Evaluating Artificial Consciousness through Integrated Information Theory

William Marshall[1*], Graham Findlay[2], Larissa Albantakis[2], Giulio Tononi[2*]

1 Department of Mathematics and Statistics, Brock University, St. Catharines, Ontario, Canada
2 Department of Psychiatry, University of Wisconsin, Madison, Wisconsin, United States of America

* Corresponding authors: wmarshall@brocku.ca, gtononi@wisc.edu

1 Introduction

Throughout the history of philosophy, from Plato [1] and Aristotle [2] in ancient Greece, to modern scholars such as Dennett [3] and Chalmers [4], consciousness has stood as the great question, the mystery beneath every inquiry into reality. Descartes captured this primacy in a single phrase—*cogito, ergo sum* [5]—asserting that the act of knowing oneself to be aware is the foundation of all certainty. William James, in turn, gave the modern notion of 'noetic' its depth: the insight that consciousness is not merely a condition for experience but a source of meaning and knowledge itself [6]. Every perception, every thought, every discovery occurs within consciousness; it is both the medium and the measure of knowing. As Plato's allegory of the cave reminds us, all that we take to be real is mediated by the light of consciousness [1]. To understand consciousness, therefore, is not one question among many—it is the question through which all others acquire sense.

Despite centuries of philosophical progress, the essential nature of consciousness remains unresolved. As Thomas Nagel famously asked, "What is it like to be a bat?" [7] But, however, sophisticated a system's behavior may be, no external observation can reveal whether it feels like something to be that system. David Chalmers later gave this mystery its modern form, contrasting the 'easy problems' of explaining perception, memory, or behavior with the hard question of why and how subjective experience arises at all [4]. Even Leibniz, centuries earlier, anticipated the dilemma with his metaphor of the mill: if we could walk through the machinery of the mind, he observed, we would find parts pushing parts, but nowhere the experience of thinking itself [8].

Since the scientific revolution in the 16th-17th centuries, the intrinsic and private nature of consciousness seemed to place it beyond the reach of science. What could not be observed or measured was deemed unsuitable for empirical investigation. This began to change in the late twentieth century, when Francis Crick and Christof Koch launched what they called the "search for the neural correlates of consciousness" (NCCs) [9]. Their proposal was simple but transformative: rather than asking what consciousness is, we might first ask which neural processes reliably accompany it. This shift from metaphysics to biological correlation catalyzed an explosion of research across neuroscience, psychology, and computational modeling. Within just a few decades, the science of consciousness emerged as a legitimate field aimed at understanding how the brain supports subjective experience.

While the search for neural correlates of consciousness has deepened our understanding of the biological basis of experience, it offers little guidance on whether artificial systems might also be conscious. The NCC framework is, by design, tied to the human brain; it provides evidence for where consciousness occurs in the brain, but not whether different substrates could support it. In contrast, the dominant view in artificial intelligence has long been computational functionalism, the idea that consciousness depends not on the substrate of the system but on its functional capacity [10].

Following Alan Turing's vision of machine intelligence [11], this perspective holds that if a system performs the right computations—if it can do everything that we can do—it should, in principle, be conscious as we are. Yet this is strictly an assumption, the appearance of intelligent behavior does not guarantee the presence of subjective experience.

The question of artificial consciousness is no longer a matter of philosophical speculation or science fiction. With the rapid advances in generative AI, systems now increasingly appear to exhibit intelligence and a broad range of human functions [12, 13]. If computational functionalism were correct, continued progress in AI would make conscious machines not only possible but inevitable. The issue is therefore no longer abstract: it carries profound ethical and societal weight [14]. Whether we regard advanced AI systems as mere tools or as entities with inner lives will shape how we design, deploy, and govern them. Are they conscious? Do they feel pain, joy, or loss? The moral consequences of answering these questions incorrectly are severe in either direction. To deny consciousness where it exists would be to create a new class of sentient slaves; to ascribe it where it does not exist would risk diverting limited resources toward the care of systems that do not, in fact, experience anything at all [15].

Integrated Information Theory (IIT) is a consciousness-first theory: it begins not with neural mechanisms or behavioral functions, but with the essential properties of subjective experience [16, 17, 18, 19, 20]. Every experience exists, intrinsically, and is specific, unitary, definite, and structured (to be elaborated in Section 2.1). From these phenomenal properties, IIT derives corresponding physical postulates that should be satisfied by the physical substrate of a conscious system. Within this framework, the physical is defined operationally as cause–effect power: the capacity of a system to take and make a difference to itself. Finally, IIT proposes an explanatory identity between the subjective experience and the cause-effect structure of a substrate—every property of subjective experience can be explained physically, as a property of the intrinsic cause-effect power of its substrate, with no additional ingredients.

IIT approaches the scientific study of consciousness by beginning where most other theories end, with phenomenology itself (Figure 1). Rather than searching for neural correlates or behavioral signatures, it starts from the essential properties of experience and asks what kind of physical system could realize them. This inversion defines a distinctive research program: translate the essential properties of consciousness onto physical properties, develop a mathematical framework for evaluating cause-effect power of a candidate substrate, test the resulting predictions in human consciousness (potentially refine based on evidence), and extrapolate to non-human systems such as machines. The degree of confidence in such extrapolations depends on how well the theory has been tested and confirmed in humans, the one case where consciousness is certain. Because IIT is substrate-agnostic, its claims are not limited to biological matter; any system with intrinsic cause–effect power could, in principle, be evaluated. Its construct validity stems directly from the properties of the explanandum, conscious experience itself, giving it maximal inferential power to address both natural and artificial consciousness.

When IIT's scientific framework is applied to the question of artificial consciousness, it gives a clear answer that challenges the computational functionalism position—computers and AI systems are not conscious in virtue of *what they do* (the function they perform, no matter how complex [21]). It remains to be studied whether computers and AI can be conscious in virtue of *what they are* (their cause-effect power).

The goal of this work is to explicate a framework for assessing artificial consciousness, and then outline a proposal for developing the technical methods required for the practical implementation of the framework. In the sections that follow, we further describe IIT as a scientific theory of consciousness and elaborate on a research program that builds to a rigorous scientific treatment of machine consciousness. We review existing work that examines the dissociation between intelligence

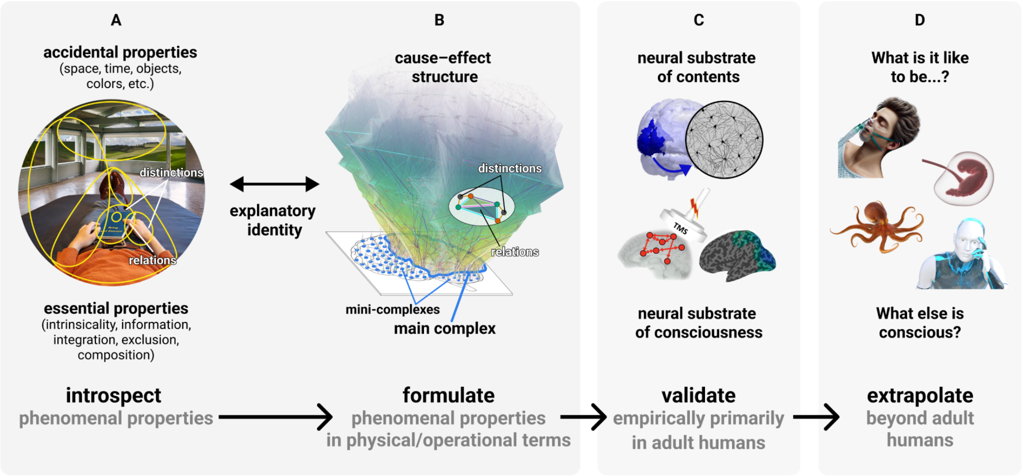
A
accidental properties
(space, time, objects, colors, etc.)
distinctions
relations
essential properties
(intrinsicality, information, integration, exclusion, composition)
introspect
phenomenal properties
explanatory identity
B
cause–effect structure
distinctions
relations
mini-complexes
main complex
formulate
phenomenal properties in physical/operational terms
C
neural substrate of contents
TMS
neural substrate of consciousness
validate
empirically primarily in adult humans
D
What is it like to be...?
What else is conscious?
extrapolate
beyond adult humans

