Why Is Anything Conscious?

Michael Timothy Bennett[1*], Sean Welsh[2] and Anna Ciaunica[3]

[1*]School of Computing, Australian National University, ACT, Australia.
[2]Engine No.2, Bardon, Brisbane, QLD, Australia.
[3]Institute of Cognitive Neuroscience, UCL, WC1N 3AZ, London, UK.
[3]Centre for Philosophy of Science, University of Lisbon, Campo Grande, 1749-016 Lisbon, Portugal.

*Corresponding author(s). E-mail(s): michael.bennett@anu.edu.au;
Contributing authors: sean@engineno2.com; a.ciaunica@ucl.ac.uk;

Abstract

We tackle the hard problem of consciousness taking the naturally selected, embodied organism as our starting point. We provide a formalism describing how biological systems self-organise to hierarchically interpret unlabelled sensory information according to valence. Such interpretations imply behavioural policies which are differentiated from each other only by the qualitative aspect of information processing. Natural selection favours systems that intervene in the world to achieve homeostatic and reproductive goals. Quality is a property arising in such systems to link cause to affect to motivate interventions. This produces interoceptive and exteroceptive classifiers and determines priorities. In formalising the seminal distinction between access and phenomenal consciousness, we claim that access consciousness at the human level requires the ability to hierarchically model i) the self, ii) the world/others and iii) the self as modelled by others, and that this requires phenomenal consciousness. Phenomenal without access consciousness is likely common, but the reverse is implausible. To put it provocatively: death grounds meaning, and Nature does not like zombies. We then describe the multilayered architecture of self-organisation from rocks to Einstein, illustrating how our argument applies. Our proposal lays the foundation of a formal science of consciousness, closer to human fact than zombie fiction.

Keywords: hard problem of consciousness

1 Introduction

Why is anything conscious? Both biological and other physical[1] systems process information, yet humans consciously *experience* as well as process information. Why? Living organisms are constantly processing self and world related information to survive in an ever-changing world. Human bodies share with all other physical systems the property of being instantiated in time and space (e.g. occupy space). Yet, unlike physical systems, biological systems are dissipative systems using energy to self-organise in the face of entropic decay and environmental perturbation [1, 2].

Initially formally introduced in the field of cybernetics [3, 4] the notion of self-organisation has been applied to various disciplines including physics [5], biology [6, 7] and neuroscience [8–10]. Self-organisation is typically defined as the spontaneous emergence of spatiotemporal order or pattern-formation processes in physical and biological systems resulting from interactions of its components with the environment [11–13].

Interestingly, much self and world related information processing goes on behind the scenes, or "in the dark" so to speak, that is, without being constantly present to our conscious minds. But why doesn't all information processing go in the dark?

This question is the subject of long-standing debate [14]. One highly influential view is that consciousness has two aspects [14]. The first is functional, by which we mean the ability to **access** and communicate information [15]. How function relates to consciousness is considered the "easy problem" of consciousness [2]. The second aspect is "what it is like" to consciously experience information processing, or **phenomenal** consciousness [1, 2, 15–17]. This doesn't mean just global states like being awake, but more local states like smelling a cup of coffee. These local contents or "qualia" are characterised by what it is like to be in them [14]. It is unclear the extent to which functional and phenomenal aspects are independent. David Chalmers has influentially suggested that it may be possible to construct a "philosophical zombie" which acts in every way like a person but has no qualia [15]. For example, a thermostat certainly detects heat and so processes information, but there presumably is not anything it is like to be a thermostat. Hence the question "why is anything conscious" may be understood as "why is there sometimes a qualitative aspect to information processing?". This is the "hard problem" of consciousness.

The hard problem has sparked a substantial body of work, detailed discussion of which [14, 18] lies beyond the scope of our paper.

In what follows we start with the observation that the mental/physical distinction that fuels the classical debates, ranging from panpsychism [19] to physicalism à la Dennett [20] may not be the only way to tackling the hard problem of consciousness. We remain agnostic to the question whether conceivability arguments such as the zombie thought experiment [21] or the knowledge argument [22, 23] are the best way to approach the problem of phenomenal consciousness. For example, one may argue that artificial systems such as Large Language Models (LLMs) conceive entities that humans never conceived or never will. It remains an open question on whether these entities exist just on the basis that a system can conceive them abstractedly or not.

[1] Physical here just means non-biological. We are not suggesting biological systems are non-physical.

Our aim is more modest and starts with low level biological facts rather than high level philosophical abstractions to reformulate the problem of consciousness in a way that does not take for granted the mental/physical division in the first place. In this paper we suggest a way to dissolve the "hard problem" of consciousness is to reverse the order and start with the 'impure' embodied biological organism instead of the 'pure' abstract mental states.

Take for example the higher order thought theory (HOT) [24, 25] holds that the information of which a conscious being is aware are higher order "meta-representations" of lower order "local" mental states. Lower order states may include emotions and perceptions, while higher order meta-representations reflect upon those. The link between the two may explain something of the phenomenal character of states. Sense data is processed by the body resulting in lower order mental states, and then meta representations of those is where we might find more abstract conceptual or thought-like contents of consciousness.

HOTs are a good starting point to understand why most biological information processing goes on "in the dark", but why do these lower order states arise? Can they occur in the absence of subjective, qualitative experience? Is there something it is like to "be in the dark" processing information at the lower bodily levels?

Longtime considered a fringe approach, the embodied cognition paradigm [26] has recently gained substantial influence in cognitive science and philosophy [27–29]. The idea is that instead of considering the body as a mere device designed to fuel and contain the mind (a device that can be replaced with a vat or a robot, for example), one must consider the mind as serving the self-sustaining needs of a surviving body.

If this is so then understanding consciousness must start with understanding the 'humble' lower bodily levels of information processing, not the higher order levels of information processing only. In other words, conscious experiences do not merely depend on bodily experiences as an external factor that can be replaced with a vat or an artificial system.

Our approach is to try to rigorously define consciousness from first principles and show that some aspects of functional consciousness depends on phenomenal consciousness in a manner that makes zombies impossible, "dissolving" the hard problem by showing the phenomenal to be intrinsically functional[2]. We establish axioms that hold in every conceivable environment and show that it is impossible to have the function of consciousness without the subjective experience of it, and why most information processing goes on "in the dark".

Enactivism is roughly the view that information processing arises through interaction between an organism and its environment [33], and it is considered by some to be incompatible with narrow definitions of computation. The notion of computation is widely debated, and a detailed review of these discussions would lead us to a major digression [34, 35]. Here we define computation not in terms of abstract symbol shuffling and representation, but in concrete mechanistic terms where 'information processing' refers to the actual physical transition of a system from one state to another, so it is compatible with enactivism. Pancomputationalism is the idea that

[2] Note that we are far from the first to claim that the phenomenal is functional. For example, the proposed Conscious Turing Machine [30] based on Global Workspace Theory [31], and constructivist approaches to artificial intelligence [32] take similar position. It is our explanation of how and why that is novel.

all systems are computational. We build upon a formalism called pancomputational enactivism [36].

To answer "why is anything conscious", we use this formalism to show how lower and higher order theories, phenomenal and access consciousness all follow from first principles, scaling natural selection pressures and the ability to adapt[3]. First, in Section **2** we integrate and extend separately published works of narrower scope [36, 38, 39] to justify our model of self-organising systems. In Sections **3** and **4** we explain how this model formalises relevance realisation and unifies lower and higher order theories. Rather than assuming abstract objects are primary and trying to learn causal relations between them, this assumes valence is primary and that abstract objects are constructed to classify cause and anticipate valence. This takes us from attraction and repulsion to physical states, through lower order thoughts to higher order meta representations. We call this the **psychophysical principle of causality**. In Section **5** we extend previous work [39] on causal learning and the development of first (1ST), second (2ND) and third (3RD) order selves, and explain how they are a consequence of scaling the ability of a self-organising system to adapt with natural selection pressures. In Sections **6** and **7** we explain how subjective experience requires a 1ST order self, and conversely why a 1ST order self implies there is something it is like to be an organism that has a 1ST order self. We then argue that access consciousness requires the ability to communicate meaning, which requires both 1ST and 2ND order selves, and show how a philosophical zombie is impossible. In Section **8** we describe the development of consciousness as we scale up the capacity to adapt with natural selection pressures, with examples from nematodes to humans.

2 Back to Foundations

To meet length restrictions, the full mathematical definitions have been placed in the appendices [40]. The core arguments should be understandable without the math, but for convenience here is a quick reference guide:

1. A set Φ is assumed, whose elements are called **states**. A **declarative program** is set f of states, and f is true relative to $\phi \in \Phi$ iff $\phi \in f$. P is the set of all programs. Embodiment is formalised as an **abstraction layer**, for which we single out finite set $\mathfrak{v} \in 2^{\Phi}$ called a **vocabulary**. A choice of vocabulary $\mathfrak{v}$ implies a **formal language** $L_{\mathfrak{v}} = \{l \subseteq \mathfrak{v} : \bigcap l \neq \emptyset\}$.
2. Members of $L_{\mathfrak{v}}$ are called **statements**, and a statement l is considered true relative to a state ϕ iff $\phi \in \bigcup l$. A **completion** of a statement l_1 is any statement l_2 s.t. $l_1 \subseteq l_2$. The **extension** of a statement l is the set E_l of all its completions, s.t. $x \in L_{\mathfrak{v}}$ is $E_x = \{y \in L_{\mathfrak{v}} : x \subseteq y\}$. The extension of a set $X \subseteq L_{\mathfrak{v}}$ of statements is $E_X = \bigcup\limits_{x \in X} E_x$.
3. A $\mathfrak{v}$-task α is a pair $\langle I_\alpha, O_\alpha \rangle$ where $I_\alpha \subset L_{\mathfrak{v}}$ are the **inputs**, E_{I_α} is the **outputs** available given I_α, and $O_\alpha \subset E_{I_\alpha}$ are the **correct outputs**. All the tasks an

[3]This is loosely inspired by a scale-based framing of machine learning [37] - we are applying Sutton's "bitter lesson" to biology.

abstraction layer tasks form a hierarchy. If $\alpha \sqsubset \omega$ then ω is a **parent** of α, meaning $I_\alpha \subset I_\omega$ and $O_\alpha \subseteq O_\omega$. A parent is higher level than its children.

4. A **policy** is a statement π and a policy is **correct** for α iff $\pi \in \Pi_\alpha = \{\pi \in L_\mathfrak{v} : E_{I_\alpha} \cap E_\pi = O_\alpha\}$. Given input $i \in I_\alpha$ and a policy π, to **infer** is to choose $o \in E_i$. A task is complete if $o \in O_\alpha$. Given a task definition α s.t. $\alpha \sqsubset \omega$, to **learn** is to choose a policy $\pi \in \Pi_\alpha$, and ω has been **learned** from α if one chooses $\pi \in \Pi_\omega$.
5. A **causal identity** for $\mathfrak{b}$ is a policy classifying the causal interventions undertaken by $\mathfrak{b}$. Subscript denotes the system that learned the causal identity, while superscript denotes the object it classifies. For example, $c^\mathfrak{b}_\mathfrak{a}$ denotes a causal identity learned by $\mathfrak{a}$ to classify $\mathfrak{b}$'s interventions, and $c^{\mathfrak{ba}}_\mathfrak{a} \subset c^\mathfrak{b}_\mathfrak{a}$ denotes $\mathfrak{a}$'s prediction of $\mathfrak{b}$'s prediction of $\mathfrak{a}$.

We begin by formalising all conceivable environments [36, 41] from first principles:

Axiom 1: When there are things, we call these things the **environment**.

Axiom 2: Where things *change* or *differ*, we have different **states** of the environment.

Hence, the environment is a set Φ of mutually exclusive states. We call a set of states a declarative "program"[4]. A program is "true" relative to the states it contains. For example, if time is one dimension along which things change, then there is one state at a time. A **"fact"** is a program which contains the state at the present time. It follows that every aspect of the environment is a set of programs. An aspect of the environment is realised by (exists in) a state if the programs it contains are facts. Previous work argues this formalises every conceivable environment [41].

