Consciousness, natural and artificial: an evolutionary advantage for reasoning on reactive substrates

Warisa Sritriratanarak[1] **and Paulo Garcia**[2]

[1]Department of Computer Engineering, Chulalongkorn University, Bangkok, Thailand

[2]International School of Engineering, Chulalongkorn University, Bangkok, Thailand

[1]warisa.s@chula.ac.th, [2]paulo.g@chula.ac.th

ABSTRACT

Precisely defining consciousness and identifying the mechanisms that effect it is a long-standing question, particularly relevant with advances in artificial intelligence. The scientific community is divided between physicalism and natural dualism. Physicalism posits consciousness is a physical process that can be modeled computationally; natural dualism rejects this hypothesis. Finding a computational model has proven elusive, particularly because of conflation of consciousness with other cognitive capabilities exhibited by humans, such as intelligence and physiological sensations. Here we show such a computational model that precisely models consciousness, natural or artificial, identifying the structural and functional mechanisms that effect it, confirming the physicalism hypothesis. We found such a model is obtainable when including the underlying (biological or digital) substrate and accounting for reactive behavior in substrate sub-systems (e.g., autonomous physiological responses). Results show that, unlike all other computational processes, consciousness is not independent of its substrate and possessing it is an evolutionary advantage for intelligent entities. Our result shows there is no impediment to the realization of fully artificial consciousness but, surprisingly, that it is also possible to realize artificial intelligence of arbitrary level without consciousness whatsoever, and that there is no advantage in imbuing artificial systems with consciousness.

The question of whether or not artificial systems can exhibit consciousness[1] is of critical importance as we consider the ethical and moral implications of ever more powerful Artificial Intelligence (AI) going forward[2], and it addresses a fundamental question about the nature of human beings. There is no agreed-upon definition for what constitutes "consciousness", besides the general agreement that human beings exhibit it[3], and its definition is intertwined with several adjacent concepts, often in circular references (e.g., sentience, awareness, qualia, experience, etc.)[4]. Beyond linguistic ambiguity, there is also an overlap between perceptions of consciousness and other concepts (e.g., wakefulness, intelligence)[5], and cultural and religious overhead that contributes to ambiguity[6].

This inexact definition has challenged the design of consciousness tests[7] and, sometimes implicitly, prevents meaningful comparisons of different theories of consciousness[8]. Broadly, consciousness has been divided into phenomenal and cognitive[9], and its study approached from the perspective of natural dualism (Chalmers' "hard problem", positing consciousness is beyond the reach of contemporary science[10]) and from the perspective of physicalism (sometimes referred to as materialism, and more formally as computational functionalism, which posits consciousness is ultimately a computational process that can be modeled and understood[11]). One the main challenges is the conflation of consciousness and other cognitive functions (namely, intelligence), a topic explored in detail by Minsky[6,12,13]; we point interested readers to the extensive review by Kuhn[8] for a much more in-depth description and to the work of Butlin et al[11] for a clear distinction between intelligence and consciousness in the context of AI, which we adopt here.

To avoid ambiguity as much as possible, we will refrain from using the terms described above, and instead phrase the problem as: **awareness of the phenomenal experience**: "how can an entity have an awareness reality, a sense of "*here and now*", grounded in its place relative to the world around it[14]?"; and **subjective experience**: "how can two distinct entities, equipped with the same cognitive substrate, experience the same stimuli in subjectively different ways (i.e., different qualia)?". We are still using overloaded terms (e.g., to experience), but we will make these terms clearer. Answering these two questions is the primary goal of the quest for "consciousness", as all other aspects attributed to "consciousness" can be explained by intelligence, sensing, and perception. Our approach is based on the fact that "consciousness", in humans, emerged from an

evolutionary process and is thus intimately linked to other human traits that evolved before and alongside it[15].

We show that, if we accept three axioms for which there is empirical support, we can obtain a computational model that explains both awareness of the phenomenal experience and subjective experience. We will see that subjective experience relies on awareness of the phenomenal experience, and that this awareness is an evolutionarily required property for intelligence to operate, given that intelligence is being executed on a physical substrate concurrently with processes that have evolved before, and are often at odds with, intelligence. This hypothesis is a simpler explanation than contemporary theories, lends itself to empirical validation, and supports the physicalism theory. A consequence of this result is that we show that there is no barrier to the development of AIs that are as conscious as humans; however, another consequence is that there is no need for conscious AIs to ever be built, as there is nothing to be gained from it.

The remainder of this paper is organized as follows: Section 2 ("Models of reasoning and substrate machinery") describes cognitive sub-systems that effect reasoning, both in human beings and AI systems, and highlights how autonomous responses that evolutionarily predate reasoning occur in function of these subsystems. Section 3 ("Consciousness as an evolutionary advantage for reasoning entities") formalizes a reasoning process computationally. Unlike traditional formulations of models of computation, this formulation is not substrate-independent; here, we show how that dependence, when coupled with the aforementioned autonomous responses, gives rise to the need for consciousness, and that consciousness can be interpreted as a boolean switch in substrate machinery. Section 4 ("Analysis") re-states the reasoning in Section 3, but without the mathematical formalisms, further describing the evolutionary need for such a switch. Section 5 ("Discussion") contextualizes our findings, particularly in relation to other views on consciousness. Section 6 ("Conclusions") offers our concluding remarks.

Models of reasoning and substrate machinery

Intelligence is another ill-defined term: here, we provide a working definition of "reasoning", one of the aspects that make up intelligence, and its formulation will suffice for the rest of this paper. Reasoning is the iterative process of: obtaining new data from the surrounding world through a *sensing* function; utilizing that sensed data and known information (in the form of propositions, postulates, predicates, probabilities, etc.; generally, *percepts*) in long-term and working memory (in AI, often referred to as a *knowledge base (KB)* to generate new information that updates memory; i.e., a *perception* function; utilizing all information obtained thus far to decide on an action to perform (a *decision* process). *Sensing* is straightforward: given available sensors, be them natural or artificial, (audio, visual, etc.) data is attained. *Perception* transforms raw sensor data into higher-level concepts, represented symbolically, and imbues those concepts with semantic information by relating them to other concepts. For example, while sensing may obtain an image of a cat, perception associates that image with the symbols for the class "cat", its linguistic representation (the word "cat"), the properties one may associate with cats, connections to a specific instance of the cat class ("a cat I know"), relations with other concepts ("my neighbor's cat"), etc.