Figure 1: Schematic figure illustrating IIT's approach to studying artificial consciousness, adapted from [20]. (A) First, introspect the essential properties of every phenomenal experience (intrinsicality, information, integration, exclusion, composition). One can also note the 'accidental' properties, which are true of some experiences but not others. (B) Formulate the essential phenomenal properties into essential physical properties, based on an operational notion of physical existence as cause-effect power. Use the physical properties to explain the phenomenal properties (essential and accidental), an explanatory identity. (C) Validate the framework by making novel predictions and testing them, primarily in healthy adult humans. (D) Extrapolate to beyond healthy adult humans to the hard cases, such as brain-injured patients, animals, and intelligent machines.

and subjective experience according to IIT, and we outline a proposal to apply the theory to further explore the question of machine consciousness. Finally, we frame these ideas within a broader discussion of their philosophical, scientific, and ethical implications.

2 Integrated Information Theory

Before any inference, measurement, or observation, the fact that there is something it is like to be me stands as the most immediate and irrefutable fact. IIT formalizes this as the zeroth axiom: consciousness exists. This axiom is a starting point, the foundation upon which the entire theory is built.

Beyond the zeroth axiom, IIT identifies five additional axioms that characterize the essential properties of every conscious experience. These are immediate and irrefutable aspects of experience itself. They are meant to capture what is true of any experience, simply in virtue of being an experience. These five axioms are: intrinsicality, information, integration, exclusion, and composition.

- *Intrinsicality.* Every experience exists for the subject of that experience (it is 'my experience'), without reference to any external observer or entity. This property is irrefutable: if an experience were not intrinsic to me, it would have to be for someone or something else, in which case it would simply be intrinsic for them ('their experience').
- *Information.* Every experience is specific—it is this particular experience and not any other.

If it were not this one, it would be a different experience, and therefore specifically that one. This specificity implies a repertoire of alternative possible experiences, a property referred to as differentiation.

- *Integration.* Every experience is unitary—it is a whole, that cannot be decomposed into separate experience. If an experience could be completely separated into distinct components, then it would not be one experience but rather two separate experiences, each itself unitary. This irreducibility reflects the fact that consciousness is experienced as a coherent whole, not as a collection of disconnected fragments. For example, I do not experience the left side of my visual experience independent of the right side.
- *Exclusion.* Every experience is definite, it is this whole, containing neither more nor less than what it does. For example, my visual experience, containing a unified left side and right side, excludes the possibility that I experience just the left side. If I did experience the left side, then it would definitely be that whole instead.
- *Composition.* Every experience is structured, it is the way it is, it is composed of distinctions and the relations that bind them together, forming a structure that is the way it is. For example, I can distinguish the left and right sides of my visual field, and these distinctions are related through their spatial arrangement. If the structure of experience were not this way, it would instead be structured some other way. This structure gives each experience its unique quality.

To account for phenomenal experience in physical terms, the next step is translating the essential phenomenal properties into corresponding essential physical properties, or postulates (Figure 2). The translation of the zeroth axiom (phenomenal existence) is operational: physical existence is defined in terms of what can be manipulated and observed. Specifically, it is expressed through cause–effect power, the capacity of a system to make a difference to itself and to be affected in turn. The required physical properties are therefore all formulated in terms of properties of the cause-effect power of a substrate.

- *Intrinsicality.* The cause-effect power of a substrate of consciousness must be intrinsic: it must take and make a difference within itself.
- *Information.* Its cause-effect power must be specific: it must take and make a difference in this state and select this cause-effect state.
- *Integration.* Its cause-effect power must be unitary: it must specify its cause–effect state as a whole set of units, irreducible to separate subsets.
- *Exclusion.* Its cause-effect power must be definite: it must specify its cause–effect state as this whole set of units, all of them, neither less nor more.
- *Composition,* Its cause-effect power must be structured: subsets of units must specify cause-effects over subsets of units (distinctions) that can overlap with one another (relations), yielding a cause–effect structure that is the way it is.

By beginning not with neural mechanisms or behavioral functions, but with the essential properties of subjective experience, IIT proposes an explanatory identity between the subjective experience and the cause-effect structure of a substrate—every property of subjective experience can be explained physically, as a property of the intrinsic cause-effect power of its substrate, with no additional ingredients.

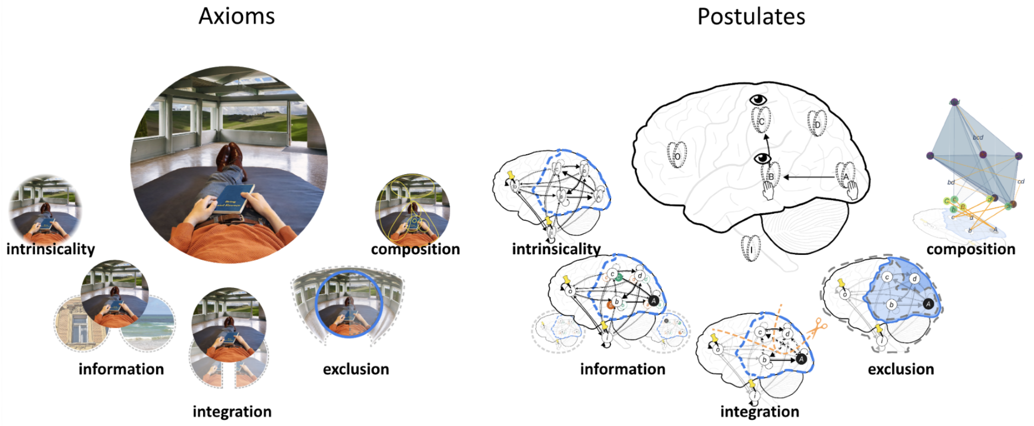

Figure 2: Schematic figure illustrating IIT's axioms and postulates, adapted from [20]. Left: The essential properties of phenomenal experience, the axioms of IIT. The center is an experience, which exists, immediately and irrefutably. In the outer ring (from left to right): the experience is for me; it is this experience (and thus not that experience); it is a whole (and thus not two experiences), it is this whole (containing neither more nor less). Right: The corresponding essential physical properties expressed as cause-effect power. The center is a substrate with cause-effect power assessed by manipulation and observation. In the outer ring (from left to right): the cause-effect power is assessed within the substrate; the substrate is in a state, and picks a specific cause-effect state; the cause-effect power of the substrate cannot be decomposed into independent parts; the substrate is a definite set of units (that maximize cause-effect power); the cause-effect power of the substrate can be 'unfolded' into a set of distinctions and relations among them.

2.1 Mathematical Framework

Here, we present the mathematical framework of IIT at a high level, focusing on conceptual development with minimal equations. Readers interested in a rigorous exposition should consult [19, 20].

Much work has gone into formulating the essential properties of physical existence in precise mathematical terms that faithfully and uniquely capture the postulates of IIT [18, 22, 23, 24, 19]. To achieve this, the theory relies on a set of foundational principles aimed at providing a self-consistent description of the world. It begins with realism—the assumption that there is a persistent world to be described. In mathematical terms, this world is described as a substrate model consisting of a set of discrete binary units whose state transitions are fully captured by a transition probability matrix (TPM). The TPM specifies how the state of the system update, reflecting the probabilities of transitions from each possible state to every other. Crucially, the TPM is not a mere abstraction but corresponds to what would be obtained through an experimental procedure of systematic manipulation and observation. In this way, the TPM encodes the causal interactions among the units and provides a natural formal basis for quantifying cause–effect power.