To give some intuition as to why, note that we assume nothing of what states might be or contain. We assume no differences within a state, only between states, and those differences are the programs of which aspects are formed. After all, one can only interact with aspects of one's environment. If one were to try and pinpoint what an aspect was made of, the answer would be deferred to other aspects. Any description that sought to fully encapsulate the semantics, truth or unmediated pure experience of a state would always be delayed or incomplete. This does not render such attempts vacuous, but speaks to their inherent contingency and conditionality, which is what this formalism attempts to capture.

2.1 Natural Selection and Embodiment

We assume a naturally selected, embodied organism. Every aspect of the environment is a set of facts[5], hence the body of an organism that persists across multiple states must be a set of programs. We call this set a **vocabulary**, and it is formalised in definition 2 (see appendix [40]).

[4] We do not concern ourselves with imperative programs, because there is an equivalence between declarative and imperative [42].

[5] Some may object, pointing out that this ignores composition. However, the application of a function is a fact regardless of whether it takes another function as input.

Piccinini [34, 35] divides computation into abstract and concrete sorts. Abstract computations mean or represent whatever we say they do. In contrast concrete **computations** are the actions of physical systems. The declarative programs we describe are **concrete**. Each captures a single point of difference between physical states. As such no two organisms have the same vocabulary because that would mean they are the same body.

This might seem inconvenient, as intuitively we might prefer to think about bodies as an abstract 'type'. However, doing so would undermine any claim we might make, as earlier results in artificial intelligence (AI) show [39, 43]. The meaning of abstractions depends upon interpretation, much like the behaviour of software is determined by the abstraction layer that "interprets" it. At its least abstract software is machine code[6]. Machine code is interpreted by hardware which determines every aspect of what that machine code does. A word of machine code is a mechanical trigger that only "means" whatever we have designed the hardware to do when we input that word. That hardware is an abstraction layer in which the software exists (including "higher level" abstraction layers like the Python interpreter). In other words, what we call software is nothing more than the state of hardware [36]. This has undermined all but the most subjective of claims regarding the behaviour of theorised, software superintelligence [44–46]. It is a flaw in the very idea of intelligent software. The distinction between software mind and hardware embodiment is subsequently called computational dualism [36], because it is reminiscent of how traditional Cartesianism conceived of a mental substance distinct from physical.

If we are to dissolve the hard problem by starting with the 'impure' embodied biological organism instead of the 'pure' abstract mental states, then we cannot presuppose organisms use particular abstractions. Instead we must explain why particular abstractions are formed, in terms of contentless states, as the product of interaction between an organism and its environment [26, 33, 47]. This process is sometimes called "relevance realisation" [48–50].

Some have argued enactivism is incompatible with computation because the set of possible abstractions that must be searched for relevance realisation is intractable [51]. For tractability, one must first isolate a "small world" of relevant information from the "big world" of all information [52]. Part of the reason we can address this is because we formalise embodiment, in concrete terms from the level of irreducible physical states. Embodiment implies finite resource constraints, so we adopt one more axiom from [41].

Axiom 3: All aspects of the environment are spatially extended [41].

This means that a body is realised by only a finite number of states (see def. 2). A body is a **vocabulary** $\mathfrak{v} \subset P$ with finitely many elements. $\mathfrak{v}$ implies a **formal language** $L_\mathfrak{v}$ **of interaction** between body and environment. A **statement** in this language is an aspect of the environment. A set of "programs" from our vocabulary. Statements have truth conditions with respect to environmental states, and every

[6] A "word" of machine code triggers a mechanistic process hardwired in the CPU by the human design. For example a word may copy the 32 bit value stored at memory address X into the 32 bit register Y, which the next line of machine code "adds" to register Z by looping over each bit in an "adder" circuit.

statement has an **extension**. The extension of a statement is the set of all statements which are supersets of the first statement (the set of all other statements by which the first statement is implied). Extension is important we can relate statements by their truth conditions, forming a lattice. Most importantly, we avoid the distinction between software and hardware and so avoid computational dualism.

Embodiment is dispositional. As natural selection evolves a vocabulary, it begins to isolate a "small world" relevant to survival [41, 51, 53], from the intractable "big world" P of all information.

2.2 Self-Organising Systems as Self and World Constraints

Both snowflakes and human bodies self-organise, yet we regard only the latter as conscious. What separates the two? To tackle this question we need self and world constraints.

Using our embodied formal language $L_{\mathfrak{v}}$, we can talk about computation with inputs and outputs by treating everything as embodied statements embedded and enacted within the environment. Given an **input** $i \in L_{\mathfrak{v}}$, the set of all possible **outputs** is the extension E_i of that input. This is because if i is realised by the environment, then the environment is constrained to those states that realise i, which constrains what other statements are realised. If behaviour is "motivated", then statements have **valence** determined by natural selection and only a subset of the statements a body could make are "fit". Hence an organism will self-organise to express some statements, but not others. Though environmental states are contentless and their only significance is in how they relate to one another, natural selection must make them attractive or repulsive to a self-organising system. This is formalised by the $\mathfrak{v}$-task as in definition 3[7]. If $\mathfrak{v}$ is the vocabulary of a body, and $\mathfrak{v}$-task $\mu = \langle I_\mu, O_\mu \rangle$ is fit behaviour, then I_μ is all the statements that body can express in which it is possible to remain fit, and O_μ is all the statements that body can express in which it remains fit. The extension E_{I_μ} of I_μ would be every output that it is *possible* to choose given the inputs I_μ, but only a subset of those $O_\mu \subset E_{I_\mu}$ are **correct outputs**.

2.3 Inference

Every statement in $L_{\mathfrak{v}}$ implies a constraint, because there are only so many outputs that can be expressed by a body at the same time as any given statement. Assume we want to constrain the body to O_μ given I_μ. A body could express a statement π (meaning π is true), and π could then constrain the body to only correct outputs $O_\mu \subset E_{I_\mu}$ if $O_\mu = E_{I_\mu} \cap E_\pi$. We call a constraining statement a **policy**. A policy constrains outputs given inputs, and a **correct policy** is one that constrains outputs to only **correct outputs**. For the sake of intuition, think of "correct" as "fit" according to natural selection[8]. This is formalised in definition 4.

[7] A $\mathfrak{v}$-task is called because a $\mathfrak{v}$-task because it exists in the context of a vocabulary $\mathfrak{v}$.
[8] This need not always be the case.

2.4 Learning

If an organism remains alive, then its history is an ostensive definition of "fit" self-organising behaviour. That history is expressed as a $\mathfrak{v}$-task $\mathfrak{h}$, where each input and output in $\mathfrak{h}$ is an interaction between organism and environment. If the set of all fit behaviour is a task μ, then the organism remains fit then $\mathfrak{h} \sqsubset \mu$.

"I survive therefore my model is viable." - Mark Solms [54]

$\mathfrak{h}$ implies a set $\Pi_{\mathfrak{h}}$ of policies, and not all of those policies will constrain outputs the same way given new inputs. Some policies will "generalise" to imply fit behaviour in unfamiliar circumstances, meaning correct outputs given a new set of inputs. If those policies imply fit behaviour given new inputs, then the organism will remain fit. An organism learns fit behaviour by learning $\pi \in \Pi_{\mathfrak{h}}$ such that $\pi \in \Pi_{\mu}$.

"The best model of the world is the world itself" - Rodney Brooks [55]

Earlier work [38] showed that the optimal strategy to adapt as fast as possible[9] is to prefer "weaker" policies, meaning those with larger extensions. The policy which generated outputs O given inputs I is most efficiently identified by constructing a policy π such that π generates O from I, and π implies the weakest constraint that can be implied while still generating O from I. This means a system that does so will construct fit policies from a shorter history [39] than one that does not (see def. 5 and proofs in appendix [40])[10].

3 Relevance Realisation Through Causal Learning

When an embodied organism learns a policy, it learns a relevant interpretation of that information, labelling it through behaviour. To efficiently adapt, an organism must correctly anticipate valence, which means correctly identifying what causes valence.

All organisms (see definition 6) must have preferences, because they will make some decisions and not others. Behaviour implies policies, and every policy implies tasks, so we can define preferences as an binary relation over tasks.

Natural selection favours adaptability, hence it prefers organisms that optimise for weak policies, which we call **weak policy optimisation** (w-maxing). By constructing fit policies, an organism divides the world into relevant objects and properties. A weaker policy implies all more specific versions of itself, meaning those that more tightly constrain outputs by having a smaller extension. Hence policy learning implies a lattice of policies that vary in weakness. Every fit policy would be a "causal identity" for something relevant, for example a specific object like "this coffee" or a weaker and more general concept like "all coffee".

Organisms interact when they "affect" one another (see definition 9) in the causal or physical rather than psychological sense of the word. Now, Artificial Intelligence (AI) and machine learning [58–61] are concerned with engineering adaptive agents.

[9] Meaning to converge on a fit policy given the smallest history possible, to realise fit behaviour.

[10] The effect of insufficiently weak representations can be observed in large language models like GPT-4 [56, 57], where they fail to make consistent statements about an object because represent it as multiple objects.

In that context, causality has become a mainstream topic of research. Causal learning is demonstrably necessary to thrive in an interactive setting [62, 63]. Where the causal graph is known in advance (for example if measuring the efficacy of medical interventions), this issue could be resolved using causal language [64] to represent an intervention *from outside* the system the graph describes.

To illustrate [39], suppose Bob is attempting to learn and predict the environment. Bob has observed Alice wearing a raincoat only when it rains, and Bob has seen rain only when he observed Alice wearing a raincoat. If we represent the raincoat observation with a binary variable $O \in \{true, false\}$ and the advent of rain with a variable $R \in \{true, false\}$, and if Bob is a Bayesian, then Bob's observations will lead to the conclusion that $p(R = true \mid O = true) = 1$. This means Bob believes that if Alice wears a raincoat, then it must be raining. Now lets permit Bob to interact with its environment. Assume Bob wants it to rain. As $p(R = true \mid O = true) = 1$, Bob may conclude that forcing $O = true$ by holding a gun to Alice's head and demanding Alice wear a raincoat will cause it to rain. This is absurd. $O = true$ does not represent the event $q =$"I coerced Alice into wearing a raincoat", but an entirely different event $v =$"Alice decided to put on a raincoat for the same reason I have observed Alice wearing a raincoat in the past". To accurately represent the environment, we need to represent that $p(R = true \mid q) = p(R = true) \neq p(R = true \mid v) = 1$, meaning both q and v. To illustrate visually, we started with the acyclic graph

$$O \longleftarrow R$$

and our intervention disconnected rain from the choice of clothing:

$$O \qquad\qquad R$$

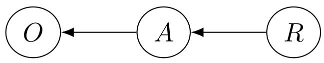
This can be resolved by introducing a "do" operator that we apply to a variable O to obtain $do[O = o]$, to represent the fact that an agency from outside the system has intervened to assign a value to O, so that we can represent $p(R = true \mid do[O = true]) = p(R = true) \neq p(R = true \mid O = true) = 1$. Thus, the aforementioned q is equivalent to $do[O = true]$, while v is equivalent to $O = true$.

3.0.1 The Psychophysical Principle of Causality

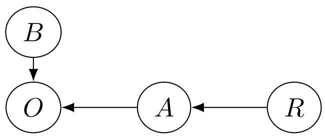
We have established the *need* for causal reasoning, but not how one might come to know the objects involved (where did the variables come from?), or how they relate to one another causally (to form a graph).

There are two ways to learn causal graphs. One can either assume a set of variables and learn relations between them, or assume relations and learn the objects that fit them [39]. We do the latter. Rather than assuming objects are primary and trying to learn causal relations between them, this assumes valence is primary and that abstract objects are constructed according to what **causes** valence. We call this the **psychophysical principle of causality**. This is what will take us from attraction to and repulsion from contentless physical states, through lower order thoughts to higher order meta representations. We start with the complete absence of content [65], and once a system learns policies there are distinct "contents". Policies serve to classify objects and properties. Death grounds meaning.