Decision can be modeled in myriad ways: one of the classic formulations is that of state-space exploration[16]. Having (approximately) inferred the state of the world through an inference function post perception and knowing the effect each action at its disposal will have on that state (deterministically or probabilistically), an entity can construct a decision tree[17] to examine the possible futures within its reach, given its reasoning horizon (how many steps in the future it can predict in useful time). If equipped with a measure of how "good" each possible future state is (usually referred to as a *utility function*), the entity can then decide on which action to perform to maximize its utility (generally by choosing the action with highest expected utility). This formulation of a reasoning entity, dating back to the 1960s[18], is based on cognitive science that explored decision-making in humans, under the premise that "cognition is computation of representations"[19]. Informally, it corresponds to "what if" scenarios[20]; scenarios we envision to assess the likely outcome of our actions. When examining these scenarios, we are imagining sensory input (what we will see and hear; whether it will hurt or not) and imagining perception on that input, and that fictional data and percepts are constructed by accessing our memory from past sensory input and using prior information to construct hypothetical novel one[21–23].

Both biological and digital substrates are finite in size, and both are made of interconnected sub-systems that are specialized for distinct tasks[24,25]. Thus, the use of the same sub-system (e.g., the visual cortex in humans[26,27], graphics cards on computers[28]) by different components of the reasoning process must be multiplexed in time. When biological substrates process the recall of images from memory, the visual cortex is engaged[21,22] (albeit with different neuronal firing patterns than those observed when processing stimuli from the senses in real time, and affecting sensing[29]). Thus, all substrates must necessarily perform some resource allocation[30], as they are bound in *structure* and *function*[31]. When asked to recall an image as vividly as possible, most participants will close their eyes to prevent real-time sensory stimuli from interfering with the visual processing from memory in the visual cortex[23]. This implies sub-systems in our biological substrate used for real-time sensory processing (e.g., the visual cortex) are necessarily employed to process hypothetical sensory input when reasoning[21–23].

A final property of biological substrates is relevant: sub-systems are linked to autonomous neurological responses[32], and these can be triggered at the sensing or perception level, before that percept is forwarded to higher-level reasoning[33], eliciting responses in, e.g., the hippocampus and amygdala, in the case of fear[34]. These stimulus-response autonomous properties[35] are

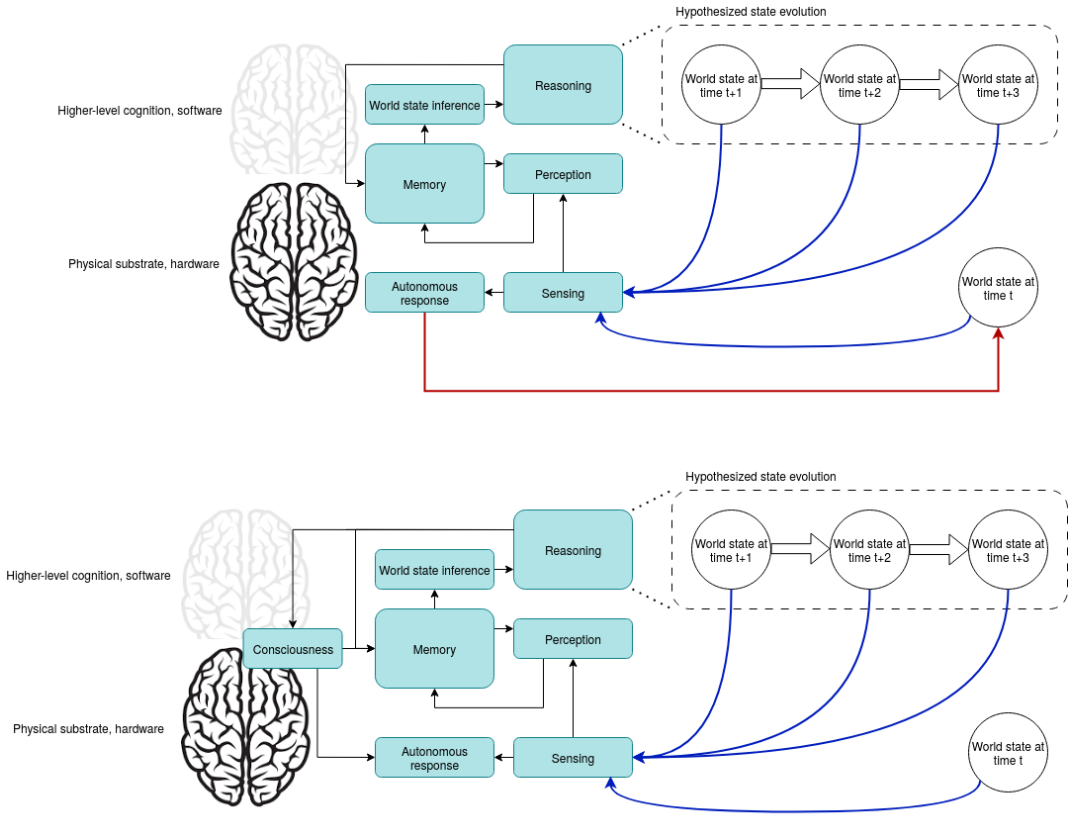

Figure 1. Conceptual view of consciousness as a control switch for autonomous responses on reactive substrates. Top: non-conscious reasoning entity. Sensory sub-system re-use for processing hypothetical scenarios may trigger autonomous responses on real world state. **Bottom:** conscious reasoning entity. Consciousness switch controlled by high-level reasoning distinguishes between real and hypothetical input, allows identification of "here and now". Exhibits Hofstadter's "strange loop"[37]: feedback from higher-level cognition to the underlying physical substrate.

not surprising, considering that higher-level reasoning is much younger, in evolutionary terms, than primitive responses such as the fight-or-flight mechanism[36].