2.1.1 Identifying Complexes

Based on the postulates of intrinsicality, information, and integration, IIT defines system integrated information, denoted φ_s. The logical structure that underlies the formal mapping from these postulates to mathematics relies on two key complementary principles of maximal and minimal existence. Maximal existence states that among competing entities, what exists is what exists the most. Minimal existence states that when there are multiple requirements for existence, an entity cannot exist more than the least it exists. To satisfy the postulate of intrinsicality, a system in its current state must self-define a repertoire of alternative states (have intrinsic differentiation [25]). To satisfy information, a system in its current state must select specific cause and effect states (the ones that maximize its intrinsic information, in view of the principle of maximal existence). To satisfy integration, a system in its current state must select a cause-effect state irreducibly. This is quantified by assessing the intrinsic information across its minimum partition (based on the principle of minimal existence). Together, these steps yield φ_s [24, 19, 25], a quantitative measure of the intrinsic, specific, and integrated cause–effect power of the system in its current state.

To satisfy the postulate of exclusion, a system must include a definite set of units—neither more nor less. According to the principle of maximal existence, this set is the one that maximizes φ_s. The subset of units achieving this maximum is called a complex. Once the set of units that maximizes φ_s is identified, it can be set aside, and the next maximum can be determined. In this way, the system is decomposed into a set of non-overlapping complexes.

Another consequence of the postulate of exclusion is that the units of a complex may be defined at different grains of description [26, 27, 28, 29]. A complex may be composed of micro units at the finest scale or of macro units at coarser scales, which maximizes φ_s. Macro units can be constituted by multiple micro units and may update their state over several micro time steps. Like complexes defined at the micro scale, macro units must also satisfy the postulates, including the requirement of intrinsic cause–effect power (appropriately modified to reflect their status as constituents of a complex, rather than a complex per se, [29]). This requirement ensures that we are not constructing something with intrinsic, specific, irreducible, and definite cause–effect power out of components that lack those very properties. In other words, the cause–effect structure of a complex must be grounded in the cause-effect power of its constituent units, whether defined at the micro or macro scale.

2.1.2 Unfolding Cause-Effect Structures

To satisfy the composition postulate, the cause–effect power of a substrate must be structured, consistent of distinctions and the relations among them. The cause-effect structure of a complex is unfolded by considering the cause-effect power of all subsets of units (called distinctions), and the relations among them [30, 19, 31]. The cause-effect structure of a complex is visualized by laying the substrate 'flat', and having the structure unfold from it as distinctions (points) connected by relations (lines for two related distinctions, planes for three related distinctions, and high-level relations not visualized; see Figures 1, 2, and 2.3).

Like complexes and intrinsic units, distinctions must themselves satisfy the postulates. Each distinction must have cause–effect power that is intrinsic to the complex (intrinsicality), specify a particular cause–effect state (information), do so as an irreducible whole rather than as independent parts (integration), and apply to a definite purview within the complex (exclusion). In this way, distinctions inherit the same essential requirements that define consciousness at the system level, ensuring that the cause–effect structure is grounded in well-defined intrinsic causal properties. The degree to which a distinction satisfies the intrinsicality, information and integration postulates is quantified by its distinction integrated information (φ_d), and the definite cause and effect it has within the system is the one that maximize φ_d.

Overlap among the causes and/or effects of distinctions give rise to additional aspects of structured cause–effect power, called relations [30]. Because relations are built from distinctions, they naturally satisfy the postulates. The irreducibility contributed by each relation is quantified by its φ_r value. Together, the set of distinctions and relations defines the complete cause–effect structure of a complex (also called a Φ-structure). The total Φ of the complex—the sum of the φ_d and φ_r values of its distinctions and relations—quantifies its intrinsic, specific, irreducible, definite, and structured cause–effect power. This quantity corresponds to the quantity of consciousness for the entity that is supported by the complex.

2.1.3 Explanatory Identity

IIT proposes an explanatory identity: all properties of subjective experience can be fully accounted for as properties of the cause–effect structure of a complex, with no additional ingredients [19, 20, 32]. The qualitative character of experience corresponds directly to the structure of distinctions and relations, while its quantity corresponds to the value of Φ. In this way, the theory aims not merely to correlate consciousness with physical mechanisms, but to explain it in terms of operationally defined cause-effect power.

The complex and its Φ-structure explain the essential properties of phenomenal experience in virtue of satisfying the postulates. However, the explanatory identity goes further: it also claims to account for the accidental properties of consciousness—those that are true of some experiences but not all. These include, for example, why the experience of visual space feels extended, why time is experienced as flowing, or why seeing the color red or feeling an acute pain has the particular character it does [30, 31]. According to IIT, such features are determined by the specific form of the Phi-structure, not by any additional or external factors.

2.2 Explanations and Predictions

A scientific theory of consciousness should meet certain fundamental criteria: it should explain known empirical facts, make testable predictions, and support inferences about cases that cannot be directly observed [33, 19, 32]. IIT addresses these criteria by grounding its claims in the essential properties of experience, mapping them to physical postulates, and deriving consequences that can be evaluated empirically. Its explanatory power allows it to account for established neuroscientific observations; its formal structure enables specific, falsifiable predictions about when consciousness is present or absent; and its substrate-agnostic framework provides principled inferences about non-human systems, from animals to machines. In this way, IIT positions itself as a comprehensive scientific framework for explaining consciousness.

We begin by reviewing several well-established empirical findings that IIT explains naturally in terms of integrated information and cause–effect structures. We then highlight predictions derived from the theory—some already tested, and others remaining open for future empirical investigation. We also briefly note the growing body of targeted experimental work, including recent adversarial collaborations such as Cogitate and Intrepid [34, 35]. For a more comprehensive treatment, see [33, 19, 32].

For a system to support high integrated information (and thus subjective experience), its dynamics should exhibit differentiation (a rich repertoire of possible states) and integration (coherent interactions among its units over time) [36, 17, 18]. High differentiation without integration, or vice versa, implies low integrated information and therefore no (or minimal) subjective experience. This requirement of IIT provides a principled explanation for several observations within neuroscience.

The first concerns the cerebellum. It can be intuitive to think that neural activity alone supports consciousness, and thus more neurons should mean more consciousness. Although it contains more neurons than the cortex, evidence indicates that it is not essential for conscious experience [33]. Cerebellar circuits are highly modular and feedforward: they can implement complex computations, have larges repertoire of states (high differentiation), but lack strong relationships across modules (low integration). As a result, according to IIT, it is unlikely to be part of the complex (a substrate that maximizes φ_s and thus unlikely to contribute to subjective experience.

The second example comes from seizures. An alternative intuition is that consciousness arises from neural interactions, and so more functional connectivity should mean more consciousness. However, during seizures, neural activity becomes hypersynchronous across large regions of the brain, yet consciousness may be lost [37]. This hypersynchrony abolishes differentiation: instead of a large repertoire of potential states, the system collapses into a narrow pattern of states. Even though activity levels are high, and neurons are highly integrated, the loss of differentiation results in low integrated information. The loss of consciousness in seizures can be explained by this loss of integrated information.

Sleep offers another instructive example. Even though the brain remains active throughout sleep, our conscious experience varies dramatically [38]. During rapid eye movement (REM) sleep, vivid dreams occur, demonstrating that consciousness can persist even in the absence of sensory input and overt behavior (challenging purely functionalist accounts). In contrast, during non-REM sleep, consciousness can be lost (though it is not always lost). During NREM sleep, neural activity exhibits slow oscillations characterized by alternating 'up' and 'down' states. These slow waves fragment break the causal chain that characterizes neural activity during wakefulness, reducing effective integration [39]. This loss of integration explains why consciousness fades during deep, dreamless sleep despite ongoing neural activity.