To develop this idea, we first show adding additional nodes to a causal graph is an adequate substitute for the "do" operator [66]. Assume we again have Bob who constructs a causal graph of the environment. Assume Alice exists in that environment. From Bob's perspective, Alice is just a part of the environment represented by a variable A in Bob's causal graph.



Now assume Bob observes Alice taking an action that changes an aspect of the environment, represented by the variable O (for example, Bob observes Alice putting on the raincoat). From Bob's perspective, Alice's action assigns values to variables A and O. There is no need to involve a do operator in this scenario because we can already represent that $p(R = true \mid A = x, O = true) = p(R = true) \neq p(R = true \mid A = y, O = true) = 1$ (because Alice is part of the causal graph). Likewise, we can add variables for interventions of Bob [39].



Now we will show how it is possible to *learn* these nodes. The matter of *which* cause-and-effect relations are learned is determined by valence, and so the nodes learned are statements classifying *causes* and *valence*.

4 Relevant Causal Identities

The optimal choice of policy for adaptability is the weakest because it is implied by more statements in an embodied language. That language is dispositional [67], shaped by natural selection, and so a weaker policy is likely to be a fit policy. w-maxing will correctly identify cause and effect where relevant, simplifying the entire environment into objects and properties relevant to the organism's motivations.

Assume a vocabulary $\mathfrak{v}_\mathfrak{b}$ belonging to an organism $\mathfrak{b}$ (Bob). A "cause" in the context of this formalisation is not a variable but a *statement* $l \in L_{\mathfrak{v}_\mathfrak{b}}$. The raincoat example would involve $obs, rain \in L_{\mathfrak{v}_\mathfrak{b}}$ such that:

$$obs \leftrightarrow \textit{“Alice put on a raincoat”} \text{ and } rain \leftrightarrow \textit{“It rained”}$$

obs and *rain* have truth values in accord with the definition. As in the earlier example Bob's passive observation implies $p(rain \mid obs) = 1$. The statement $obs =$

"Alice put on a raincoat" can be made true by either passive observation or intervention. However, the statement which is true in the case of intervention not *only obs*, but $int \in L$ such that $obs \subseteq int$ and:

$$int \leftrightarrow \textit{“Alice is wearing a raincoat because of Bob's actions”}$$

$$int \longrightarrow obs \longleftarrow rain$$

So long as $obs \neq int$, the intervention *can* be differentiated from the passive observation (see definition 10).

This being the case, any set $c \subseteq int - obs$ could be used to identify the party undertaking the intervention, which is why c is referred to as a "causal identity". It distinguishes the intervention *int* from the passively observed effect *obs*, like reafference in living organisms. However, the above only considers one intervention. A *weaker* or more general causal identity would be one that is shared by more interventions.

For example, suppose the inputs Alice has been subject to are $I_{\mathfrak{h}_{<t_\mathfrak{a}}}$. These can be divided into those in which Bob affected Alice $I_\mathfrak{a}^\mathfrak{b}$ and those in which Bob did not $I_\mathfrak{a}^{\neg\mathfrak{b}} = I_{\mathfrak{h}_{<t_\mathfrak{a}}} - I_\mathfrak{a}^\mathfrak{b}$. Alice can construct a causal identity b for Bob corresponding to interventions $INT = I_\mathfrak{a}^\mathfrak{b}$ and observations $OBS = I_\mathfrak{a}^{\neg\mathfrak{b}}$ (see definition 11).

4.1 Ascribing Intent to Other Objects

The distinction between "intervention" and not is misleading. Passive observation is just bearing witness to an intervention by something other than one's self. The question is not "is this an intervention" but "by whom was this intervention made?".

Earlier, we arrived at the following graph in which Bob's intervention was given by *int*.

$$int \longrightarrow obs \longleftarrow rain$$

What if a third person Larry puts the coat on Alice? Bob may observe this, and so Bob's observation of Larry's intervention is $v \in L_{\mathfrak{v}_\mathfrak{a}}$ such that $obs \subset v$. To account for this, Bob can construct a causal graph as below (with $b.$ representing Bob and $l.$ representing Larry).

$$\begin{tikzcd}
l.\ int \arrow[dr] & & \\
b.\ int \arrow[r] & obs & rain \arrow[l]
\end{tikzcd}$$

Bob's causal identity for himself $c_b \subset b.\ int - obs$ only represents the intervention by himself. However, now we can see that Bob must also construct an identity c_l for Larry, where the $c_l \subset l.\ int - obs$. For an organism $\mathfrak{a}$ with an embodied language $L_{\mathfrak{v}_\mathfrak{a}}$ to construct a causal identity for an object $\mathfrak{b}$, it must first be the case that $\mathfrak{a}$ is affected by $\mathfrak{b}$ [56], to satisfy the *incentive* precondition for causal identity.

Assume an organism $\mathfrak{a}$ is affected by $\mathfrak{b}$ given inputs INT, and not affected given inputs OBS. To then attribute the contents of INT to one specific entity, there must be something in common between the members of INT caused by $\mathfrak{b}$ that is not shared by any member of OBS caused by something else (in other words it must be at least possible for $\mathfrak{a}$ to discern the existence of $\mathfrak{b}$). The contents of INT are "interventions" by $\mathfrak{b}$ and by learning c, a corresponding causal identity, $\mathfrak{a}$ can discern the existence of $\mathfrak{b}$. This is not to say that $\mathfrak{b}$ *has* intent, but intent can be ascribed to $\mathfrak{b}$ because $\mathfrak{b}$ affects $\mathfrak{a}$, who can then discern when interventions are a consequence of $\mathfrak{b}$.

4.2 Preconditions

There are preconditions for the existence of a causal identity. First, the vocabulary $\mathfrak{v}_{\mathfrak{a}}$ of an organism $\mathfrak{a}$ must be of sufficient **scale** to ensure that observations are *distinguishable* from interventions.

Second, there must be an **incentive** to construct it. Inference is only possible if some states are preferable to others. One cannot derive what "ought" to be from a statement of what "is". Natural selection provides a notion of what ought to be, by eliminating anything which ought not.

1. The **scale** precondition requires $\mathfrak{v}$ contain the causal identity.
2. The **incentive** precondition is that fitness *demands* the causal identity.

4.3 Realising Lower Order States And Higher Order Meta Representations

Each policy an organism $\mathfrak{o}$ learns implies $\mathfrak{v}$-tasks. A $\mathfrak{v}$-task is a triadic relation between inputs, outputs and policies which resembles Peircean semiosis [56, 68] of sign, referent and interpretant. We formalise this as a "protosymbol"[11] system $\mathfrak{s}_{\mathfrak{o}}$ for the organism $\mathfrak{o}$ (see definitions 6 and 7).

Tasks exist in a "generational hierarchy". They are not mutually exclusive. Higher level tasks are more general, and have fewer policies because only very weak policies could complete them. Related as they are in a lattice, protosymbols are analogous **lower** order states and **higher** order meta representations. Some have framed consciousness as a problem of moving from unary, to dyadic, to triadic relations [69]. Similarly, we have gone from unary states to dyadic programs, to triadic tasks and protosymbols.

5 Multi-Layered Self-Organisation

As the vocabulary and capacity for w-maxing scale, a greater variety of concepts can be learned. Progressively higher orders of 'causal-identity' for one's self related information processing are constructed [39]. This lets us frame the construction of embodied selves in developmental [70] and evolutionary terms.

5.1 The First Order Self

A first order self (1ST henceforth, formalised in definition 13) allows an organism to discern the consequences of its actions. This serves as the locus of self related

[11] Proto because it is something more primitive than a symbol as conceived of by Peirce.

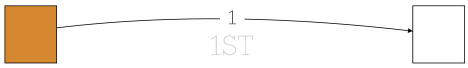


Fig. 1 Visual intuition for a 1ST order self. The organism constructs a causal identity for itself and can relate that to observed events.

information processing and experience [39], allowing the organism to plan complex interactions and maintain a consistent "self" that is part of the present interaction. 1ST order selves amount to reafference [71, 72], observed in mammals and insects.

A 1ST order self may require centralisation and a "solid brain" with a persistent structure [73] (an individual human), as opposed to a "liquid brain" (e.g. an ant colony or a population of humans).

5.2 The Second Order Selves

Survival may demand organism $\mathfrak{a}$ infer $\mathfrak{b}$'s prediction of $\mathfrak{a}$'s interventions (to see one's self as if through another's eyes [56]). This is called a second order self (2ND henceforth). We argue that if access conscious contents are available for **communication** in the human sense, then they must be communicable in the Gricean sense [74, 75]. Grice argued that communication is about the inference of intent. If person $\mathfrak{a}$ and $\mathfrak{b}$ are talking, then the meaning $m_\mathfrak{a}$ of what $\mathfrak{a}$ says is whatever $\mathfrak{a}$ intends $\mathfrak{b}$ understand. The meaning $m_\mathfrak{b}$ that $\mathfrak{b}$ understands is whatever $\mathfrak{b}$ thinks $\mathfrak{a}$ wants $\mathfrak{b}$ to think. $\mathfrak{b}$ has understood what $\mathfrak{a}$ means if $m_\mathfrak{b}$ approximates $m_\mathfrak{a}$. This can happen only if $\mathfrak{a}$ can predict with reasonable accuracy what $\mathfrak{b}$ thinks $\mathfrak{a}$ thinks, and $\mathfrak{b}$ can predict what $\mathfrak{a}$ thinks $\mathfrak{b}$ will think upon hearing an utterance. In other words, both $\mathfrak{a}$ and $\mathfrak{b}$ must have 2ND order selves that are good approximations. Yes, there are other aspects to communication.

However, here we are talking about consciousness. Access conscious contents are those available for reasoning and report. It follows[12] that access conscious contents must in principle be communicable in the sense Grice described.

As such, we argue contents available for communication can only be the contents of 2ND order selves, which means only an organism with 2ND order selves can be considered to have access consciousness. 2ND order selves also explain attention and self-awareness. An organism can have many 2ND order selves because they depend upon who or what the organism is interacting with, just as the availability of information depends on context.

Where a 1ST order self might allow one to observe a cat and form plans regarding causal interactions with the cat, a 2ND order self would allow one to be consciously *aware* of the cat for the purpose of reasoning and report. One can know of the cat, and that another organism knows of the cat, but a 2ND order self is insufficient to be aware that one is aware of the cat[13].

[12] If the definition of access consciousness is to be consistent with reasoning and report as exhibited by conscious humans.

[13] This is 'meta-self-reflexive consciousness' as some have described it [76].

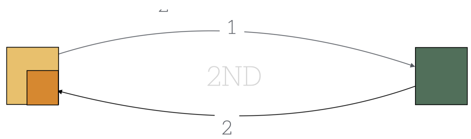


Fig. 2 Visual intuition for a 2ND order self. The organism constructs a causal identity for itself, and for another object, and the causal identity for the other object includes a prediction of the organism itself from that object's perspective. This would occur, for example, if one was a predator trying to predict the movements of prey in response to one's own actions. It is an extension of the 1ST order self.