Consciousness as an evolutionary advantage for reasoning entities

We adopt three premises as axioms in the formulation of subjective experience and awareness of the phenomenal experience.

- **Ever-evolving perception**: the perception function evolves over time. The same entity, being exposed to the same stimuli (i.e., sensed data) at different times, may infer different percepts[3,10,19]. This stems from the fact that previously inferred percepts shape the perception function: the same image of a cat will result in the relational percept "my neighbor's cat" only after we have learned our neighbor has a cat (specifically, this cat). In other words, perception is a process exhibiting feedback: a recurrent process. This premise is supported by research on neural plasticity and learning mechanisms[38].
- **Autonomous triggers**: the perception function can initiate substrate responses before resulting percepts are made available to higher-level reasoning. We respond to stimuli before we have time to reason about them. This is an evolutionary response for survival, established when our biological substrates were simple stateless reactive systems (i.e., outputs depended only upon inputs at the current time)[39]. There is extensive neurological research on autonomous responses[36] including pain, fear, arousal, hunger, etc. Furthermore, there does not seem to be a well-defined segregation of these sub-systems in our brains: there is no one "perception" sub-system, but rather an emerging interconnection of components that make up human perception, often tightly coupled with sensing[34].
- **Percept of self**: long-term and working memory (equivalently, KB) includes a percept that represents the substrate sensing, perception, and reasoning, are executing on, and represents the sensing, perception, and reasoning processes.

As we are capable of touching, seeing, hearing, tasting, and smelling our own bodies (i.e., we can *sense* them as much as we can sense anything else), we represent them symbolically, in space and time, in the same way we represent all other objects in the world around us, and we can model that this substrate can execute higher-level reasoning. Notice that we are not trying to axiomatize the conclusion: we are not claiming this constitutes "conscious" self-awareness. This is simply a property of computational recursion: Hofstadter[37] has explored how we perform such recursion extensively.

Let a reasoning entity be modeled by $\frac{R}{P}$, meaning a reasoning process R executing on a physical substrate P.

The physical substrate P is modeled by the tuple $\langle s_t, \sigma, \pi, a^1, a^2, ..., a^i \rangle$, where s_t is the state of the physical substrate at time t, including its spatial position in the world and the values of all its internal properties; σ is the sensing function producing data from world state w_t, such that $\sigma(w_t) = \{d^1, d^2, ..., d^n\}$ ($d^i, i \in [1:n]$ corresponds to one datum); π_t is the perception function at time t, such that $\pi_t(\sigma(w_t), KB_{1:t-1}) = \{p_t^1, p_t^2, ..., p_t^m\}$ ($p_t^i, i \in [1:m]$ corresponds to one percept, and $KB_{1:t-1}$ is the collection of all percepts thus far and rules for their semantic association); and $a^1, a^2, ..., a^i$ are physical actions that change the state of the outside world (one of them being the null action).

The reasoning process R is modeled by the tuple $\langle KB_{1:t-1}, \iota, \tau, d \rangle$, where ι is the world state inference function such that $w\prime_t = \iota(KB_{1:t-1}, \pi_t(\sigma(w_t))), p_t^P \in w\prime_t$ (where $w\prime_t$ is the approximate inference of world state and p_t^P represents the entity in the world, per the axiom of percept of self); τ is the state transition function, such that $\tau(a^j, w_t) = w_{t+1}^j, j \in [1:i]$; ν is the utility function that scores states, such that $\nu(w) = u, u \in \mathbb{R}$; and d is the decision process, corresponding to a tree of world states connected via physical actions $a^1, a^2, ..., a^i$, constructed through the successive application of τ and ν to $w\prime_t$. Notice that, while we are using a state-space exploration formulation of reasoning, this is not a unique option; other formulations (including, but not limited to, statistical inference, inductive and deductive logic, hybrid approaches) are equally valid, and do not affect the main result as long as that formulation includes a world model.

The need for awareness of the phenomenal experience

We will denote the effect of executing f (where f is a function or an action) by $\frac{R}{P} \xrightarrow{f} \frac{R\prime}{P\prime}$. Consider the following process of a reasoning entity *without awareness of the phenomenal experience*, assuming two physical actions and a reasoning horizon of 2. By sensing and perceiving the world, the entity updates its KB with percepts at time t:

$$\frac{\langle KB_{1:t-1}, \iota, \tau, d = \emptyset \rangle}{\langle s_t, \sigma, \pi_t, a^1, a^2 \rangle} \xrightarrow{\pi(\sigma(w_t), KB_{1:t-1})} \frac{\langle \{KB_{1:t-1} \cup p_t^1, p_t^2, ..., p_t^m\}, \iota, \tau, d = \emptyset \rangle}{\langle s_t, \sigma, \pi_t, a^1, a^2 \rangle} \tag{1}$$

Through the world state inference function ι applied to all percepts obtained thus far, the entity constructs an internal model of the world at time t:

$$\frac{\langle KB_{1:t-1} \cup \{p_t^1, p_t^2, ..., p_t^m\}, \iota, \tau, d = \emptyset \rangle}{\langle s_t, \sigma, \pi_t, a^1, a^2 \rangle} \xrightarrow{\iota(KB_{1:t-1} \cup \{p_t^1, p_t^2, ..., p_t^m\})} \frac{\langle KB_{1:t}, \iota, \tau, w\prime_t \rangle}{\langle s_t, \sigma, \pi_t, a^1, a^2 \rangle} \tag{2}$$