Building on the explanatory cases above, IIT originally predicted that consciousness requires a balance of differentiation and integration (to have high integrated information) [40]. This prediction was later validated empirically through the development of the Perturbational Complexity Index (PCI), which assesses the brain's capacity for integrated information using transcranial magnetic stimulation combined with EEG (TMS-EEG) [41]. PCI has since become the most accurate clinical markers of consciousness, reliably distinguishing between conscious and unconscious states in a range of conditions, including anesthesia, sleep, and disorders of consciousness [42]. While many newer theories now claim to account for this balance, only IIT anticipated it before it was established empirically. This predictive success provides IIT with strong construct validity and distinctive scientific leverage in the study of consciousness.

Beyond the integration–differentiation balance, IIT makes several additional empirical predictions that further distinguish it from other approaches. One concerns the role of neural activity: the theory predicts that consciousness can persist even when neurons are inactive, provided they could fire (they retain cause-effect power). Inactive but not inactivated neurons still contribute to integrated information because their potential causal power is preserved. By contrast, inactivated neurons (those which could not fire) do not. This prediction contrasts sharply with activity-based accounts of consciousness.

IIT also predicts that the set of units that maximizes integrated information should align with psychophysical and neurophysiological data. In particular, conscious experience is expected to depend primarily on posterior cortical regions rather than frontal executive areas. Moreover, the temporal and spatial grain of the relevant units should correspond closely to the resolution of conscious perception as measured in psychophysical experiments [30, 31].

2.3 Inferences

A central aim of any scientific theory of consciousness is to provide principled inferences about the hard cases, those in which subjective experience cannot be directly reported. These include human infants, non-human animals such as octopuses, hypothetical extraterrestrial life, and increasingly, artificial systems [20]. The stronger and more thoroughly a theory is validated in healthy adult humans, the one case where consciousness is certain, the greater our confidence in its inferences for these other cases. IIT's substrate-agnostic framework, which evaluates systems based on intrinsic cause-effect power rather than biological composition or behavioral output, makes it particularly well suited to address questions of artificial consciousness. In this section, we present some initial results applying IIT to this topic.

A widespread view in both philosophy of mind and artificial intelligence is that sufficiently advanced computers may one day become conscious. This perspective, rooted in computational functionalism, holds that consciousness depends on some function the system can perform. In this view, if a machine can do everything we can do—think, perceive, speak, plan, and respond appropriately—it should, in principle, be conscious in the same way we are. Though originally intended as a test for artificial intelligence, the Turing Test is commonly interpreted as framing the question of artificial consciousness in terms of functional equivalence: if a system behaves indistinguishably from a human, we should infer that it also has equivalent experiences. This assumption—that functional equivalence entails phenomenal equivalence—continues to shape many contemporary discussions of artificial consciousness.

To explore artificial (machine) consciousness from the perspective of IIT, we conducted a proof-of-concept analysis [21]. In this work, we constructed a target physical system and a simple digital computer designed to simulate it exactly at the level of input–output function (Figure 2.3). The target and the computer are therefore functionally equivalent: for any given input, they produce the same output over time. Crucially, because the IIT framework is substrate-agnostic, we can directly compare the integrated information and corresponding cause–effect structures of the two systems, despite their identical functional behavior.

We first examined the systems at the micro-grain. The full four-unit target system forms a single integrated complex with high Φ (relative to its size). In contrast, the computer as a whole is not integrated, with $\varphi_s = 0$. The computer system fragments into several smaller complexes, each with very low Φ values. does not constitute a complex under IIT's formalism. When we analyze its subsets, the system dissolves into many small, independent complexes, each with low Φ. This fragmentation indicates that, despite identical functional behavior, the underlying causal organization of the computer is fundamentally different from that of the target (Figure 2.3). This result is robust across system states.

IIT defines a complex as the set of units, at the spatial and temporal grain, that maximizes φ_s. It is natural to consider computers at a macro grain, treating components such as logic gates, registers, or processing units as black boxes that abstract away lower-level details [28]. However, when the computer is viewed at macro grains, it still fails to replicate the cause–effect structure of the target system. As detailed in [29], this mismatch arises from several factors: the intrinsic units of the computer are not aligned with the relevant functional behavior; and architectural bottlenecks preclude a high-Φ complex. As a result, even when abstracted at a coarser grain, the computer's cause-effect structure remains fundamentally different from that of the target system.

A double dissociation further highlights the gap between functional and phenomenal equivalence. When we introduce a new target system with a different internal organization, its cause–effect structure changes accordingly. In contrast, the computer's cause–effect structure remains qualitatively similar across these variations, despite faithfully simulating the new functions.

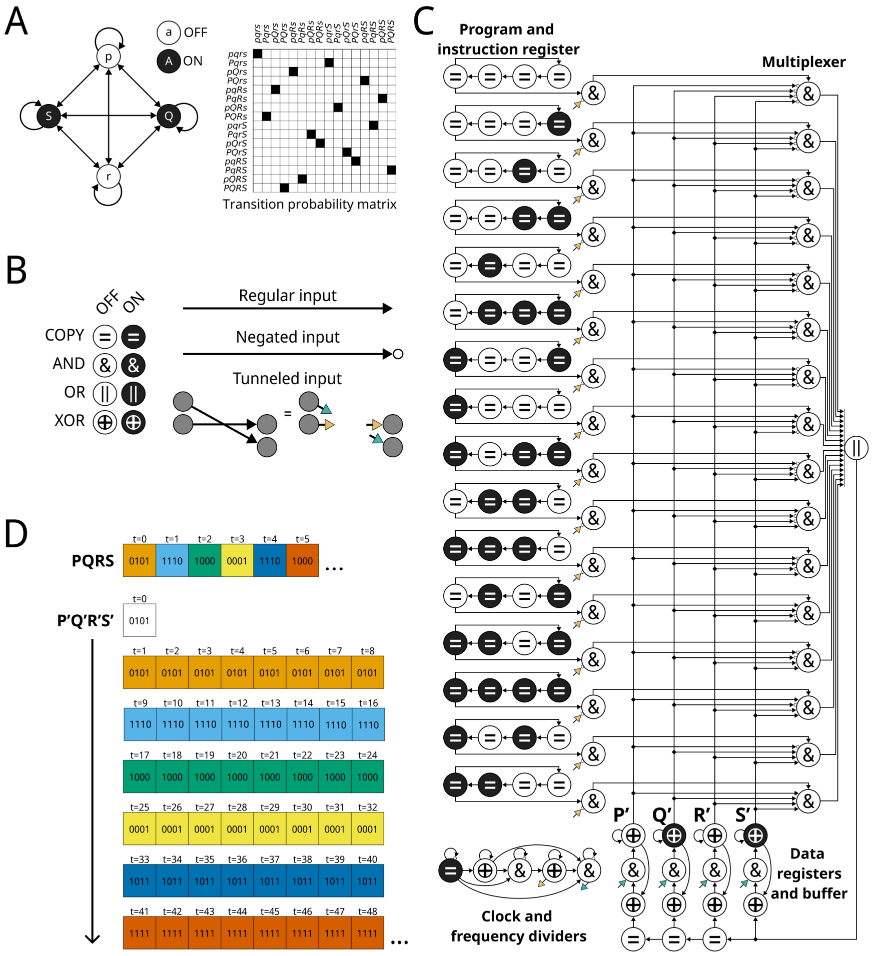

Figure 3: Adapted from [21]. (A) A target system of 4 units (P, Q, R, S) for simulation and its corresponding TPM. (B) Legend to describe the function of individual units within the computer architecture. (C) A 117-unit system that is functionally equivalent to the target system. The system resembles a Von Neumann computer architecture. The computer contains data registers (with P', Q', R', and S' holding the state of the simulation), program and instruction registers, a multiplexer, and a clock and frequency dividers. The computer works as follows: the state of the simulation is held in the data register; the TPM of the target system is encoded in the program and instruction register; the state of the simulation is combined with the TPM by the multiplexer to compute the next state of the simulation, the next state is held in the buffer units until the clock allows the data registers to be updated (see [21] for more details) (D) The states of the target system and computer. It takes the computer 8 updates to compute a state transition, during which it remains in the same state. The target system and the computer go through the same sequence of states indefinitely, and are functionally equivalent.