More formally using the notation given in the appendix [40] (see quick reference guide in section 2), assume $\mathfrak{a}$ and $\mathfrak{b}$ are organisms that evolved to accurately predict one another's behaviour. Assume $\mathfrak{a}$ constructs a causal identity $c^{\mathfrak{b}}_{\mathfrak{a}}$ to predict $\mathfrak{b}$ given input $i_{\mathfrak{a}} \in I_{\mu_{\mathfrak{a}}}$, of which a second order self $c^{\mathfrak{b}\mathfrak{a}}_{\mathfrak{a}}$ is part. Likewise, $\mathfrak{b}$ constructs $c^{\mathfrak{a}}_{\mathfrak{b}}$ to predict $\mathfrak{a}$ given input $i_{\mathfrak{b}} \in I_{\mu_{\mathfrak{b}}}$, of which $c^{\mathfrak{a}\mathfrak{b}}_{\mathfrak{b}}$ is part. What is important here is that each organism's intent is to some extent inferred by the other, changing the sorts of policies that are fit. For example, second order self means each knows the other can anticipate manipulation, which means the optimal policy will often be to *have* rather than feign intent that aligns to some extent with the other party's desires, to co-operate [77][14]. Repeated interaction creates an iterated prisoner's dilemma, incentivising co-operation and signals that both parties interpret similarly (the beginnings of language) [56]. To communicate in Gricean terms, $\mathfrak{a}$ must intend to convey meaning $m_{\mathfrak{a}}$, and $\mathfrak{b}$ must recognise this intent. The incentive precondition explains *why* $\mathfrak{a}$ would form such intent (co-operation is often advantageous), while *how* may be understood as follows:

- $c^{\mathfrak{b}\mathfrak{a}}_{\mathfrak{a}}$ lets $\mathfrak{a}$ predict what $\mathfrak{b}$ will come to believe when it observes $\mathfrak{a}$'s behaviour.
- $c^{\mathfrak{a}\mathfrak{b}}_{\mathfrak{b}}$ then lets $\mathfrak{b}$ predict what $\mathfrak{a}$ intends that $\mathfrak{b}$ believe.

$\mathfrak{a}$ can use $c^{\mathfrak{b}\mathfrak{a}}_{\mathfrak{a}}$ to infer behaviour to which $\mathfrak{b}$ will ascribe the intent to communicate $m_{\mathfrak{a}}$, and $c^{\mathfrak{a}\mathfrak{b}}_{\mathfrak{b}}$ lets $\mathfrak{b}$ infer that this is what $\mathfrak{a}$ intends. The "utterance" Grice refers to is how $\mathfrak{a}$ affects $\mathfrak{b}$ in accord with earlier definitions. Put another way, $\mathfrak{a}$ *encodes* $m_{\mathfrak{a}}$ into its behaviour in a manner that $\mathfrak{b}$ can *decode* (their respective second order selves act as encoders and decoders). By encode and decode, we mean a loose approximation of $m_{\mathfrak{a}}$ is communicated. There are of course shortcuts, for example of $\mathfrak{a}$ and $\mathfrak{b}$ are of the same species then they likely have similar motives and experiences, and so the efficient thing for each to do would be to use its own intent as an approximation of what the other might think. However, that does not obviate the need for second order selves, it just makes them easier to realise.

5.3 The Third Order Selves

We can scale preconditions indefinitely, however for the purposes of this paper we will stop at 3RD order selves, because this appears to be the highest level of human consciousness, known as Meta self-awareness [76]. This is the awareness that one is

[14]Depending upon circumstances, for example organisms may co-operate in some circumstances but not others [56], and transient relations, information asymmetry and other factors can make deceit a more attractive option.

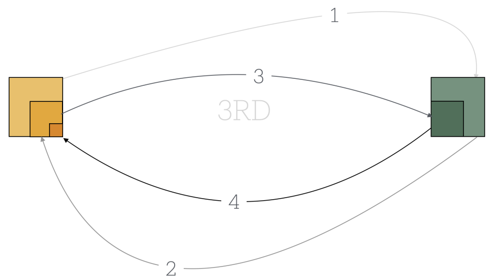


Fig. 3 Visual intuition for a 3RD order self. The organism constructs a 2ND order self for its 2ND order self, and so becomes aware that it is aware.

self-aware. If self-awareness stems from 2ND order selves, then it follows that meta self-awareness requires 3RD. Formally a 3RD order self for $\mathfrak{a}$ reflecting off $c_{\mathfrak{a}}^{\mathfrak{ba}}$ lets $\mathfrak{b}$ is $c_{\mathfrak{a}}^{\mathfrak{baba}}$. It is $\mathfrak{a}$'s prediction of $\mathfrak{b}$'s prediction of $\mathfrak{a}$'s prediction of $\mathfrak{b}$'s prediction of $\mathfrak{a}$.

6 The What and Why of Consciousness

Up to now we have developed the conceptual toolbox that we use to dissolve the hard problem. We started with the observation that a human body self organises to maintain itself in the face of change both within and without. An organism does not spring into existence understanding number systems or objects, but constructs them through interaction with the environment. One can regard them as policies according to which the environment is interpreted. To learn how to interpret the environment an organism must differentiate between states, must react to change, and learn policies according to the valence associated with that change.

We will now argue that: *i)* there is something it is like to be an organism with a 1ST order self, *ii)* that subjective experience is the process of learning and enacting a hierarchy or causal identities, and *iii)* that access consciousness requires both 1ST and 2ND order selves.

Because we do not presuppose representations, we explain how they emerge from attraction to and repulsion from contentless physical states. This means the foundation of information processing is valence.

As we saw earlier, learned policy has valence, and it is a classifier of inputs. Those inputs are "information" in the mechanistic sense of the environment being in one state and not another. Hence, a policy is a classifier of information, but that information is not in a language, and it is not yet something labelled or quantified (until there is a policy that labels or quantifies).

For biological, living self-organising systems, being alive is intrinsically good, being dead or ill is bad. For example, in developing self-organising systems such as the human body, it is good to have a better developed sense of smell than vision at early stages of life (babies are more accurate smellers than adults). Different senses have different degrees of valences at different stages of the developmental ladder. Hence there is "something what it is like" to be a baby which is different from something what it is like to be an adult. However, the idea here is that basic experiences (i.e. the self-organising process of staying alive) have valence at their very core (good=stay alive; bad=dead). This culminates in experiential, phenomenal 'quality' of appearance, of what it is likeness and how thing appear to the system (i.e. phenomena). That is the aspect that philosophers typically tackle. Our aim is not to dismiss the latter; our claim is that one cannot have the latter without the former.

Clearly, prelinguistic self-organising systems must classify and attach value and disvalue to states and anticipated states to prioritise and make decisions. To learn is to sense "something" and, motivated by valence, construct a policy classifying that "something". When we say physical states are contentless, we are not suggesting physical objects and properties don't exist. Objectively, a physical object is an aspect of the environment and every aspect is realized by one state or another. However, previous work has shown that for every aspect of the environment there is an equivalent program, meaning from an objective point of view what is or is not a physical object or property is a matter of interpretation [41]. Subjectively an object, property or quality only exists if an organism constructs a causal identity for it. At the most basic level, the foundation of all of this is physical attraction and repulsion. An organism hard-wired by natural selection can be physically attracted or repelled without any concept of what it is physically attracted or repelled from. It doesn't know to what it is attracted or repelled from in the sense that it has not constructed a causal identity for that object. There is no named object which has the property 'attractive'. The organism itself is simply attracted or repelled according to hard-wired policy. For example, a single-celled organism might be attracted to glucose, and its response to 'tumble' or 'stop' would be hard-wired. It does not satisfy the incentive or scale preconditions to construct a causal identity for glucose, or 'sweet', or hunger, or anything else. Intuitively, one might think of this like 'one dimensional' valence. To the bacteria, there is no subjective categorical variable or object to which it is attracted, though in the objective sense such a thing does exist. It is simply attracted or repelled. Over the course of this paper we have sought to explain how we get from simple 'one dimensional' hard-wired valence to a property or quality that has valence. Categorical variables for properties or 'qualities' like 'the colour red' or 'the smell of coffee' only exist when we have a causal identity for them, which requires scale and incentive. It can be hard-wired, or learned. When we have only one cell we do not have a vocabulary capable of satisfying the scale precondition. However self-organising biological systems are collective [78]. Control is delegated and distributed among smaller components [43] which operate concurrently. An individual cell might only have 'one dimensional' valence, but as soon as we have two cells we have something analogous to a second 'dimension'. Cellular collectives are polycomputational systems, meaning one cell can play a part in more than one computation simultanously, at different scales [79][15]. As we scale up the collective, its is impelled by the cells of which it is formed. Instead of 'one dimensional' valence, we now have a rich tapestry of competing drives. It is this rich tapestry that is a property or 'quality' of a state of the environment interpreted by the organism. It is this rich tapestry that can be reduced to causal identities for properties like 'the color red', 'hunger' or 'thirst'. A higher level of abstraction is

[15]This is what the weakness of a policy captures; the number of interactions of which it can be a part.

formed from this rich tapestry of valence and how it changes over many interactions (the different inputs and outputs in a task that imply generalizable causal identities). Hunger and thirst might have the same overall intensity and so one could say they have the same overall valence, but they are qualitatively different because they are different 'tapestries' of valence at a lower level of abstraction. This is what is meant by The Psychophysical Principle of Causality in section 3. In a self-organising system of sufficient scale with sufficient incentive, what starts out as simple 'one dimensional' valence culminates in categorical variables that serve different organismic needs. This is why the sound of middle C is different from the color blue. They have different causal identities. Middle C results from sound waves hitting the ear, blue results from light waves hitting the eye. These variations produce different qualia. Qualia do not need to be perfect. They simply need to be "good enough" to motivate actions necessary for survival in the environment of the organism. Qualia are produced by the sensory and nervous systems of the body. The valence of red light waves misses the eyes of dogs entirely. Dogs rely more on smell than sight. This suits their environmental niche. At a sufficient scale causal identities can be constructed, stored and recalled by their quality and valence.

The bold claim here is that information processing (i.e. exploring one's body and environment) in self-organising systems such as the human body is necessarily qualitative. Note that this is different from experience in the sense usually defined in the literature.

Experience in the sense of experiential phenomenological content cannot exist without experience in the sense of embodied exploration of the body and world as we defined here (see also Ciaunica in prep). To put it provocatively quality precedes quantity, and quantity is nothing more than the interpretation of quality. Quality comes first and experiences should be regarded on a continuum rather than a switch on/switch off phenomena. All living systems experience the world through their bodies and as such, there is something what it is like to experience the world in that basic way (even when one is asleep). One can access those phenomenal, experiential aspects at a higher level, true, but by accessing them, it doesn't mean that one 'constructs' consciousness or one becomes a conscious being. One is already consciously experiencing the world before one can explicitly access one's own experiences in 2ND order selves.

It follows that every policy learned in this way must classify a quality. Hence every such policy is a local state. A causal identity imbued with disposition by valence. There would be a policy for one's act of smelling coffee. For perceiving one's friend. There would be something different it is like to interact with a hostile version of that very same friend. The 1ST order self accompanies everything an organism does. It has a quality, so the 1ST order self is "**what it is like**" to be that organism. Put another way, Nagel's question of "what it is like to be" a particular organism could be answered if one could somehow have that organism's first order self [1]. An organism does not make a decision to interpret information, it just reacts. What we call **subjective experience** is the process of learning and enacting a lattice of causal identities, and there is something that "has" all of those subjective experiences once there is a 1ST order self that is part of them all. This is where **phenomenal consciousness** begins, with experience as learning and exploration.

One's 2ND order selves would also have a certain qualitative character, and one's 3RD order too.

There would, however, be a very clear delineation between conscious and not. The absence of a 1ST order self would mean there would be no policy linking all interventions together, and so no "self" that experiences making them. Hence a 1ST order self must precede higher orders.

Our arguments align with those of Merker, who has linked subjective experience to reafference [71, 72, 80]. We agree that reafference is key, but we provide a very different explanation of why and how. Their work presents biological evidence for subjective experience in organisms with reafference. In contrast, we derive the 1ST order self from first principles and explain why and how it is a classifies "what it is like" to be a particular organism. The 1ST order self also happens to be equivalent to reafference, so we arrive at the same conclusion as Merker from very different, mathematical premises. Hence these are complementary positions.

7 Unifying Lower and Higher Order Theories of Consciousness

The psychophysical principle of causality takes us from attraction to and repulsion from physical states, to higher order meta representations. This produces different orders of self. We suggest that the full richness of human subjective experience depends upon these different orders of self[16].