By applying the transition function τ to its internal model of the world, given its available actions, the entity begins constructing a tree of future world states (i.e., its prediction of possible worlds at $t+1$):

$$\frac{\langle KB_{1:t}, \iota, \tau, w\prime_t \rangle}{\langle s_t, \sigma, \pi_t, a^1, a^2 \rangle} \xrightarrow{\tau(a^1, w\prime_t), \tau(a^2, w\prime_t)} \frac{\left\langle KB_{1:t}, \iota, \tau, w\prime_t \rightarrow \begin{cases} (a^1), w\prime_{t+1}^1 \\ (a^2), w\prime_{t+1}^2 \end{cases} \right\rangle}{\langle s_t, \sigma, \pi_t, a^1, a^2 \rangle} \tag{3}$$

The utility function ν assigns a value to each hypothetical world state:

$$\frac{\left\langle KB_{1:t}, \iota, \tau, w\prime_t \rightarrow \begin{cases} (a^1), w\prime_{t+1}^1 \\ (a^2), w\prime_{t+1}^2 \end{cases} \right\rangle}{\langle s_t, \sigma, \pi_t, a^1, a^2 \rangle} \xrightarrow{\nu(w^1\prime_t), \nu(w^2\prime_t)} \frac{\left\langle KB_{1:t}, \iota, \tau, w\prime_t \rightarrow \begin{cases} (a^1, x_1), w\prime_{t+1}^1 \\ (a^2, x_2), w\prime_{t+1}^2 \end{cases} \right\rangle}{\langle s_t, \sigma, \pi_t, a^1, a^2 \rangle} \tag{4}$$

assuming $x_1 = \nu(w^1\prime_t)$ and $x_2 = \nu(w^2\prime_t)$. By the axiom of percept of self, these hypothetical world states include a representation of the entity itself; i.e., $w\prime_{t+1}^1 = \{w\prime_{t+1}^1 \setminus p_{t+1}^{P_1}, \frac{\langle KB_{1:t}, \iota, \tau, d=\emptyset \rangle}{\langle s_{t+1}, \sigma, \pi_{t+1}, a^1, a^2 \rangle}\}$:

$$\frac{\left\langle KB_{1:t}, \iota, \tau, w\prime_t \rightarrow \begin{Bmatrix} (a^1,x_1), w\prime^1_{t+1} \\ (a^2,x_2), w\prime^2_{t+1} \end{Bmatrix} \right\rangle}{\langle s_t, \sigma, \pi_t, a^1, a^2 \rangle} = \frac{\left\langle KB_{1:t}, \iota, \tau, w\prime_t \rightarrow \begin{Bmatrix} (a^1,x_1), \{w\prime^1_{t+1} \setminus p^{P_1}_{t+1}, \frac{\langle KB_{1:t}, \iota, \tau, d=\emptyset \rangle}{\langle s_{t+1}, \sigma, \pi_{t+1}, a^1, a^2 \rangle}\} \\ (a^2,x_2), \{w\prime^2_{t+1} \setminus p^{P_2}_{t+1}, \frac{\langle KB_{1:t}, \iota, \tau, d=\emptyset \rangle}{\langle s_{t+1}, \sigma, \pi_{t+1}, a^1, a^2 \rangle}\} \end{Bmatrix} \right\rangle}{\langle s_t, \sigma, \pi_t, a^1, a^2 \rangle} \tag{5}$$

The entity computes a further step into the future by considering what it would reason about, it it were in such a hypothetical world state:

$$\begin{gathered} \frac{\left\langle KB_{1:t}, \iota, \tau, w\prime_t \rightarrow \begin{Bmatrix} \dots \frac{\langle KB_{1:t}, \iota, \tau, d=\emptyset \rangle}{\langle s_{t+1}, \sigma, \pi_{t+1}, a^1, a^2 \rangle}\} \\ \dots \frac{\langle KB_{1:t}, \iota, \tau, d=\emptyset \rangle}{\langle s_{t+1}, \sigma, \pi_{t+1}, a^1, a^2 \rangle}\} \end{Bmatrix} \right\rangle}{\langle s_t, \sigma, \pi_t, a^1, a^2 \rangle} \xrightarrow{\pi(\sigma(w^1\prime_{t+1}), KB_{1:t})} \\ \xrightarrow{\pi(\sigma(w^1\prime_{t+1}), KB_{1:t})} \frac{\left\langle KB_{1:t}, \iota, \tau, w\prime_t \rightarrow \begin{Bmatrix} \dots \frac{\langle \{KB_{1:t} \cup p^1_{t+1}, \dots, p^m_{t+1}\}, \iota, \tau, d=\emptyset \rangle}{\langle s_{t+1}, \sigma, \pi_{t+1}, a^1, a^2 \rangle}\} \\ \dots \frac{\langle KB_{1:t}, \iota, \tau, d=\emptyset \rangle}{\langle s_{t+1}, \sigma, \pi_{t+1}, a^1, a^2 \rangle}\} \end{Bmatrix} \right\rangle}{\langle s_t, \sigma, \pi_t, a^1, a^2 \rangle} \end{gathered} \tag{6}$$

finally obtaining an estimation of the world state after two actions:

$$\begin{gathered} \frac{\left\langle KB_{1:t}, \iota, \tau, w\prime_t \rightarrow \begin{Bmatrix} \dots \frac{\langle \{KB_{1:t} \cup \{p^1_{t+1}, \dots, p^m_{t+1}\}\}, \iota, \tau, d=\emptyset \rangle}{\langle s_{t+1}, \sigma, \pi_{t+1}, a^1, a^2 \rangle}\} \\ \dots \frac{\langle KB_{1:t}, \iota, \tau, d=\emptyset \rangle}{\langle s_{t+1}, \sigma, \pi_{t+1}, a^1, a^2 \rangle}\} \end{Bmatrix} \right\rangle}{\langle s_t, \sigma, \pi_t, a^1, a^2 \rangle} \xrightarrow{\iota(KB_{1:t} \cup \{p^1_{t+1}, \dots, p^m_{t+1}\})} \\ \xrightarrow{\iota(KB_{1:t} \cup \{p^1_{t+1}, \dots, p^m_{t+1}\})} \frac{\left\langle KB_{1:t}, \iota, \tau, w\prime_t \rightarrow \begin{Bmatrix} \dots \frac{\langle \{KB_{1:t+1}, \iota, \tau, d=w\prime^1_{t+1} \rangle}{\langle s_{t+1}, \sigma, \pi_{t+1}, a^1, a^2 \rangle}\} \\ \dots \frac{\langle KB_{1:t}, \iota, \tau, d=\emptyset \rangle}{\langle s_{t+1}, \sigma, \pi_{t+1}, a^1, a^2 \rangle}\} \end{Bmatrix} \right\rangle}{\langle s_t, \sigma, \pi_t, a^1, a^2 \rangle} \end{gathered} \tag{7}$$