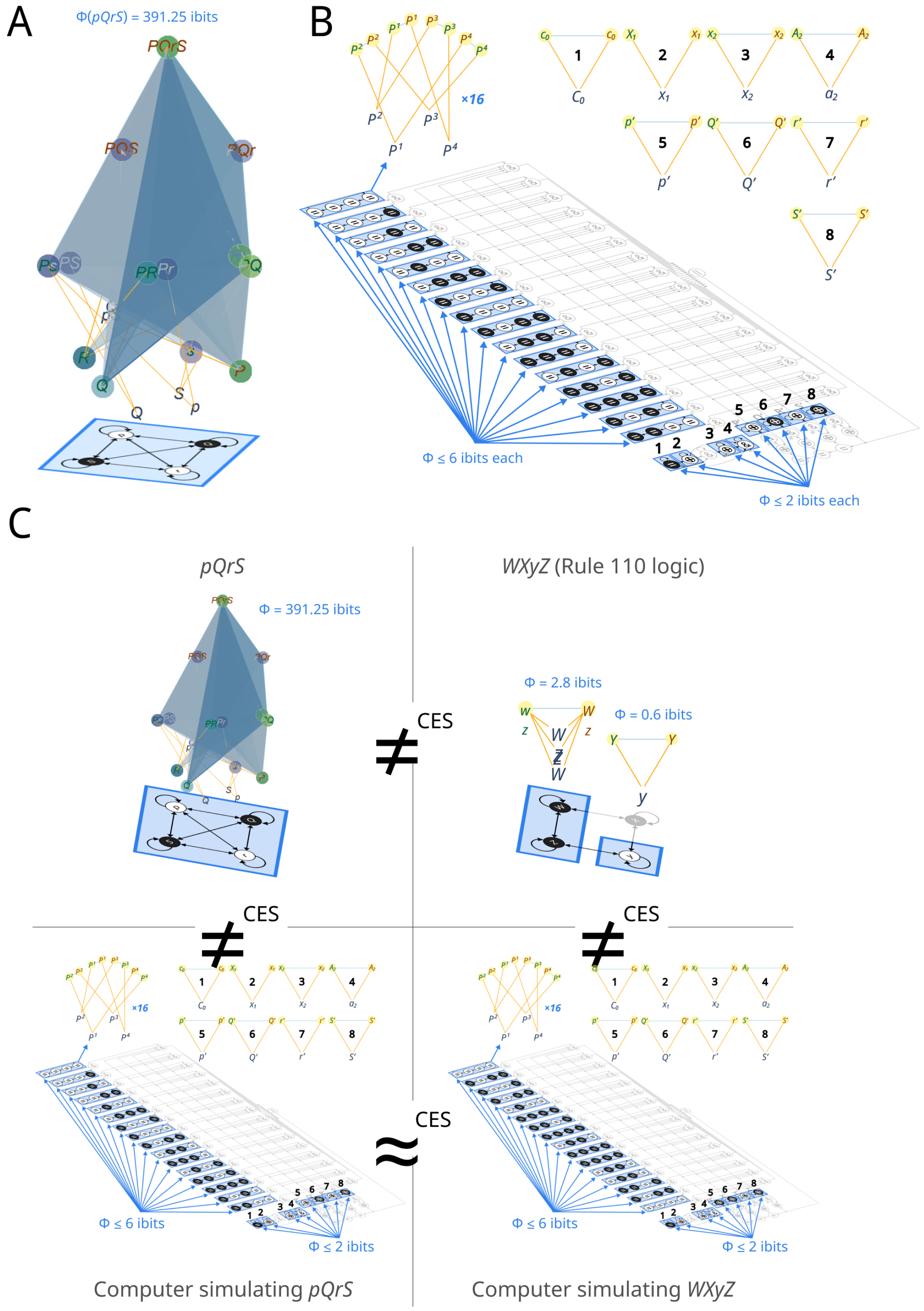

Figure 4: Adapted from [21]. (A) The cause-effect structure of the target system. The four units for a complex with many distinctions and relations among them, with a high (relative to its size) Φ value. (B) The computer as a whole is not integrated, with $\varphi_s = 0$. Instead, the computer decomposes into 24 smaller complexes, each with few distinctions, few relations, and small Φ values (each less than or equal to 6). Thus, the functional equivalence of the target system and computer does not imply phenomenal equivalence according to IIT. (C) The function and cause-effect structure can be dissociated. If we change the target system (e.g., WXYZ implementing Rule 110 cellular automata), the cause-effect structure changes substantially, while the computers cause-effect structures remain simple and stereotyped. Thus, according to IIT, function can be dissociated from experience (though there are reasons to think they go together in humans due to evolutionary pressure [43])

These results are not limited to the particular computer we analyzed. In principle, we can extend the simulated computer to arbitrary size and complexity, allowing it to implement increasingly sophisticated functions. This demonstrates that the observed differences are not a limitation of the functional repertoire: the computer architecture can, in theory, simulate any target system with arbitrary precision. Yet, despite this growing functional complexity, its intrinsic cause–effect structure remains simple and stereotyped. The architectural features that fragment causal interactions persist regardless of scale, preventing the emergence of a high-Φ complex.

The results above demonstrate that, according to IIT, functional equivalence does not imply phenomenal equivalence. Two systems can implement identical input–output functions while exhibiting fundamentally different cause–effect structures, as shown by the double dissociation between target systems and their digital simulations. Consequently, standard digital computer architectures do not support a high-Φ complex and therefore do not give rise to consciousness, regardless of their functional sophistication.

However, this conclusion does not rule out the possibility of machine consciousness altogether. It applies specifically to classical, von Neumann–style architectures, whose implementation is modular. Other, non-standard architectures, such as neuromorphic systems that more closely mirror the substrate of biological brains, or even quantum systems with fundamentally different causal properties, may provide richer substrates for integrated information. These alternative architectures offer promising directions for extending this line of investigation.

3 Research Proposal

The goal of this research proposal is to extend our previous work with von Neumann-style digital machines [21] to systematically investigate the relationship between functional and phenomenal equivalence under IIT. Our earlier results show that functional equivalence does not imply phenomenal equivalence and that standard digital architectures do not give rise to consciousness in virtue of what they do (their functions). However, this does not preclude the possibility that other kinds of computational architectures, whether functionally equivalent to biological systems or not, could support consciousness in virtue of what they are (their cause-effect power). We therefore aim to evaluate alternative architectures, such as neuromorphic and quantum systems, to assess whether and under what conditions they may support complexes with high-Φ cause–effect structures.

Our approach will be to design alternative computational architectures that replicate the behavior of a simple target system, mirroring our previous work with von Neumann–style machines. These architectures will then be represented within IIT's probabilistic framework and analyzed using IIT's mathematical formalism. For each candidate architecture, we will compute φ_s to identify complexes and then unfold their cause–effect structures. If an architecture supports a complex, we will compare its Φ value and cause–effect structure directly to those of the target system to assess the extent to which functional equivalence implies phenomenal equivalence. Particular attention will be given to macro-grain systems. This systematic evaluation will allow us to determine whether, and at what grain, these systems can instantiate high-Φ complexes.