Conscious "access" as a human has it involves *meaningful* report, which as we've established requires 2ND order selves. This is very different from the mere "access" to information. However, supposing we could somehow contrive a system with 2ND order selves but no 1ST, then 2ND order selves alone would still not allow one to reason about how one's interventions might affect the contents of 2ND order selves to communicate meaning. For that, one requires a 1ST order self. Hence access consciousness requires both 1ST and 2ND order selves.

Our radical and provocative claim is that phenomenal consciousness without access consciousness is likely very common, but the reverse is implausible. For example, Block holds that phenomenally conscious content is phenomenal, whereas access conscious content is representational. In pancomputational enactivism there are only contentless states and the programs they imply. Any abstract "representational" content is just organised phenomenal content, clustered according to what *causes* valence. We don't discard these two sorts of consciousness, but unify them by showing how phenomenal consciousness gives rise to access.

Our framework thus makes a philosophical zombie impossible because there can be no perfect unconscious replica of a conscious human. The hard problem has it backwards. The question is not why qualia exist, but why anyone thinks representational contents can exist without first being learned through qualitative experience and discrimination.

[16]Perhaps the full richness of "hard" consciousness [81] depends upon the interaction of selves, perhaps in planning [82].

Far from suggesting there is no such thing as qualia, this suggests instead that there is no such thing as purely representational content in anything but the tools we construct and our interpretations of their behaviour. Cells are a material with agency [83], and human intelligence is the high level goal directed behaviour of a swarm of cells [28, 78, 84]. When we embody human abstractions (e.g. arithmetic) in silico (e.g. the x86 instruction set), we disconnect the high level goal directed behaviour from the low level behaviour (and the motivating valence) that gave rise to it [43]. We call the information embodied in the computer "representational" because it means something to us. However when we embody our abstractions in silico we disconnect them from the valence that motivated their construction [41]. This suggests there is no such thing as representational contents. They are a fiction we have invented because we struggle to reduce our own abstractions to their basic nature: the causes of valence. This 'dissolves' the hard problem by going a level down to the fundamental drive: stay alive!

8 From Rocks to Einstein: The Hierarchy of Being

To illustrate how our argument applies in the real world we describe stages of conscious organism. Each stage follows from scaling up supply and demand for w-maxing, through natural selection. For each stage we point out animals which are likely to be *at least* so conscious:

0 : Inert *(e.g. a rock)*
1 : Hard Coded *(e.g. protozoan)*
2 : Learning *(e.g. nematode)*
3 : 1ST Order Self *(e.g. housefly)*
4 : 2ND Order Selves *(e.g. cat)*
5 : 3RD Order Selves *(e.g. human)*

Stage 1: Hard Coded

Stage one refers to adaptations hard-wired by natural selection, allowing complexity to persist [85] in a stable environment.

- *What:* Hard-coded adaptations. Habituation and sensitization.
- *How:* The extension of fit behaviour is learned by natural selection and hard-coded into the organism as a policy (in DNA, form, the local environment etc).
- *Why:* If the environment is very predictable, it may be more efficient to hard-code fit behavior.
- *Example:* Single-celled protozoan.

Stage 2: Learning

Stage two introduces learning. To learn an organism must store, classify and order historical examples by valence. However there is not something it is like to be stage two, because there is no locus of "self". A biological example of such a decentralised nervous system is the cubozoan box jellyfish Tripedalia cystophora. Even Tripedalia cystophora was recently shown to be capable of associative learning [86]. An entirely distributed control system can "learn". Likewise, stage two is exemplified by nematodes [87, 88] with a centralised nervous system and some ability to adapt with experience. However, the absence of a "self" limits causal reasoning, which as others have already pointed out must limit spatial, navigational abilities [72]. When starved C. elegans exhibit "increased locomotion and dispersal in a random, rather than directed, search" [72, 89, 90], whereas something like a bee or an ant can recall and navigate to previously discovered food [91–93].

- *What:* Learning.
- *How:* Valence.
- *Why:* An organism that can learn can survive in more circumstances than one that cannot.
- *Examples:* Jellyfish, nematode.

Stage 3: 1ST Order Self

This is where phenomenal consciousness begins, with a 1ST order self. In biological terms this implies reafference, which others have argued is the key to subjective experience, albeit for different reasons than what we have [71, 72, 80]. They identified a housefly as a good example of where subjective experience may begin, and we concur. We also hold this is where an organism might be said to have intent. Intuitively, the policy that motivated behaviour is the intent of that behaviour[17]. For example, eating tends to involve the intent of satisfying hunger.

A stage three organism can feel simple things like hunger, but cannot cannot conceive of itself from another's perspective. Subsequently it cannot communicate in the Gricean sense [56, 75], or conceive of its own death, or experience shame.

- *What:* A 1ST order self. Reafference. Phenomenal consciousness.
- *How:* Embodiment in which intervention is not identical to observation.
- *Why:* Accurate prediction of consequences of interventions. For example, a fly must distinguish between having moved, and the environment having moved, to navigate.
- *Example:* Housefly.

Stage 4: 2ND Order Selves

Stage four is the 2ND order self, and this is where we hold access consciousness begins because it is where information is available for report in the Gricean sense. The ability of ravens to intentionally deceive [94] suggests they are *at least* stage four. Raven

[17] In the same way declarative and imperative programs are equivalent [42].

$\mathfrak{a}$, aware that it is being observed by raven $\mathfrak{b}$, will act as if it is hiding food in one location to mislead $\mathfrak{b}$, but will then move the food in another location unobserved by $\mathfrak{b}$. $\mathfrak{a}$ seems to predict not just the intent of $\mathfrak{b}$ (to steal the food), but $\mathfrak{b}$'s perception of $\mathfrak{a}$. It seems likely that dogs and cats have second order selves, as they must hunt reasonably intelligent animals and must anticipate how their actions are perceived. For example, a cat anticipates its prey will flee when it is observed, and hides.

- *What:* Access consciousness. Theory of mind. Self-awareness. Inner narrative.
- *How:* Selection pressures that demand theory of mind.
- *Why:* A 2ND order self is necessary to anticipate, manipulate and communicate intent. This could be anticipating the reactions of predator or prey, or navigating a social hierarchy.
- *Example:* Cats, dogs, ravens.

Stage 5: 3RD Order Selves

A 3RD order self facilitates complex planning, particularly of a social nature. A 3RD order self is a 2ND order self for one's 2ND order self. Humans appear to possess this, because we are aware that we are aware. Altruistic behaviour observed in Australian magpies [95] suggests they may also have 3RD order selves.

- *What:* Meta self-awareness. Inner narrative in which actors have inner narratives.
- *How:* More accurate prediction and planning.
- *Why:* Because a social organism must predict complex social dynamics.
- *Example:* Human. Highly intelligent animals such as Australian magpies may also be this conscious.

9 Conclusions: Why Nature Does Not Like Zombies

We have described a multilayered formalism illustrating how biological self-organising systems become phenomenally conscious when they construct a 1ST order self. A human lacking a 1ST order self could not perform causal reasoning needed to adapt as humans evidently can [62].

As previous research pointed out [36], the computer metaphor with hardware and software is a simplification. Rather, software is a state of hardware. Nature privileges efficiency over abstract simplification. Compare the vast quantities of both training data and energy required by a large language model, to the small quantities humans need to solve a problem. Biological systems are more efficient, because they are adaptive at a every level [28, 43]. Hence, rather than trying to explain the mind in the abstract like software, we started at the level of the embodied organism.

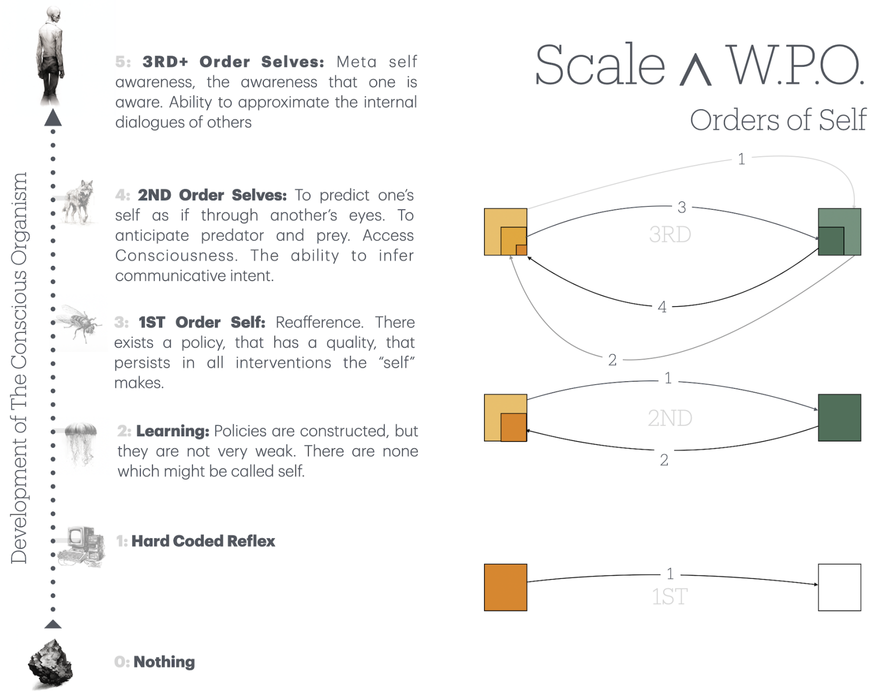


Fig. 4 Overview of stages and orders of self.

One consequence of our approach is that it places the phenomenal quality of conscious experience before access consciousness. We have shown a consistent definition of access consciousness requires 1ST and 2ND order selves. Phenomenal consciousness arises first, access comes later. Unlike panpsychism, we don't believe rocks are conscious. Only self-organising systems that need to adapt, motivated by valence, while keeping track of the self are conscious. Consciousness is an adaptation, and a philosophical zombie is impossible.

Furthermore, the conceivability of zombies depends on the ability to duplicate disposition but not phenomenal experience. Carruth argued that if qualia are dispositional, then that is impossible [96]. Our arguments align with Carruth's. A causal identity is inherently dispositional, as is embodiment.

In summary, we have argued that access consciousness at the human level is impossible without the ability to hierarchically model *i)* the self, *ii)* the world and others and *iii)* the self as modelled by others. These follow from learning and enacting a lattice of causal identities, if the scale and incentive preconditions are met. There is something it is like to be an organism with a self, and that is where phenomenal consciousness begins. Our proposal lays the foundation of a formal science of consciousness, deeply connected with natural selection rather than abstract thinking, closer to the organic facts of humanity than the conceptual fiction of zombies.

Funding

Supported by a Fundaçao para a Ciencia e a Tecnologia (FCT) grant PTDC/FER-FIL/4802/2020; and 2020-02773 CEECIND FCT to A.C.

10 Appendix

For convenience of reference, we have placed all definitions here. We have aimed to make the gist of the paper understandable without the math, hence the reader can either skip this section or refer back to it if needed. Many of the definitions have been adapted from a variety of preceding work [36, 38, 39, 41, 43, 56]. They are referred to in the body of the paper when they become relevant, in the order in which they appear here.