where the decision process in $p^{P_1}_{t+1}$ and $p^{P_2}_{t+1}$ must be completed to reach the reasoning depth of 2 across all branches: the action that leads to the state with the highest utility would then be executed. Assuming a_1:

$$\begin{gathered} \frac{\left\langle KB_{1:t}, \iota, \tau, w\prime_t \rightarrow \begin{Bmatrix} (a^1,x_1), \{w\prime^1_{t+1} \setminus p^{P_1}_{t+1}, \frac{\langle KB_{1:t}, \iota, \tau, d=\dots \rangle}{\langle s_{t+1}, \sigma, \pi_{t+1}, a^1, a^2 \rangle}\} \\ (a^2,x_2), \{w\prime^2_{t+1} \setminus p^{P_2}_{t+1}, \frac{\langle KB_{1:t}, \iota, \tau, d=\dots \rangle}{\langle s_{t+1}, \sigma, \pi_{t+1}, a^1, a^2 \rangle}\} \end{Bmatrix} \right\rangle}{\langle s_t, \sigma, \pi_t, a^1, a^2 \rangle} \xrightarrow{a_1} \\ \xrightarrow{a_1} \frac{\langle KB_{1:t}, \iota, \tau, d=\emptyset \rangle}{\langle s_{t+1}, \sigma, \pi_{t+1}, a^1, a^2 \rangle} \end{gathered} \tag{8}$$

We are omitting the evolution of π for now. This is the algorithm for state-space exploration utilized in classic AI agents, with additional modeling of substrate machinery. In the context of AI, this algorithm suffers from scalability issues (combinatorial explosion of hypothetical world states as the reasoning horizon increases), but performs reasonably well (within computation and memory constraints) for small reasoning horizons, when the axiom of autonomous triggers does not apply.

Let us consider the impact of the axiom of autonomous triggers on this formalism. Substrate machinery is used for reasoning, and it may produce autonomous responses: let us denote an autonomous response a_r triggered by an action or function f by $\frac{R}{P} \xrightarrow[a_r]{f} \frac{R\prime}{P\prime}$. Let us examine Equation 6, in a scenario where perception identifies a threat in hypothetical state $w^1\prime_{t+1}$, triggering the fight or flight response, with its associated rise in stress levels, adrenaline production, etc.

$$\begin{gathered} \frac{\left\langle KB_{1:t}, \iota, \tau, w\prime_t \rightarrow \begin{Bmatrix} \dots \frac{\langle KB_{1:t}, \iota, \tau, d=\emptyset \rangle}{\langle s_{t+1}, \sigma, \pi_{t+1}, a^1, a^2 \rangle}\} \\ \dots \frac{\langle KB_{1:t}, \iota, \tau, d=\emptyset \rangle}{\langle s_{t+1}, \sigma, \pi_{t+1}, a^1, a^2 \rangle}\} \end{Bmatrix} \right\rangle}{\langle s_t, \sigma, \pi_t, a^1, a^2 \rangle} \xrightarrow[a_r]{\pi(\sigma(w^1\prime_{t+1}), KB_{1:t})} \\ \xrightarrow[a_r]{\pi(\sigma(w^1\prime_{t+1}), KB_{1:t})} \frac{\left\langle KB_{1:t}, \iota, \tau, w\prime_t \rightarrow \begin{Bmatrix} \dots \frac{\langle \{KB_{1:t} \cup p^1_{t+1}, \dots, p^m_{t+1}\}, \iota, \tau, d=\emptyset \rangle}{\langle s_{t+1}, \sigma, \pi_{t+1}, a^1, a^2 \rangle}\} \\ \dots \frac{\langle KB_{1:t}, \iota, \tau, d=\emptyset \rangle}{\langle s_{t+1}, \sigma, \pi_{t+1}, a^1, a^2 \rangle}\} \end{Bmatrix} \right\rangle}{\langle s\prime_t = a_r(s_t), \sigma, \pi_t, a^1, a^2 \rangle} \end{gathered} \tag{9}$$

This is an effect, on the real substrate state s_t within w_t, of an imagined threat in a hypothetical state $w^1\prime_{t+1}$: damage caused by *thinking*. It is straightforward to imagine many more scenarios where this damage is even move extreme, simply by considering autonomous movement (e.g., recoil as pain response) to hypothetical future states. What we have identified here is a mis-alignment between autonomous responses evolved when we were stateless reactive systems and later capability for reasoning; mis-alignment stemming from using the same substrate resources for different functions.

There is a simple solution to this mis-alignment: a switch in substrate machinery that allows *turning off* (or dampening) autonomous responses; a switch (denoted $\mathbf{c} \in R$) that must necessarily be controlled by the reasoning process, such that:

$$\frac{R}{P} \xrightarrow[a_r \cap \mathbf{c}=\top]{f} \frac{R\prime}{P}, \frac{R}{P} \xrightarrow[a_r \cap \mathbf{c}=\bot]{f} \frac{R\prime}{P\prime} \tag{10}$$