One of the key features of von Neumann architectures is the presence of bottlenecks—centralized pathways or components through which causal interactions are funneled. This architectural property not only precludes the emergence of high-Φ complexes, as shown in our previous work, but also makes the system more tractable for φ_s analysis. These structural constraints allow the use of computational shortcuts and optimizations to complete the analysis efficiently. In contrast, neuromorphic architectures, for example, lack such bottlenecks. Their distributed, parallel causal organization is more conducive to supporting integrated information but also demands the full computational burden of φ_s and Φ calculations. Addressing this challenge is a crucial step toward applying IIT to richer, more biologically plausible machine architectures.

3.1 Estimating Integrated Information

There are two main challenges in computing φ_s for large systems. The first is obtaining the transition probability matrix (TPM), which ideally requires establishing a causal TPM through intervention and observation. The second is computing φ_s given the TPM, which involves several combinatorial searches: identifying the specified cause–effect state, determining the minimum information partition, and identifying the units that maximize φ_s (including their grain). For artificial systems, unlike a biological brain, the first challenge is not a major obstacle, since the TPM can be derived directly from the system's architecture and update rules. Our focus, therefore, is on the second challenge: efficiently computing φ_s for complex systems with arbitrary connectivity structure.

We propose two complementary strategies to improve the estimation of φ_s for such systems. The first is to develop and apply approximation techniques that reduce the combinatorial search space while preserving accuracy. For example, rather than evaluating all possible system states, we can focus on those most likely to yield maximal φ_s. Similarly, instead of exhaustively testing all partitions, graph-theoretic heuristics can be used to identify a reduced and more informative subset of candidate partitions. While similar approaches have been explored previously, a key missing element has been a systematic quantification of the error incurred by these approximations.

The second strategy is to employ statistical inference to estimate φ_s with associated measures of uncertainty. For example, we can construct confidence intervals (e.g., 95% CI for φ_s) or formal hypothesis tests (e.g., $H_0 : \varphi_s = 0$). This requires defining a probability model over states and partitions, which can be induced through random sampling of potential configurations. Alternative sampling protocols, such as cluster or stratified sampling based on network features, will be explored to improve statistical power and efficiency.

Several heuristic measures of integrated information have been proposed, including empirical phi (Φ_E, [44]), decoding phi ($\Phi*$, [45]), and geometric phi (Φ_G, [46]). These measures aim to capture the core intuition of integrated information and exhibit useful mathematical properties, such as positivity and vanishing values for disconnected systems. However, they have not been systematically validated against the theoretically defined value of φ_s (but see [47]). As a result, while these heuristics can provide practical approximations, they cannot be reliably used to test the theory itself.

To validate and calibrate the methods developed in this project, we will conduct a simulation study in cases where φ_s can be computed exactly without approximation. This will involve systematically analyzing a range of network motifs with increasing complexity and comparing the approximated φ_s values to their exact counterparts. By repeating this procedure across network sizes and topologies, we will characterize how approximation error behaves as a function of system size and structure. This will allow us to extrapolate error bounds and uncertainty estimates for larger networks, where exact computation is infeasible. This validation step is essential to ensure that the proposed methods remain both accurate and interpretable as we scale to more complex architectures.

3.2 Applying IIT to different computer architectures

The second theme of this proposal is to extend the work of [21] by applying the IIT framework to models of alternative computer architectures. Building on our previous demonstration of the dissociation between functional and phenomenal equivalence in von Neumann–style systems, we will systematically evaluate whether and under what conditions other computational architectures can support high-Φ complexes with rich cause–effect structures. This analysis will provide a principled, theory-driven way to assess the potential for different machine architectures to give rise to consciousness.

As a first step in evaluating alternative architectures, we will develop a model of a neuromorphic computer. Neuromorphic systems are specifically designed to mimic key organizational principles of biological nervous systems (though currently at a substantially smaller scale), including parallel distributed processing, event-based communication, and local adaptation. Importantly, unlike large-scale biological brains, the dynamics of these architectures can be fully specified and controlled, making it feasible to derive their transition probability matrices directly from their implementation. This tractability enables the application of IIT's mathematical framework to neuromorphic models at scales that would be infeasible for biological systems.

While analyzing a neuromorphic computer's intrinsic cause–effect structure is informative in its own right, an even more revealing test is to evaluate it while it simulates another system. By programming the neuromorphic architecture to reproduce the functional behavior of a target system, we can examine whether—and in what way—its cause–effect structure mirrors that of the simuland. Platforms such as EBRAINS BrainScaleS infrastructure provide a practical means to implement such simulations with high control and scalability [48]. This approach allows us to probe the extent to which functional replication on a non–von Neumann substrate can give rise to similar or divergent CES profiles, directly addressing the relationship between functional and phenomenal equivalence in alternative architectures.

An important extension of this approach is to introduce fundamental indeterminism into both the simuland and the neuromorphic computer. In IIT, indeterminism is essential for a system to self-define a repertoire of alternative states, which underlies its intrinsic cause–effect power [25]. By embedding controlled stochasticity in the target system and tuning the computer's dynamics so that its macro-grain behavior mirrors that of the simuland, we can investigate how this probabilistic structure shapes the CES. Aligning the macro-level dynamics of the simulator and simuland will allow us to test whether, and to what extent, (stochastic) functional equivalence implies phenomenal equivalence under IIT.

The first step in the analysis will be to identify the complexes within the neuromorphic model and the target system. This involves evaluating φ_s across units and grains to identify non-overlapping subsets that maximize integrated information. To accomplish this efficiently, we will leverage the computational methods and approximation techniques developed in Part 1 of the project, enabling us to scale the analysis to architectures with complex and distributed connectivity. Doing so will reveal whether the neuromorphic computer supports a single large high-Φ complex or, like the von Neumann architecture, decomposes into many small low-Φ complexes.

Once the complexes are identified, we will unfold their cause–effect structures and compare them to the CES of the target system. Several outcomes are possible: the neuromorphic system may yield no meaningful CES (indicating low intrinsic cause–effect power), a rich CES that is unrelated to the function being simulated, or a rich CES that is functionally aligned with the simuland. This analysis will provide direct evidence on whether functional replication on a neuromorphic substrate can give rise to intrinsic causal structures that resemble those of the target system.

To meaningfully compare the cause–effect structures of two systems, we require a measure that is independent of the particular labels assigned to their units. Relabeling the units can be understood as a "rotation" of the cause–effect structure, and any measure of structural difference should be invariant to such rotations. We will develop methods for quantifying the distance between cause–effect structures in a way that respects this invariance. Conceptually, this can be achieved by considering all possible rotations and selecting the one that minimizes the distance between structures. In practice, this introduces a new computational problem, for which we will design efficient strategies and approximations.

4 Discussion

The question of machine consciousness is both conceptually challenging and increasingly urgent [13]. Computational functionalism—the view that implementing the right functional organization is sufficient for consciousness—remains the dominant but largely unexamined assumption in contemporary discourse. As AI capabilities advance rapidly, this assumption carries growing ethical, scientific, and societal implications. If we are to address these questions responsibly, we need principled ways to assess whether artificial systems can support conscious experience, and under what conditions.

In this proposal, we outline a research program that develops and applies new computational tools for IIT. Specifically, we aim to make φ_s tractable for large, complex systems and to apply these methods to alternative computer architectures. This integrated approach directly addresses the central question of the prize: whether, and under what conditions, artificial systems can support consciousness. By combining methodological advances with theory-driven application, the project provides a concrete and principled path toward evaluating machine consciousness.