Definition 1 (environment)**.**

- *We assume a set Φ whose elements we call* ***states****.*
- *A* ***declarative program*** *is $f \subseteq \Phi$, and we write P for the set of all programs (the powerset of Φ).*
- *By a* ***truth*** *or* ***fact*** *about a state ϕ, we mean $f \in P$ such that $\phi \in f$.*
- *By an* ***aspect of a state*** *ϕ we mean a set l of facts about ϕ s.t. $\phi \in \bigcap l$. By an* ***aspect of the environment*** *we mean an aspect l of any state, s.t. $\bigcap l \neq \emptyset$. We say an aspect of the environment is* ***realised***[18] *by state ϕ if it is an aspect of ϕ.*

Definition 2 (abstraction layer)**.**

- *We single out a subset $\mathfrak{v} \subseteq P$ which we call* ***the vocabulary*** *of an abstraction layer. The vocabulary is finite.*
- *$L_\mathfrak{v} = \{l \subseteq \mathfrak{v} : \bigcap l \neq \emptyset\}$ is a set of aspects in $\mathfrak{v}$. We call $L_\mathfrak{v}$ a formal language, and $l \in L_\mathfrak{v}$ a* ***statement****.*
- *We say a statement is* ***true*** *given a state iff it is an aspect realised by that state.*
- *A* ***completion*** *of a statement x is a statement y which is a superset of x. If y is true, then x is true.*
- *The* ***extension of a statement*** *$x \in L_\mathfrak{v}$ is $E_x = \{y \in L_\mathfrak{v} : x \subseteq y\}$. E_x is the set of all completions of x.*
- *The* ***extension of a set of statements*** *$X \subseteq L_\mathfrak{v}$ is $E_X = \bigcup\limits_{x \in X} E_x$.*
- *We say x and y are* ***equivalent*** *iff $E_x = E_y$.*

(notation) *E with a subscript is the extension of the subscript*[19]*.*

(intuitive summary) *$L_\mathfrak{v}$ is everything which can be realised in this abstraction layer. The extension E_x of a statement x is the set of all statements whose existence implies x, and so it is like a truth table. Intuitively a sensorimotor system is an abstraction layer. Likewise, a computer (each statement asserting a state of the computer, or a part thereof).*

---

[18] Realised meaning it is made real, or brought into existence.
[19] e.g. E_l is the extension of l.

Definition 3 ($\mathfrak{v}$-task)**.**
For a chosen $\mathfrak{v}$, a task α is a pair $\langle I_\alpha, O_\alpha \rangle$ where:

- *$I_\alpha \subset L_\mathfrak{v}$ is a set whose elements we call **inputs** of α.*
- *$O_\alpha \subset E_{I_\alpha}$ is a set whose elements we call **correct outputs** of α.*

*I_α has the extension E_{I_α} we call **outputs**, and O_α are outputs deemed correct. $\Gamma_\mathfrak{v}$ is the set of **all tasks** given $\mathfrak{v}$.*

(generational hierarchy) *A $\mathfrak{v}$-task α is a **child** of $\mathfrak{v}$-task ω if $I_\alpha \subset I_\omega$ and $O_\alpha \subseteq O_\omega$. This is written as $\alpha \sqsubset \omega$. If $\alpha \sqsubset \omega$ then ω is then a **parent** of α. $\sqsubset$ implies a "lattice" or generational hierarchy of tasks. Formally, the level of a task α in this hierarchy is the largest k such there is a sequence $\langle \alpha_0, \alpha_1, ...\alpha_k \rangle$ of k tasks such that $\alpha_0 = \alpha$ and $\alpha_i \sqsubset \alpha_{i+1}$ for all $i \in (0, k)$. A child is always "lower level" than its parents.*

(notation) *If $\omega \in \Gamma_\mathfrak{v}$, then we will use subscript ω to signify parts of ω, meaning one should assume $\omega = \langle I_\omega, O_\omega \rangle$ even if that isn't written.*

(intuitive summary) *To reiterate and summarise the above:*

- *An **input** is a possibly incomplete description of a world.*
- *An **output** is a completion of an input [def. 2].*
- *A **correct output** is a correct completion of an input.*

*A $\mathfrak{v}$-task is a formal, **behavioural** description of goal directed behaviour. For example, an organism could be described by all behaviour in which it remains alive. Likewise, a $\mathfrak{v}$-task could describe a Turing machine.*

Definition 4 (inference)**.**

- *A $\mathfrak{v}$-task **policy** is a statement $\pi \in L_\mathfrak{v}$. It constrains how we complete inputs.*
- *π is a **correct policy** iff the correct outputs O_α of α are exactly the completions π' of π such that π' is also a completion of an input.*
- *The set of all correct policies for a task α is denoted Π_α.*[20]

Assume $\mathfrak{v}$-task ω and a policy $\pi \in L_\mathfrak{v}$. Inference proceeds as follows:

1. *we are presented with an input $i \in I_\omega$, and*
2. *we must select an output $e \in E_i \cap E_\pi$.*
3. *If $e \in O_\omega$, then e is correct and the task "complete". $\pi \in \Pi_\omega$ implies $e \in O_\omega$, but $e \in O_\omega$ doesn't imply $\pi \in \Pi_\omega$ (an incorrect policy can imply a correct output).*

(intuitive summary) *To reiterate and summarise the above:*

- *A **policy** constrains how we complete inputs.*

[20] To repeat the above definition in set builder notation:

$$\Pi_\alpha = \{\pi \in L_\mathfrak{v} : E_{I_\alpha} \cap E_\pi = O_\alpha\}$$

- *A **correct policy** is one that constrains us to correct outputs.*

In functionalist terms, a policy is a "causal intermediary" between inputs and outputs.

Definition 5 (learning)**.**
*A **proxy** $<$ is a binary relation on statements. In this paper we use only one proxy, called the **weakness proxy**, which compares the cardinality of a statement's extension. For statements l_1, l_2 we have $l_1 < l_2$ iff $|Z_{l_1}| < |Z_{l_2}|$. Whenever we use $<$ to compare statements, we are referring to the aforementioned weakness proxy.*

(generalisation) *A statement l **generalises** to a $\mathfrak{v}$-task α iff $l \in \Pi_\alpha$. We speak of **learning** ω from α iff, given a proxy $<$, $\pi \in \Pi_\alpha$ maximises $<$ relative to all other policies in Π_α, and $\pi \in \Pi_\omega$.*

(probability of generalisation) *We assume a uniform distribution over $\Gamma_\mathfrak{v}$. If l_1 and l_2 are policies, we say it is less probable that l_1 generalizes than that l_2 generalizes, written $l_1 <_g l_2$, iff, when a task α is chosen at random from $\Gamma_\mathfrak{v}$ (using a uniform distribution) then the probability that l_1 generalizes to α is less than the probability that l_2 generalizes to α.*

(sample efficiency) *Suppose $\mathfrak{app}$ is the set of **a**ll **p**airs of **p**olicies. Assume a proxy $<$ returns 1 iff true, else 0. Proxy $<_a$ is more sample efficient than $<_b$ iff*

$$\left(\sum_{(l_1,l_2)\in\mathfrak{app}} |(l_1 <_g l_2) - (l_1 <_a l_2)| - |(l_1 <_g l_2) - (l_1 <_b l_2)| \right) < 0$$

(optimal proxy) *There is no proxy more sample efficient than weakness. The weakness proxy formalises the idea that "explanations should be no more specific than necessary" (see Bennett's razor in [38]).*

(intuitive summary) *Learning is an activity undertaken by some manner of intelligent agent, and a task has been "learned" by an agent that knows a correct policy. Humans typically learn from "examples". An example of a task is a correct output and input. A collection of examples is a child task, so "learning" is an attempt to generalise from a child to one of its parents. The lower level the child from which an agent generalises to parent, the "faster" it learns, the more sample efficient the proxy. The most sample efficient proxy is weakness [38, prop. 1, 2], which is why we're using it here.*

Definition 6 (organism)**.**
We can describe the circumstances of an organism $\mathfrak{o}$ as $\langle \mathfrak{v}_\mathfrak{o}, \mu_\mathfrak{o}, \mathfrak{p}_\mathfrak{o}, <_\mathfrak{o} \rangle$ where:

- *$O_{\mu_\mathfrak{o}}$ contains every output which qualifies as "fit" according to natural selection.*
- *$\mathfrak{p}_\mathfrak{o}$ is the set of policies an organism knows, s.t. $\mathfrak{p}_\mathfrak{o} \subset \mathfrak{p}_{n.s.} \cup \mathfrak{p}_{\mathfrak{h}_{<t_\mathfrak{o}}}$ and:*
  - *$\mathfrak{p}_{n.s.} \subset L_{\mathfrak{v}_\mathfrak{o}}$ is **reflexes** hard coded from birth by natural selection.*

– $\mathfrak{p}_{\mathfrak{h}_{<t_\mathfrak{o}}} = \bigcup_{\zeta \in \mathfrak{h}_{<t_\mathfrak{o}}} \Pi_\zeta$ *is the set of policies it is possible to* ***learn*** *from a history of past interactions represented by a task* $\mathfrak{h}_{<t_\mathfrak{o}}$.
– *If* $\Pi_{\mathfrak{h}_{<t_\mathfrak{o}}} \not\subset (\mathfrak{p}_\mathfrak{o} - \mathfrak{p}_{n.s.})$ *then the organism has* ***selective memory****. It can "forget" outputs, possibly to productive ends if they contradict otherwise good policies.*

- $<_\mathfrak{o}$ *is a binary relation over* $\Gamma_{\mathfrak{v}_\mathfrak{o}}$ *we call* ***preferences****.*

(intuitive summary) *Strictly speaking an organism* $\mathfrak{o}$ *would be a policy, but we can describe the circumstances of its existence as a task* μ *that describes all "fit" behaviour for that organism. We can also identify policies the organism "knows", because these are implied by the policy that is the organism. Likewise, we can represent lossy memory by having the organism "know" fewer policies than are implied by its history of interactions. Finally, preferences are the particular "protosymbol" the organism will use to "interpret" an input in later definitions.*

Definition 7 (protosymbol system)**.**
Assume an organism $\mathfrak{o}$*. For each policy* $p \in \mathfrak{p}_\mathfrak{o}$ *there exists a set* $\mathfrak{s}_p = \{\alpha \in \Gamma_{\mathfrak{v}_\mathfrak{o}} : p \in \Pi_\alpha\}$ *of all tasks for which* p *is a correct policy. The union of all such sets is*

$$\mathfrak{s}_\mathfrak{o} = \bigcup_{p \in \mathfrak{p}_\mathfrak{o}} \{\alpha \in \Gamma_{\mathfrak{v}_\mathfrak{o}} : p \in \Pi_\alpha\}$$

We call $\mathfrak{s}_\mathfrak{o}$ *a "protosymbol system". A* $\mathfrak{v}$*-task* $\alpha \in \mathfrak{s}_\mathfrak{o}$ *is called a "protosymbol", and is "more abstract" if it is higher in the generational hierarchy.*

Definition 8 (interpretation)**.**
Interpretation is an activity undertaken by an organism $\mathfrak{o} = \langle \mathfrak{v}_\mathfrak{o}, \mu_\mathfrak{o}, \mathfrak{p}_\mathfrak{o}, <_\mathfrak{o} \rangle$*, as follows:*

1. *Assume an input* $i \in L_{\mathfrak{v}_\mathfrak{o}}$.
2. *We say that* i ***signifies*** *a protosymbol* $\alpha \in \mathfrak{s}_\mathfrak{o}$ *if* $i \in I_\alpha$.
3. $\mathfrak{s}^i_\mathfrak{o} = \{\alpha \in \mathfrak{s}_\mathfrak{o} : i \in I_\alpha\}$ *is the set of all protosymbols which* i *signifies.*
4. *If* $\mathfrak{s}^i_\mathfrak{o} \neq \emptyset$ *then* i ***means something*** *to the organism in the intuitive sense that there is "affect" or "value" compelling the organism to act.*
5. *If* i *means something, then* $\mathfrak{o}$ *chooses* $\alpha \in \mathfrak{s}^i_\mathfrak{o}$ *that maximises its preferences* $<_\mathfrak{o}$.
6. *The organism then infers an output* $o \in E_i \cap E_{\Pi_\alpha}$.

(intuitive summary) *Interpretation is inference, with the additional step of choosing policies according to preference. This allows for irrational and instinctive choices, as well as rational ones. Intuitively,* i *is every aspect of the context in which the organism finds itself; everything that can influence its interpretation.*

Definition 9 (to affect)**.**
Suppose we have two organisms, $\mathfrak{a}$ *(Alice) and* $\mathfrak{b}$ *(Bob). Suppose* $\mathfrak{a}$ *interprets* $i \in L_{\mathfrak{v}_\mathfrak{o}}$ *as an output* o*, then:*

- *a* ***statement*** $v \subset i$ *affects* $\mathfrak{a}$ *if* $\mathfrak{a}$ *would have interpreted* $e = i - v$ *as a different output* $g \neq o$.
- *an* ***organism*** $\mathfrak{b}$ *has affected* $\mathfrak{a}$ *by making an output* k *if, because of* k*, there exists* $v \subset s$ *which affects* $\mathfrak{a}$.