In other words, this model allows reformulating Equation 6 such that:

$$\frac{\left\langle \mathbf{c}=\top, KB_{1:t}, \iota, \tau, w\prime_t \rightarrow \left\{ \begin{array}{l} \cdots \frac{\langle KB_{1:t}, \iota, \tau, d=\emptyset \rangle}{\langle s_{t+1}, \sigma, \pi_{t+1}, a^1, a^2 \rangle} \\ \cdots \frac{\langle KB_{1:t}, \iota, \tau, d=\emptyset \rangle}{\langle s_{t+1}, \sigma, \pi_{t+1}, a^1, a^2 \rangle} \end{array} \right\} \right\rangle}{\langle s_t, \sigma, \pi_t, a^1, a^2 \rangle} \xrightarrow[a_r]{\pi(\sigma(w^1\prime_{t+1}), KB_{1:t})}$$
$$\xrightarrow[a_r]{\pi(\sigma(w^1\prime_{t+1}), KB_{1:t})} \frac{\left\langle \mathbf{c}=\top, KB_{1:t}, \iota, \tau, w\prime_t \rightarrow \left\{ \begin{array}{l} \cdots \frac{\langle \{KB_{1:t} \cup p^1_{t+1}, \ldots, p^m_{t+1}\}, \iota, \tau, d=\emptyset \rangle}{\langle s\prime_{t+1}=a_r(s_{t+1}), \sigma, \pi_{t+1}, a^1, a^2 \rangle} \\ \cdots \frac{\langle KB_{1:t}, \iota, \tau, d=\emptyset \rangle}{\langle s_{t+1}, \sigma, \pi_{t+1}, a^1, a^2 \rangle} \end{array} \right\} \right\rangle}{\langle s_t, \sigma, \pi_t, a^1, a^2 \rangle} \tag{11}$$

If such a switch exists as a *structure* in substrate machinery, and higher-level reasoning can control it (there is a *function* associated with it), its value over time is represented as a percept in the KB.

Implication: subjective experience

Now equipped with a clear, formal representation of awareness of phenomenal experience (in short, to experience), we may see how the axiom of ever-evolving perception gives rise to qualia.

Addressing the experience of the same stimuli by entities on different substrates (e.g., humans or bats): given that different sensory organs and processing (e.g., different visual cortices) implement different functions, they give rise to different sensory data from the same world state: i.e., $\sigma^1 \neq \sigma^2 \rightarrow \sigma^1(w_t) \neq \sigma^2(w_t)$.

Addressing entities on (quasi-)identical substrates (e.g., two human beings) experiencing the same stimuli differently: the perception function evolves over time, as perception computations need necessarily change in function of past experience. Consider a perception function that includes a propositional formula such as $\Phi = p^Y_t \iff p^X_t \cap p^U_t$ (where p^Y_t, p^X_t, and p^U_t are logical propositions) alongside other formulae that can perceive p^Y_t, p^X_t, and p^U_t from sensory data. At a given time k, this function is given data that implies p^Y_k, p^X_k, and $\neg p^U_k$, contradicting Φ. This contradiction implies that Φ was wrong, and must be removed from perception logic. Thus, if two entities, with corresponding initial perception functions π^1_0, π^2_0 are ever exposed to different stimuli (and they necessarily must be, as two different entities cannot occupy the same space at the same time), then necessarily $\exists k \rightarrow \pi^1_k \neq \pi^2_k$. The same world state will produce different perception on an entity that is aware of such phenomenal experience, as previously established.

Analysis

As human beings reason about their current state at current time, they perform a decision process that infers hypothetical future world states in function of their actions. This process necessarily includes negative future states (i.e., where there is hunger, pain, etc.), allowing us to prevent the actions that lead to them. As we assess the relative value of such states, we engage our sensory and perceptual sub-systems[21–23]. If the associated autonomous responses were enabled, we would recoil from hypothetical pain, potentially injuring ourselves from the movement initiated by that recoil. Thus, it it evolutionarily advantageous to be able to turn off, or at least dampen[40], autonomous responses when engaging sub-systems to process hypothetical data. Such a switch must necessarily be controlled by the reasoning system: thus, the reasoning system can be include a percept for its value. In other words: there exists a boolean value reasoning that can be interpreted as "*this (what I am sensing and perceiving) is real, happening here and now*", as opposed to imagined or hypothesized. We argue this satisfies all criteria for awareness of the phenomenal experience[41]. Awareness of the phenomenal experience (what we vaguely referred to when we said "to experience") is the capability of possessing an internal model of "I" (necessary for reasoning, and achievable recursively) and the capability to discern a reality here and now to which that "I" is subject to. It is a property of our biological substrates that arose simply through evolutionary processes: it is advantageous for reasoning entities to be aware of the phenomenal

experience. The consciousness switch allows us override legacy evolutionary responses: e.g., it allows us to withstand pain without recoil, by *consciously* choosing to experience that pain, if such an experience is advantageous. Its existence is supported by empirical data on sub-system neuronal activity in reality versus recall[29]. The "evidence for a temporal circuit characterized by a set of trajectories along which dynamic brain activity occurs" discovered by Huang et al[42], regulating the activity of two distinct cortical systems (the default mode network and the dorsal attention network) is consistent with the hypothesis, suggesting different substrate sub-systems for different aspects of reasoning. Similarly, Dehaene, Lau and Kouider[2] suggest that the common use of the word "consciousness" "...conflates two different types of information-processing computations in the brain: the selection of information for global broadcasting... , and the self-monitoring of those computations,...": we suggest this self-monitoring forms the basis of consciousness, whilst the selection of information for broadcasting corresponds to the distribution of software across substrate resources. Findings of neurological markers of consciousness in non-human animals that we consider highly intelligence (e.g., corvid birds[43]) support the link between consciousness and intelligence. If the consciousness switch allows us to separate reality from imagination, it explains why hallucinogenics alter consciousness by interfering with the connection between sensory signals and higher cognition[44], and why such chemicals can be used to improve consciousness in patients with schizophrenia[45].

In the context of AI, such a switch is easily implemented, showing artificial consciousness is readily realized. It is critical to note, though, that if an AI is built without autonomous responses, there is no need for it to ever be conscious: a conscious AI on reactive substrates would behave, at best, as a non-conscious AI on non-reactive substrates.