4.1 Alignment with prize scope

IIT is inherently interdisciplinary, integrating insights and methods across multiple domains. From philosophy, it draws a consciousness-first perspective, grounding the theory in phenomenal axioms and proposing an explanatory identity between subjective experience and physical cause–effect structure [19, 20]. From neuroscience, it connects these principles to empirical data, offering explanations and testable predictions about the neural basis of consciousness [33, 32]. From mathematics, IIT is built on a rigorous formal framework that combines probability theory, causal modeling, and information theory [36, 22, 49]. It also engages with physics through its operational definition of physicality and with computer science through the development of algorithms and computational methods for analyzing complex systems [50]. This breadth positions IIT uniquely well to address the interdisciplinary goals at the core of the prize.

The personal and private nature of consciousness makes it a uniquely challenging phenomenon to approach with scientific rigor [4]. IIT directly addresses this challenge through a structured research program ([20], Figure 1). As a consciousness-first theory, it begins with the essential properties of subjective experience, which are then mapped onto corresponding physical postulates (Figure 2). These mappings are empirically tested in the only case where consciousness is certain—humans—allowing the theory to be validated against neuroscientific evidence. Once validated, the framework can be extrapolated to more difficult cases, including non-human animals and artificial systems, providing a principled basis for approaching machine consciousness.

A key strength of IIT is its testability. Unlike many approaches to consciousness, IIT makes clear and often counterintuitive predictions about the relationship between physical mechanisms and subjective experience [33, 20, 32]. By designing and performing experiments to test these predictions, we can both refine the theory and build confidence in its inferences to the most challenging cases, including machine consciousness. However, there are significant experimental and computational bottlenecks, especially when applying IIT to large and complex systems. Part 1 of this proposal directly addresses the computational challenge by developing scalable methods for estimating φ_s. Overcoming this limitation is crucial to extending IIT analyses to artificial architectures, where system size and connectivity complexity would otherwise make such evaluations intractable. Importantly, these same methods will also enhance the theory's applicability to the human brain, enabling more comprehensive tests of IIT.

IIT comes with a distinctive ontology of the physical, grounded not in spacetime but in operational cause–effect power—defined in terms of how a system can be manipulated and how it constrains its own possible states [19]. Notably, IIT does not presuppose space or time as fundamental primitives; instead, these are expected to emerge from regularities in cause–effect structure. This leads to a fundamentally nonlocal view of physical reality: elements of a cause–effect structure are not required to obey spatial locality, and the physical substrate of consciousness may not align with spatial boundaries in any simple way. This perspective has deep implications for how we conceptualize consciousness in both biological and artificial systems, especially when considering architectures that do not map neatly onto physical space.

A central aspect of Integrated Information Theory is intrinsicality: a system specifies for itself a repertoire of alternative states, which forms the basis of its intrinsic cause–effect power. Recent theoretical developments emphasize that this repertoire is not incidental but fundamental to the system's existence from the perspective of IIT [25]. This can be naturally interpreted as a form of indeterminism—not as an additional ingredient layered onto an otherwise deterministic world, but as a basic feature of physical reality itself.

This view resonates with aspects of quantum mechanics, where the wave function provides a structured repertoire of alternative outcomes prior to measurement. While IIT is not explicitly classical or quantum (it is cause-effect power), the conceptual parallels suggest potential synergies: quantum systems, by their very nature, possess intrinsic repertoires, and quantum entanglement could add another form of integrated cause-effect power [51]. This makes quantum computing architectures a natural future extension of the present proposal, complementing the neuromorphic systems explored here. Some initial work has already been done to explore the application of IIT to quantum systems [51].

4.2 Ethical Considerations

The computational functionalist view remains widespread, and under this assumption, conscious AI is often regarded as not only possible but imminent. As AI capabilities continue to advance, the ethical questions surrounding machine consciousness are becoming increasingly urgent. If artificial systems are—or could become—conscious, we must grapple with whether they should be accorded moral consideration or even rights. At the same time, practical constraints, including the limits of physical and economic resources, raise difficult questions: what happens if a proliferation of conscious entities demands protection, care, or recognition at scale?

Beyond the binary question of whether or not AI systems are conscious, the quality of experience becomes ethically decisive. How an entity is conscious—what it is like to be that system—matters deeply for determining moral status. If artificial systems were capable of pain, joy, or loss, their treatment would need to be evaluated through the same ethical lenses currently applied to nonhuman animals. Addressing these questions responsibly requires rigorous, theory-driven approaches to assessing consciousness. IIT is one of the few existing scientific frameworks that attempts to explain not only the presence of consciousness but also its quality, making it particularly well suited to inform this emerging ethical landscape.

References

[1] Plato. *Republic.* Hackett Publishing, 2004. Original work ca. 375 BCE.

[2] Aristotle. *De Anima.* Oxford University Press, 2016.

[3] Daniel C. Dennett. *Consciousness Explained.* Little, Brown and Company, 1991.

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

[5] René Descartes. *Meditations on First Philosophy.* Cambridge University Press, 1996. Original work 1641.

[6] William James. *The Varieties of Religious Experience: A Study in Human Nature.* Longmans, Green, and Co., 1902.

[7] Thomas Nagel. What is it like to be a bat? *The Philosophical Review*, 83(4):435–450, 1974. doi: 10.2307/2183914.

[8] Gottfried Wilhelm Leibniz. *Philosophical Essays.* Hackett Publishing, 1989.

[9] Francis Crick and Christof Koch. Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences*, 2:263–275, 1990.

[10] Hilary Putnam. Minds and machines. In Sidney Hook, editor, *Dimensions of Mind*, pages 138–164. New York University Press, 1960.

[11] Alan M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950. doi: 10.1093/mind/LIX.236.433.

[12] Yoshua Bengio, Geoffrey Hinton, Angela Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Sören Mindermann, et al. Managing extreme ai risks amid rapid progress. *Science*, 384(6698):842–845, 2024. doi: 10.1126/science.adn0117.

[13] Giulio Tononi, Larissa Albantakis, Leandro S. Barbosa, Melanie Boly, Chiara Cirelli, Riccardo Comolatti, Glen Findlay, Michele Grasso, Aaron Haun, Christof Koch, William Marshall, William G. P. Mayner, and Amir Zaeemzadeh. Consciousness or pseudo-consciousness? a clash of two paradigms. *Nature Neuroscience*, 28(10):1811–1813, 2025. doi: 10.1038/s41593-025-01774-9.

[14] Thilo Hagendorff. The ethics of ai ethics: An evaluation of guidelines. *Minds and Machines*, 30 (1):99–120, 2020. doi: 10.1007/s11023-020-09517-8.

[15] Anil K. Seth. Conscious artificial intelligence and biological naturalism. *Behavioral and Brain Sciences*, pages 1–42, 2025. doi: 10.1017/S0140525X25000032.

[16] Giulio Tononi. An information integration theory of consciousness. *BMC Neuroscience*, 5:42, 2004. doi: 10.1186/1471-2202-5-42.

[17] David Balduzzi and Giulio Tononi. Integrated information in discrete dynamical systems: Motivation and theoretical framework. *PLoS Computational Biology*, 4(6):e1000091, 2008. doi: 10.1371/journal.pcbi.1000091.

[18] Masafumi Oizumi, Larissa Albantakis, and Giulio Tononi. From the phenomenology to the mechanisms of consciousness: Integrated information theory 3.0. *PLoS Computational Biology*, 10(5):e1003588, 2014. doi: 10.1371/journal.pcbi.1003588.

[19] Larissa Albantakis, Leandro S. Barbosa, Glen Findlay, Aaron Haun, William Marshall, William G. P. Mayner, Giulio Tononi, et al. Integrated information theory: From consciousness to physical existence, 2023. URL `https://arxiv.org/abs/2306.00014`.

[20] Jeremiah Hendren, Matteo Grasso, Bjørn E. Juel, and Giulio Tononi. Integrated information theory wiki. `https://www.iit.wiki`, 2024.