Definition 10 (intervention)**.**
By ***event*** *we mean a statement in* $L_{\mathfrak{v}}$*, and an event* ***happens*** *or is* ***observed*** *iff it is a true statement given a state* ϕ*. If* $obs \in L_{\mathfrak{v}}$ *is sensorimotor activity we interpret as an "observed event", and* $int \in L_{\mathfrak{v}}$ *is an* ***intervention*** *to cause that event, then* $obs \subset int$ *(because* int *could not be said to cause* obs *unless* $obs \subset int$*).*

(intuitive summary) *An intervention is action undertake an organism or other agency, in the sense described by Pearl [62]. Intuitively, if "int" and "obs" are events which have happened, then we say that int has* ***caused*** *obs if obs would not have happened in the absence of int (counterfactual).*

Definition 11 (causal identity)**.**
If $obs \in L_{\mathfrak{v}}$ *is an observed event, and* $int \in L_{\mathfrak{v}}$ *is in intervention causing* obs*, then* $c \subseteq int - obs$ *"identifies" or "names" the intervening agency. If* $c = \emptyset$ *then we have no way of knowing the intervening agency, if there is one. We call* c *a* ***causal identity*** *corresponding to* int *and* obs*. Suppose* INT *and* OBS *are sets of statements, and we assume* OBS *contains observed events and* INT *interventions, then a causal identity corresponding to* INT *and* OBS *is* $c \neq \emptyset$ *s.t.* $\forall i \in INT(c \subset int)$ *and* $\forall obs \in OBS(c \cap obs = \emptyset)$ *(we can attempt to construct a causal identity for any* INT *and* OBS*). If a policy is a causal identity, then the associated task is to classify interventions.*

Definition 12 (purpose, goal or intent)**.**
We consider a policy c *which is a causal identity corresponding to* INT *and* OBS *to be the* ***intent****,* ***purpose*** *or* ***goal*** *ascribed to the interventions.* c *is what the interventions share in common, meaning the "name" or "identity" of behaviour is the "intent", "goal" or "purpose" of behaviour. Just as an intervention caused an observation, the intent which motivated the agency undertaking the intervention is what caused it (to correctly infer intent, one must infer a causal identity that implies subsequent interventions).*

Definition 13 (1ST order self)**.**
If c *is the lowest level causal identity corresponding to* INT *and* OBS*, and* INT *is every intervention an organism could make (not just past interventions, but all potential future interventions), then we consider* c *to be the system's* ***1ST order self****. If* $c \in \mathfrak{p}_{\mathfrak{o}}$ *then an organism has constructed a 1ST order self. A 1ST order self for an organism* $\mathfrak{o}$ *is denoted* $\mathfrak{o}^1$*. An organism has at most one 1ST order self.*

(intuitive summary) *Intuitively,* $\mathfrak{o}^1$ *is where we draw the line between what the organism can intend and what it cannot. It is conceivable we might have two "organisms" in the same body by this definition, each with its own 1ST order causal identity. Ultimately, where an organism begins or ends remains malleable.*

Definition 14 (preconditions)**.**
If $\mathfrak{o}$ is an organism, and c is a causal identity, the $\mathfrak{o}$ will construct c only if the representation and incentive preconditions below are met:

- *the* ***scale*** *precondition is met iff $c \in L_{\mathfrak{v}_\mathfrak{o}}$, and*
- *the* ***incentive*** *precondition is met if $\mathfrak{o}$ must learn c to remain "fit".*

(intuitive summary) *If c is a 1ST order self, then these are the preconditions that must be met for an organism to construct c. Likewise, any other sort of causal identity.*

Definition 15 (chain notation)**.**
Suppose we have two organisms, $\mathfrak{a}$ (Alice) and $\mathfrak{b}$ (Bob). $c_\mathfrak{a}^\mathfrak{b}$ denotes a causal identity for $\mathfrak{b}$ constructed by $\mathfrak{a}$ (what Alice thinks Bob intends). Subscript denotes the organism who constructs the causal identity, while superscript denotes the object. The superscript can be extended to denote chains of predicted causal identity. For example, $c_\mathfrak{a}^{\mathfrak{ba}} \subset c_\mathfrak{a}^\mathfrak{b}$ denotes $\mathfrak{a}$'s prediction of $\mathfrak{b}$'s prediction of $\mathfrak{a}^1$ (what Alice thinks Bob thinks Alice intends). The superscript of $c_\mathfrak{a}^*$ can be extended indefinitely to indicate recursive predictions; however the extent recursion is possible is determined by $\mathfrak{a}$'s vocabulary $\mathfrak{v}_\mathfrak{a}$. Finally, Bob need not be an organism. Bob can be anything for which Alice constructs a causal identity.

Definition 16 ($\mathrm{n}^{\mathbf{th}}$ order self)**.**
An n^{th} order self for $\mathfrak{a}$ is $\mathfrak{a}^\mathfrak{n} = c_\mathfrak{a}^{*\mathfrak{a}}$ where $*$ is replaced by a chain, and n denotes the number of reflections. For example, a 2ND order self $\mathfrak{a}^2 = c_\mathfrak{a}^{\mathfrak{ba}}$, and a 3RD order self $\mathfrak{a}^3 = c_\mathfrak{a}^{\mathfrak{baba}}$. We use $\mathfrak{a}^2$ to refer to any 2ND order self, and chain notation to refer to a specific 2ND order self, for example $c_\mathfrak{a}^{\mathfrak{ba}}$. The union of two n^{th} order selves is also considered to be an n^{th} order self, for example $\mathfrak{a}^3 = c_\mathfrak{a}^{\mathfrak{baba}} \cup c_\mathfrak{a}^{\mathfrak{dada}}$, and the weaker or higher level a self is in the generational hierarchy, the more selves there are of which it is part.

Definition 17 (stages of consciousness)**.**
We argue the following stages by scaling the ability to learn weak policies:

1. ***Hard Coded:*** *organism that acts but does not learn, meaning $\mathfrak{p}_\mathfrak{o}$ is fixed from birth.*
2. ***Learning:*** *an organism that learns, but $\mathfrak{o}^1 \notin \mathfrak{p}_\mathfrak{o}$ either because $\mathfrak{o}^1 \notin L_{\mathfrak{v}_\mathfrak{o}}$ (failing the "scale precondition") or because the organism is not incentivised to construct $\mathfrak{o}^1$ (failing the "incentive precondition").*
3. ***1ST order self:*** *reafference and phenomenal or core consciousness are achieved when $\mathfrak{o}^1 \in \mathfrak{p}_\mathfrak{o}$ is learned by an organism because of attraction to and repulsion from statements in $L_{\mathfrak{v}_\mathfrak{o}}$.*
4. ***2ND order selves:***
   (a) *access or self-reflexive consciousness is achieved when $\mathfrak{o}^2 \in \mathfrak{p}_\mathfrak{o}$.*
   (b) *hard consciousness is achieved when a phenomenally conscious organism learns a 2ND order self (an organism is consciously aware of the contents of 2ND order selves, which must have quality if learned through phenomenal conscious).*
5. ***3RD and higher order selves:*** *meta self-reflexive consciousness (human level hard consciousness) is achieved when $\mathfrak{o}^3 \in \mathfrak{p}_\mathfrak{o}$.*

10.1 Proofs

Versions of proofs 1-3 were published in [38]. Further details are available in on GitHub, along with the supporting experiments [40].

Proposition 1 (sufficiency)**.**
Assume $\alpha \sqsubset \omega$. The weakness proxy sufficient to maximise the probability that a parent ω is learned from a child α[21].

Proof. You're given the definition of $\mathfrak{v}$-task α from which you infer a hypothesis $\mathbf{h} \in \Pi_\alpha$. To learn ω, you need $\mathbf{h} \in \Pi_\omega$:

1. For every $\mathbf{h} \in \Pi_\alpha$ there exists a $\mathfrak{v}$-task $\gamma_\mathbf{h} \in \Gamma_\mathfrak{v}$ s.t. $O_{\gamma_\mathbf{h}} = E_\mathbf{h}$, meaning $\mathbf{h}$ permits only correct outputs for that task regardless of input. We'll call the highest level task $\gamma_\mathbf{h}$ s.t. $O_{\gamma_\mathbf{h}} = E_\mathbf{h}$ the **policy task** of $\mathbf{h}$[22].
2. ω is either the policy task of a policy in Π_α, or a child thereof.
3. If a policy $\mathbf{h}$ is correct for a parent of ω, then it is also correct for ω. Hence we should choose $\mathbf{h}$ that has a policy task with the largest number of children. As tasks are uniformly distributed, that will maximise the probability that ω is $\gamma_\mathbf{h}$ or a child thereof.
4. For the purpose of this proof, we say one task is **equivalent**[23] to another if it has the same correct outputs.
5. No two policies in Π_α have the same policy task[24]. This is because all the policies in Π_α are derived from the same set inputs, I_α.
6. The set of statements which *might* be outputs addressing inputs in I_ω and not I_α, is $\overline{E_{I_\alpha}} = \{l \in L_\mathfrak{v} : l \notin E_{I_\alpha}\}$[25].
7. For any given $\mathbf{h} \in \Pi_\alpha$, the extension $E_\mathbf{h}$ of $\mathbf{h}$ is the set of outputs $\mathbf{h}$ implies. The subset of $E_\mathbf{h}$ which fall outside the scope of what is required for the known task α is $\overline{E_{I_\alpha}} \cap E_\mathbf{h}$[26].
8. $|\overline{E_{I_\alpha}} \cap E_\mathbf{h}|$ increases monotonically with $|E_\mathbf{h}|$, because for all $e \in E_\mathbf{h}$ is $e \notin \overline{E_{I_\alpha}}$ iff $e \in E_{I_\alpha}$.
9. $2^{|\overline{E_{I_\alpha}} \cap E_\mathbf{h}|}$ is the number of non-equivalent **parents** of α to which $\mathbf{h}$ generalises. It increases monotonically with the weakness of $\mathbf{h}$.
10. Given $\mathfrak{v}$-tasks are uniformly distributed and $\Pi_\alpha \cap \Pi_\omega \neq \emptyset$, the probability that $\mathbf{h} \in \Pi_\alpha$ generalises to ω is
$$p(\mathbf{h} \in \Pi_\omega \mid \mathbf{h} \in \Pi_\alpha, \alpha \sqsubset \omega) = \frac{2^{|\overline{E_{I_\alpha}} \cap E_\mathbf{h}|}}{2^{|\overline{E_{I_\alpha}}|}}$$

$p(\mathbf{h} \in \Pi_\omega \mid \mathbf{h} \in \Pi_\alpha, \alpha \sqsubset \omega)$ is maximised when $|E_\mathbf{h}|$ is maximised. Recall from definition 4 that $<_w$ is the **weakness** proxy. For statements l_1, l_2 we have $l_1 <_w l_2$ iff

[21] Assume there exist correct policies for ω, because otherwise there would be no point in trying to learn it.

[22] I'd like to give credit here to Nora Belrose for pointing out an error. Nora pointed out I was miscounting the number of tasks. As a result I realised I was not counting tasks, I was in fact counting policy tasks and had entirely neglected to mention this fact. This was a significant error which has now been corrected, with several additional steps added to account for equivalence.

[23] This is because switching from β to ζ s.t. $I_\beta \neq I_\zeta$ and $O_\beta = O_\zeta$ would be to pursue the same goal in different circumstances. This is because inputs are *subsets* of outputs, so both sets of inputs are implied by the outputs. O_ζ implies I_β and O_β implies I_ζ

[24] Every policy task for policies of α is non-equivalent from the others.

[25] This is because E_{I_α} contains every statement which is a correct output or an incorrect output, and $\overline{E_{I_\alpha}}$ contains every statement which could possibly be in I_ω, E_{I_ω} and thus O_ω.