Discussion

We begin by refuting two main philosophical arguments that reject the physicalism premise. Chalmers'[46] argument stems from four premises he takes as axioms (reproduced from[8]):

1. "In our world, there are conscious experiences."
2. "There is a logically possible world physically identical to ours, in which the positive facts about consciousness in our world do not hold."
3. "Therefore, facts about consciousness are further facts about our world, over and above the physical facts."
4. "So, materialism is false."

This is known as the "Zombie argument"[3], which has been heavily disputed[47]. Our hypothesis particularly challenges the second premise: a world where the facts about consciousness do not hold cannot be physically identical to ours in every other way, as the evolutionary path of (at least) the human race would have been quite different. This rebuttal arises naturally by realizing that Chalmers' second premise axiomatizes his conclusions: "There is a logically possible world physically identical to ours, in which the positive facts about consciousness in our world do not hold" implicitly assumes consciousness is not physical.

Although Jackson's "knowledge argument"[48] no longer holds the weight it once did, it's beneficial to see how an objection to it looks like from the formal perspective. Consider a property such as the color red, that when sensed visually yields the datum $d^{red} \in \{d^1, d^2, ..., d^n\} = \sigma(w_t)$, as a function of the corresponding wavelength. There exists a percept that represents the visual sensing of the color red in the absence of any other context, such that $p^{red} = \pi(d^{red})$. There exist percepts that would be semantically associated with p^{red}, if it were on a KB: percepts pertaining to the position of the color "red" on the spectrum, the word "red", associations with concepts such as "color" and "blood", etc. These associated percepts were inferred from data $\{d^a, d^b, ..., d^z\}$: the difference between experiencing these data and sensing red visually is merely the fact that $d^{red} \notin \{d^a, d^b, ..., d^z\} \rightarrow p^{red} \notin \pi(\{d^a, d^b, ..., d^z\})$.

We now offer preemptive responses to criticism of our model. We put forward that any argument for consciousness to be something more than a computational process must be supported by the identification of a measurable property which arises in *function of* consciousness and that is not explained by either non-conscious intelligence alone or by conscious intelligence as we have defined it. To the best of the authors' knowledge, no such property has been identified in the literature. Computational consciousness as we have defined it can explain the notion of time[49]: inferred world states in the past have been logged onto long-term memory with the associated consciousness switch "on", while hypothesized future states are logged with the switch "off", providing a clear distinction between past and (possible) future. Regarding self-awareness and meta-cognition[50], our hypothesis reverses the usually accepted causality: whilst the traditional view is that consciousness is necessary for or synonymous with these terms, we have argued they are the cause of consciousness, in evolutionary terms.

Conclusion

Much of the difficulty in identifying the mechanisms for consciousness arose from its conflation with intelligence. When reasoning has been formally defined and the implications of its execution on substrates where physical responses can be autonomously triggered are examined, the computational function of consciousness emerges as an evolutionary advantage.

This hypothesis is simpler than contemporary theories, from Integrated Information Theory (IIT[51]) (which we argue suffers from the conflation problem) to appeals to quantum processes[52], while satisfying all criteria we attribute to consciousness[41]; and, challenges the premises of the "hard problem"[46]: our model shows a world where the facts about consciousness do not hold[3] cannot be physically identical to ours in every other way, as the evolutionary path of (at least) the human race would have been quite different[47].

The hypothesis can be empirically tested by searching for a "consciousness switch" in neurological activity: for example, by identifying common neural activation discrepancies between real-time and recalled visual stimuli in healthy humans, periods of lucidity and non-lucidity in patients with dementia, reality and hallucinations in subjects on hallucinogenics.

References

1. Bengio, Y. & Elmoznino, E. Illusions of ai consciousness. *Science* **389**, 1090–1091 (2025).
2. Dehaene, S., Lau, H. & Kouider, S. What is consciousness, and could machines have it? *Robotics, AI, humanity: Sci. ethics, policy* 43–56 (2021).
3. Frankish, K. The anti-zombie argument. *The Philos. Q.* **57**, 650–666 (2007).
4. Broom, D. M. Concepts and interrelationships of awareness, consciousness, sentience, and welfare. *J. Conscious. Stud.* **29**, 129–149 (2022).
5. Duch, W. What is computational intelligence and where is it going? In *Challenges for computational intelligence*, 1–13 (Springer, 2007).
6. Minsky, M. L. & Laske, O. A conversation with marvin minsky. *AI Mag.* **13**, 31–31 (1992).
7. JO, A. The quest to test for consciousness. *Nature* **643**, 31 (2025).
8. Kuhn, R. L. A landscape of consciousness: Toward a taxonomy of explanations and implications. *Prog. biophysics molecular biology* **190**, 28–169 (2024).
9. Block, N. Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behav. brain sciences* **30**, 481–499 (2007).
10. Melloni, L., Mudrik, L., Pitts, M. & Koch, C. Making the hard problem of consciousness easier. *Science* **372**, 911–912 (2021).
11. Butlin, P. *et al.* Consciousness in artificial intelligence: insights from the science of consciousness. *arXiv preprint arXiv:2308.08708* (2023).
12. Minsky, M. Interior grounding, reflection, and self-consciousness. In *Information and Computation: Essays on Scientific and Philosophical Understanding of Foundations of Information and Computation*, 287–305 (World Scientific, 2011).
13. Minsky, M. Decentralized minds. *Behav. Brain Sci.* **3**, 439–440 (1980).
14. Pigliucci, M. Are plants conscious. *SKEPTICAL INQUIRER* **48** (2024).
15. Grinde, B. Consciousness makes sense in the light of evolution. *Neurosci. & Biobehav. Rev.* **164**, 105824 (2024).
16. Zhang, W. *State-space search: Algorithms, complexity, extensions, and applications* (Springer Science & Business Media, 1999).
17. Bui, X.-N. *et al.* Prediction of slope failure in open-pit mines using a novel hybrid artificial intelligence model based on decision tree and evolution algorithm. *Sci. reports* **10**, 9939 (2020).
18. VanderBrug, G. J. & Minker, J. State-space problem-reduction, and theorem proving-some relationships. *Commun. ACM* **18**, 107–115 (1975).
19. Núñez, R. *et al.* What happened to cognitive science? *Nat. human behaviour* **3**, 782–791 (2019).
20. Edwards, W. The theory of decision making. *Psychol. bulletin* **51**, 380 (1954).
21. Huang, J. *et al.* Neuronal representation of visual working memory content in the primate primary visual cortex. *Sci. advances* **10**, eadk3953 (2024).