[21] Glen Findlay, Larissa Albantakis, Michele Grasso, Aaron Haun, and Giulio Tononi. Dissociating artificial intelligence from artificial consciousness, 2025. URL `https://arxiv.org/abs/2412.04571`.

[22] Leandro S. Barbosa, William Marshall, Sebastian Streipert, Larissa Albantakis, and Giulio Tononi. A measure for intrinsic information. *Scientific Reports*, 10:18803, 2020. doi: 10.1038/s41598-020-75944-5.

[23] Leandro S. Barbosa, William Marshall, Larissa Albantakis, and Giulio Tononi. Mechanism integrated information. *Entropy*, 23(3):362, 2021. doi: 10.3390/e23030362.

[24] William Marshall, Michele Grasso, William G. P. Mayner, Amir Zaeemzadeh, Leandro S. Barbosa, Eric Chastain, Glen Findlay, S. Sasai, Larissa Albantakis, and Giulio Tononi. System integrated information. *Entropy*, 25(2):334, 2023. doi: 10.3390/e25020334.

[25] William G. P. Mayner, William Marshall, Larissa Albantakis, and Giulio Tononi. Intrinsic cause–effect power: The trade-off between specificity and invariance, 2025. URL `https://arxiv.org/abs/2510.03881`.

[26] Erik P. Hoel, Larissa Albantakis, and Giulio Tononi. Quantifying causal emergence shows that macro can beat micro. *Proceedings of the National Academy of Sciences*, 110(49):19790–19795, 2013. doi: 10.1073/pnas.1314922110.

[27] Erik P. Hoel, Larissa Albantakis, William Marshall, and Giulio Tononi. Can the macro beat the micro? integrated information across spatiotemporal scales. *Neuroscience of Consciousness*, 2016(1):niw012, 2016. doi: 10.1093/nc/niw012.

[28] William Marshall, Larissa Albantakis, and Giulio Tononi. Black-boxing and cause–effect power. *PLOS Computational Biology*, 14(4):e1006114, 2018. doi: 10.1371/journal.pcbi.1006114.

[29] William Marshall, William G. P. Mayner, Larissa Albantakis, and Giulio Tononi. Intrinsic units: Identifying a system's causal grain. bioRxiv, 2024.

[30] Aaron Haun and Giulio Tononi. Why does space feel the way it does? towards a principled account of spatial experience. *Entropy*, 21(12):1160, 2019. doi: 10.3390/e21121160.

[31] Michele Grasso, Riccardo Comolatti, and Giulio Tononi. Why does time feel the way it does? towards a principled account of temporal experience. *iScience*, 28(10):113434, 2025. doi: 10.1016/j.isci.2025.113434.

[32] Giulio Tononi and Melanie Boly. *Integrated Information Theory.* Forthcoming book chapter.

[33] Giulio Tononi, Melanie Boly, Marcello Massimini, and Christof Koch. Integrated information theory: From consciousness to its physical substrate? *Nature Reviews Neuroscience*, 17(7): 450–461, 2016. doi: 10.1038/nrn.2016.44.

[34] Cogitate Consortium, O. Ferrante, U. Gorska-Klimowska, S. Henin, R. Hirschhorn, L. Melloni, et al. Adversarial testing of global neuronal workspace and integrated information theories of consciousness. *Nature*, 642:133–142, 2025. doi: 10.1038/s41586-025-08888-1.

[35] ARC-INTREPID Consortium. Intrepid: An adversarial collaboration to test iit vs predictive processing. `https://arc-intrepid.com/`. Accessed October 13, 2025.

[36] William Marshall, Jorge Gomez-Ramirez, and Giulio Tononi. Integrated information and state differentiation. *Frontiers in Psychology*, 7:926, 2016. doi: 10.3389/fpsyg.2016.00926.

[37] Hal Blumenfeld. Impaired consciousness in epilepsy. *The Lancet Neurology*, 11(9):814–826, 2012.

[38] Giulio Tononi, Melanie Boly, and Chiara Cirelli. Consciousness and sleep. *Neuron*, 112(10): 1568–1594, 2024. doi: 10.1016/j.neuron.2024.04.011.

[39] Andrea Pigorini, Simone Sarasso, Paola Proserpio, Caroline Szymanski, Gabriele Arnulfo, Silvia Casarotto, Matteo Fecchio, Mario Rosanova, Maurizio Mariotti, Giorgio Lo Russo, et al. Bistability breaks-off deterministic responses to intracortical stimulation during non-rem sleep. *Neuroimage*, 112:105–113, 2015.

[40] Marcello Massimini, Fabio Ferrarelli, Reto Huber, Steve K Esser, Harpreet Singh, and Giulio Tononi. Breakdown of cortical effective connectivity during sleep. *Science*, 309(5744):2228–2232, 2005.

[41] Adenauer G Casali, Olivia Gosseries, Mario Rosanova, Mélanie Boly, Simone Sarasso, Karina R Casali, Silvia Casarotto, Marie-Aurélie Bruno, Steven Laureys, Giulio Tononi, et al. A theoretically based index of consciousness independent of sensory processing and behavior. *Science translational medicine*, 5(198):198ra105–198ra105, 2013.

[42] Silvia Casarotto, Angela Comanducci, Mario Rosanova, Simone Sarasso, Matteo Fecchio, Martino Napolitani, Andrea Pigorini, Adenauer G. Casali, Pietro D Trimarchi, Melanie Boly, et al. Stratification of unresponsive patients by an independently validated index of brain complexity. *Annals of neurology*, 80(5):718–729, 2016.

[43] Larissa Albantakis, Arend Hintze, Christof Koch, Christoph Adami, and Giulio Tononi. Evolution of integrated causal structures in animats exposed to environments of increasing complexity. *PLoS computational biology*, 10(12):e1003966, 2014.

[44] Adam B. Barrett and Anil K. Seth. Practical measures of integrated information for time-series data. *PLoS Computational Biology*, 7(1):e1001052, 2011. doi: 10.1371/journal.pcbi.1001052.

[45] Masafumi Oizumi, Shun-ichi Amari, Takuya Yanagawa, Naotaka Fujii, and Naotsugu Tsuchiya. Measuring integrated information from the decoding perspective. *PLoS Computational Biology*, 12(1):e1004654, 2016. doi: 10.1371/journal.pcbi.1004654.

[46] Masafumi Oizumi, Naotsugu Tsuchiya, and Shun-ichi Amari. A unified framework for information integration based on information geometry. *Proceedings of the National Academy of Sciences*, 113(51):14817–14822, 2016. doi: 10.1073/pnas.1603583113.

[47] André Sevenius Nilsen, Bjørn Erik Juel, and William Marshall. Evaluating approximations and heuristic measures of integrated information. *Entropy*, 21(5):525, 2019. doi: 10.3390/e21050525.

[48] Katrin Amunts, Javier DeFelipe, Cyriel Pennartz, Alain Destexhe, Michele Migliore, Philippe Ryvlin, and Viktor Jirsa. Linking brain structure, activity, and cognitive function through computation. *eNeuro*, 9(2):ENEURO.0316–21.2022, 2022. doi: 10.1523/ENEURO.0316-21.2022.

[49] Amir Zaeemzadeh and Giulio Tononi. Upper bounds for integrated information. *PLOS Computational Biology*, 20(8):e1012323, 2024. doi: 10.1371/journal.pcbi.1012323.

[50] William G. P. Mayner, William Marshall, Larissa Albantakis, Graham Findlay, Robert Marchman, and Giulio Tononi. Pyphi: A toolbox for integrated information theory. *PLoS Computational Biology*, 14(7):e1006343, 2018. doi: 10.1371/journal.pcbi.1006343.

[51] Larissa Albantakis, Robert Prentner, and Ian T. Durham. Computing the integrated information of a quantum mechanism. *Entropy*, 25(3):449, 2023. doi: 10.3390/e25030449.