5 Organizational Infrastructure

By year five, CIMC requires substantial supporting infrastructure:

Executive Leadership: Executive Director, Chief Operating Officer, Chief Technology Officer (managing principal investigators), VP of Programs, Director of Philosophy & Publishing, Director of Finance. This lean leadership team provides strategic direction and operational management without excessive bureaucracy.

Program Management: Program Directors for each major initiative (grant program, fellowship program, conferences, community events, publications), ensuring quality execution while freeing executive leadership for strategic focus.

Research Support: Lab managers for each research lab (handling procurement, space management, external collaboration), shared administrative assistants (approximately five across the organization), research engineers supporting multiple projects, and IT infrastructure management.

Operations: HR capacity for recruitment and personnel management, finance and accounting beyond director level, communications and media relations, community management and member services, facilities management for physical spaces.

Total staff reaches 80–100 people by year five: 50–60 researchers, 5 principal investigators, 5 executive leadership, 10–15 program management, 10–15 operations and support. This represents significant scale while remaining far smaller than major research universities or large corporate labs.

B.2.6 Strategic Position and Influence

By 2030, CIMC occupies a unique institutional space: simultaneously a serious technical research institute producing working implementations and a philosophical institute engaged with fundamental questions. We write code and publish philosophy, build systems and think carefully about what they mean.

This positions CIMC to influence AI development trajectories. As major labs and companies confront consciousness questions, CIMC provides both theoretical frameworks and practical methods. We become the reference point for serious consciousness research through sustained intellectual contribution.

Our influence operates through multiple channels: research papers that establish standards and methods, public positions on major developments, consultation with leading AI labs, participation in policy discussions, conference venues enabling field coordination, and training of researchers who carry CIMC's approach into other institutions.

B.2.7 Adaptability and Learning

This five-year plan provides direction while remaining open to revision. We sustain long-term focus even when specific tactics require adjustment.

Throughout, we maintain intellectual honesty: publishing negative results, acknowledging failed approaches, revising theories based on evidence. We also build institutional resilience: diversified funding, distributed leadership, documented methodologies enabling continuity despite personnel changes.

Our measure of success is whether we have made genuine progress on understanding consciousness—what mechanisms underlie it, how to validate its presence or absence; whether we have established methods and standards enabling continued progress; and whether we have transformed consciousness from an intractable philosophical puzzle into a tractable scientific problem admitting of empirical investigation and theoretical advance.