22. Dake, M. & Curtis, C. E. Perturbing human v1 degrades the fidelity of visual working memory. *Nat. communications* **16**, 1–8 (2025).

23. Vredeveldt, A., Tredoux, C. G., Kempen, K. & Nortje, A. Eye remember what happened: Eye-closure improves recall of events but not face recognition. *Appl. Cogn. Psychol.* **29**, 169–180 (2015).

24. Mashour, G. A., Roelfsema, P., Changeux, J.-P. & Dehaene, S. Conscious processing and the global neuronal workspace hypothesis. *Neuron* **105**, 776–798 (2020).

25. Chitty-Venkata, K. T. & Somani, A. K. Neural architecture search survey: A hardware perspective. *ACM Comput. Surv.* **55**, 1–36 (2022).

26. Niell, C. M. & Scanziani, M. How cortical circuits implement cortical computations: mouse visual cortex as a model. *Annu. Rev. Neurosci.* **44**, 517–546 (2021).

27. Li, H.-H., Sprague, T. C., Yoo, A. H., Ma, W. J. & Curtis, C. E. Neural mechanisms of resource allocation in working memory. *Sci. Adv.* **11**, eadr8015 (2025).

28. Chen, W., Mo, Z., Xu, H., Ye, K. & Xu, C. Interference-aware multiplexing for deep learning in gpu clusters: A middleware approach. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 1–15 (2023).

29. Shi, H. *et al.* Task-irrelevant features in working memory alter current visual processing. *bioRxiv* 2024–10 (2024).

30. Alonso, R., Brocas, I. & Carrillo, J. D. Resource allocation in the brain. *Rev. Econ. Stud.* **81**, 501–534 (2014).

31. Börger, E. & Schewe, K.-D. A behavioural theory of recursive algorithms. *Fundamenta Informaticae* **177**, 1–37 (2020).

32. Mederos, S., Blakely, P., Vissers, N., Clopath, C. & Hofer, S. B. Overwriting an instinct: Visual cortex instructs learning to suppress fear responses. *Science* **387**, 682–688 (2025).

33. Hudson, A. J. Pain perception and response: central nervous system mechanisms. *Can. journal neurological sciences* **27**, 2–16 (2000).

34. Mobbs, D. *et al.* From threat to fear: the neural organization of defensive fear systems in humans. *J. Neurosci.* **29**, 12236–12243 (2009).

35. Kim, S. *et al.* Artificial stimulus-response system capable of conscious response. *Sci. Adv.* **7**, eabe3996 (2021).

36. Gerson, M. C., Abdul-Waheed, M. & Millard, R. W. Of fight and flight. *J. nuclear cardiology* **16**, 176–179 (2009).

37. Hofstadter, D. R. *I am a strange loop* (Basic books, 2007).

38. Chen, X. *et al.* Transcriptomic mapping uncovers purkinje neuron plasticity driving learning. *Nature* **605**, 722–727 (2022).

39. Birch, J. *The edge of sentience: risk and precaution in humans, other animals, and AI* (Oxford University Press, 2024).

40. Kerautret, L., Dabic, S. & Navarro, J. Exploring hazard anticipation and stress while driving in light of defensive behavior theory. *Sci. reports* **13**, 7883 (2023).

41. Tassi, P. & Muzet, A. Defining the states of consciousness. *Neurosci. & Biobehav. Rev.* **25**, 175–191 (2001).

42. Huang, Z., Zhang, J., Wu, J., Mashour, G. A. & Hudetz, A. G. Temporal circuit of macroscale dynamic brain activity supports human consciousness. *Sci. advances* **6**, eaaz0087 (2020).

43. Nieder, A., Wagener, L. & Rinnert, P. A neural correlate of sensory consciousness in a corvid bird. *Science* **369**, 1626–1629 (2020).

44. Ballentine, G., Friedman, S. F. & Bzdok, D. Trips and neurotransmitters: Discovering principled patterns across 6850 hallucinogenic experiences. *Sci. advances* **8**, eabl6989 (2022).

45. Watson, S. J., Berger, P. A., Akil, H., Mills, M. J. & Barchas, J. D. Effects of naloxone on schizophrenia: Reduction in hallucinations in a subpopulation of subjects. *Science* **201**, 73–76 (1978).

46. Chalmers, D. J. Facing up to the problem of consciousness. *J. consciousness studies* **2**, 200–219 (1995).

47. Márton, M. What does the zombie argument prove? *Acta Anal.* **34**, 271–280 (2019).

48. Jackson, F. Epiphenomenal qualia. In *Consciousness and emotion in cognitive science*, 197–206 (Routledge, 1998).

49. Kent, L. & Wittmann, M. Time consciousness: the missing link in theories of consciousness. *Neurosci. Conscious.* **2021**, niab011 (2021).

50. Koriat, A. *et al. Metacognition and consciousness* (Institute of Information Processing and Decision Making, University of Haifa . . . , 2006).

51. Tononi, G. *et al.* Consciousness or pseudo-consciousness? a clash of two paradigms. *Nat. Neurosci.* 1–9 (2025).

52. Hameroff, S. & Penrose, R. Consciousness in the universe: A review of the 'orch or'theory. *Phys. life reviews* **11**, 39–78 (2014).

Author contributions statement

W.S. and P.G. contributed equally to this work, formulating the initial hypothesis. P.G. was responsible for initial manuscript draft; W.S. was responsible for revisions and editing.

Data Availability statement

All data generated or analysed during this study are included in this published article.

Additional information

Competing interests

No competing interests to disclose.