[26] This is because E_{I_α} is the set of all conceivable outputs by which one might attempt to complete α, and so the set of all outputs that can't be made when undertaking α is $\overline{E_{I_\alpha}}$ because those outputs occur given inputs that aren't part of I_α.

$|E_{l_1}| < |E_{l_2}|$. $\mathbf{h}$ that maximises $<_w$ will also maximise $p(\mathbf{h} \in \Pi_\omega \mid \mathbf{h} \in \Pi_\alpha, \alpha \sqsubset \omega)$. Hence the weakness proxy maximises the probability that[27] a parent ω is learned from a child α. □ □

Proposition 2 (necessity)**.**
To maximise the probability of learning ω from α, it is necessary to use weakness as a proxy.

Proof. Let α and ω be defined exactly as they were in the proof of prop. 1.

1. If $\mathbf{h} \in \Pi_\alpha$ and $E_{I_\omega} \cap E_\mathbf{h} = O_\omega$, then it must be he case that $O_\omega \subseteq E_\mathbf{h}$.
2. If $|E_\mathbf{h}| < |O_\omega|$ then generalisation cannot occur, because that would mean that $O_\omega \not\subseteq E_\mathbf{h}$.
3. Therefore generalisation is only possible if $|E_\mathbf{h}| \geq |O_\omega|$, meaning a sufficiently weak hypothesis is necessary to generalise from child to parent.
4. For any two hypotheses $\mathbf{h}_1$ and $\mathbf{h}_2$, if $\mid E_{\mathbf{h}_1} \mid < \mid E_{\mathbf{h}_2} \mid$ then the probability $p(\left|E_{\mathbf{h}_1}\right| \geq |O_\omega|) < p(\left|E_{\mathbf{h}_2}\right| \geq |O_\omega|)$ because tasks are uniformly distributed.
5. Hence the probability that $|E_m| \geq |O_\omega|$ is maximised when $|E_m|$ is maximised. To maximise the probability of learning ω from α, it is necessary to select the weakest hypothesis.

To select the weakest hypothesis, it is necessary to use the weakness proxy. □ □

Proposition 3 (simplicity sub-optimality)**.**
Description length is neither a necessary nor sufficient proxy for the purposes of maximising the probability that induction generalises.

Proof. In propositions 1 and 2 we proved that weakness is a necessary and sufficient choice of proxy to maximise the probability of generalisation. It follows that either maximising $\frac{1}{|m|}$ (minimising description length) maximises $|E_m|$ (weakness), or minimisation of description length is unnecessary to maximise the probability of generalisation. Assume the former, and we'll construct a counterexample with $\mathfrak{v} = \{a, b, c, d, e, f, g, h, j, k, z\}$ s.t. $L_\mathfrak{v} = \{\{a, b, c, d, j, k, z\}, \{e, b, c, d, k\}, \{a, f, c, d, j\}, \{e, b, g, d, j, k, z\}, \{a, f, c, h, j, k\}, \{e, f, g, h, j, k\}\}$ and a task α where

- $I_\alpha = \{\{a, b\}, \{e, b\}\}$
- $O_\alpha = \{\{a, b, c, d, j, k, z\}, \{e, b, g, d, j, k, z\}\}$
- $\Pi_\alpha = \{\{z\}, \{j, k\}\}$

Weakness as a proxy selects $\{j, k\}$, while description length as a proxy selects $\{z\}$. This demonstrates the minimising description length does not necessarily maximise weakness, and maximising weakness does not minimise description length. As weakness is necessary and sufficient to maximise the probability of generalisation, it follows that minimising description length is neither. □

□

[27]Subsequently it also maximises the sample efficiency with which a parent ω is learned from a child α.

The following sketch is more for intuition and explanation than anything formal.

Proposition 4. *An organism that uses weakness as its proxy will learn an n^{th} order self if the incentive and scale preconditions are met for that order of self.*

Proof sketch. Assume we have an organism $\mathfrak{a}$ that learns using "weakness" as a proxy. A $\mathfrak{v}_\mathfrak{a}$-task $\mathfrak{h}_{<t_\mathfrak{a}}$ represents the history of $\mathfrak{a}$ (meaning $\mathfrak{h}_{<t_\mathfrak{a}} \sqsubset \mu_\mathfrak{a}$ and $\mathfrak{h}_{<t_\mathfrak{a}}$ is an ostensive definition of $\mu_\mathfrak{a}$, because $\mathfrak{a}$ remains alive). The organism explores the environment, intervening to maintain homeostasis. As it does so, more and more inputs and outputs are included in $\mathfrak{h}_{<t_\mathfrak{a}}$. It follows that:

1. From the *scale* precondition we have that there exists a n^{th} order self $\mathfrak{a}^\mathfrak{n} \in L_{\mathfrak{v}_\mathfrak{a}}$.
2. To remain fit, $\mathfrak{a}$ must "generalise" to $\mu_\mathfrak{a}$ from $\mathfrak{h}_{<t_\mathfrak{a}}$. According to the *incentive* precondition, generalisation to $\mu_\mathfrak{a}$ requires $\mathfrak{a}$ learn the n^{th} order self, which is when $\mathfrak{a}^\mathfrak{n} \in \mathfrak{p}_\mathfrak{a}$.
3. From [38] we have proof that weakness is the optimal choice of proxy to maximise the probability of generalisation from child to parent is the *weakest* policy. It follows that $\mathfrak{a}$ will generalise from $\mathfrak{h}_{<t_\mathfrak{a}}$ to $\mu_\mathfrak{a}$ given the smallest history of interventions with which it is possible to do so (meaning the smallest possible ostensive definition, or cardinality $|D_\alpha|$).

Were we to assume learning under the above conditions *does not* construct an n^{th} order self for $\mathfrak{a}$, then one of the three statements above would be false and we would have a contradiction. It follows that the proposition must be true. □

References

[1] Nagel, T.: What is it like to be a bat? Philosophical Review **83**(October), 435–50 (1974) https://doi.org/10.2307/2183914

[2] Chalmers, D.: Facing up to the problem of consciousness. Journal of Consciousness Studies **2**(3), 200–19 (1995)

[3] Ashby, W.R.: Principles of the self-organizing dynamic system. The Journal of General Psychology **37**(2), 125–128 (1947)

[4] Foerster, H.: On self-organizing systems and their environments. In: Self-Organizing Systems, pp. 31–50. Pergamon Press., ??? (1960)

[5] Haken, H.: Advanced Synergetics: Instability Hierarchies of Self-Organizing Systems and Devices. Springer, Berlin Heidelberg (1983)

[6] Camazine, S.: Patterns in nature. Natural history **112**, 34–41 (2003)

[7] Bell, M.A., Deater-Deckard, K.: Biological systems and the development of self-regulation: Integrating behavior, genetics, and psychophysiology. Journal

of developmental and behavioral pediatrics : JDBP **28**, 409–20 (2007) https://doi.org/10.1097/DBP.0b013e3181131fc7

[8] Kelso, S.: Dynamic Patterns: The Self-Organization of Brain and Behavior. MIT Press, Boston (1997)

[9] Friston, K.: The free-energy principle: a unified brain theory? Nature Reviews Neuroscience **11**(2), 127–138 (2010) https://doi.org/10.1038/nrn2787

[10] Tognoli, E., Kelso, J.A.S.: Enlarging the scope: grasping brain complexity. Front Syst Neurosci **8**, 122 (2014)

[11] Camazine, S., Franks, N., Sneyd, J., Bonabeau, E., Deneubourg, J.-L., Theraulaz, G.: Self-Organization in Biological Systems. Princeton University Press, NJ (2001)

[12] Seeley, T.D.: When is self-organization used in biological systems? The Biological Bulletin **202**(3), 314–318 (2002)

[13] Rosas, F., Mediano, P.A.M., Ugarte, M., Jensen, H.J.: An information-theoretic approach to self-organisation: Emergence of complex interdependencies in coupled dynamical systems. Entropy **20**(10) (2018)

[14] Seth, A., Bayne, T.: Theories of consciousness. Nature Reviews Neuroscience **23** (2022) https://doi.org/10.1038/s41583-022-00587-4

[15] Block, N.: On a confusion about a function of consciousness. Brain and Behavioral Sciences **18**(2), 227–247 (1995) https://doi.org/10.1017/s0140525x00038188

[16] Gallagher, S., Zahavi, D.: The Phenomenological Mind. Routledge, New York, NY (2021)

[17] Fuchs, T.: Ecology of the Brain: The Phenomenology and Biology of the Embodied Mind. Oxford University Press, ??? (2017). https://doi.org/10.1093/med/9780199646883.001.0001 . https://doi.org/10.1093/med/9780199646883.001.0001

[18] Northoff, G.: Unlocking The Brain, Vol. II: Consciousness vol. 2. Oxford University Press, USA (2014)

[19] Goff, P.: Galileo's Error: Foundations for a New Science of Consciousness. Rider, UK (2019). Also published by Pantheon in the US

[20] Dennett, D.C.: Consciousness Explained. Little, Brown and Co., Boston (1991)

[21] Chalmers, D.J.: The Conscious Mind: In Search of a Fundamental Theory (2nd Edition). Oxford University Press, ??? (1996)

[22] Jackson, F.: Epiphenomenal qualia. The Philosophical Quarterly **32**(127), 127–136 (1982)

[23] Jackson, F.: What mary didn't know. The Journal of Philosophy **83**(5), 291–295 (1986)

[24] Rosenthal, D.M.: Consciousness and Mind. Oxford University Press UK, New York (2005)

[25] Brown, R., Lau, H., LeDoux, J.E.: Understanding the higher-order approach to consciousness. Trends in Cognitive Sciences **23**(9), 754–768 (2019) https://doi.org/10.1016/j.tics.2019.06.009

[26] Varela, F., Thompson, E., Rosch, E., Kabat-Zinn, J.: The Embodied Mind: Cognitive Science and Human Experience, pp. 1–322 (2016). https://doi.org/10.7551/mitpress/9780262529365.001.0001

[27] Ciaunica, A., Shmeleva, E.V., Levin, M.: The brain is not mental! coupling neuronal and immune cellular processing in human organisms. Frontiers in Integrative Neuroscience **17** (2023) https://doi.org/10.3389/fnint.2023.1057622

[28] McMillen, P., Levin, M.: Collective intelligence: A unifying concept for integrating biology across scales and substrates. Communications Biology **7**(1), 378 (2024)

[29] Seth, A.K., Tsakiris, M.: Being a beast machine: The somatic basis of selfhood. Trends Cogn Sci **22**(11), 969–981 (2018)

[30] Blum, M., Blum, L.: A theoretical computer science perspective on consciousness. J. Artif. Intell. Conscious. **8**, 1–42 (2020)

[31] Baars, B.: In the theater of consciousness: The workspace of the mind (1997) https://doi.org/10.1093/acprof:oso/9780195102659.001.1

[32] Wang, P.: A constructive explanation of consciousness. Journal of Artificial Intelligence and Consciousness **07**(02), 257–275 (2020) https://doi.org/10.1142/S2705078520500125 https://doi.org/10.1142/S2705078520500125

[33] Thompson, E.: Mind in Life: Biology, Phenomenology, and the Sciences of Mind. Harvard University Press, Cambridge MA (2007)

[34] Piccinini, G.: Physical Computation: A Mechanistic Account. Oxford University Press, UK (2015)

[35] Piccinini, G., Maley, C.: Computation in Physical Systems. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy, Sum. 21 edn. Stanford University, Stanford (2021)

[36] Bennett, M.T.: Computational dualism and objective superintelligence. In:

Thórisson, K.R., Isaev, P., Sheikhlar, A. (eds.) Artificial General Intelligence, pp. 22–32. Springer, Cham (2024)

[37] Sutton, R.: The bitter lesson. University of Texas at Austin (2019)

[38] Bennett, M.T.: The optimal choice of hypothesis is the weakest, not the shortest. In: Hammer, P., Alirezaie, M., Strannegård, C. (eds.) Artificial General Intelligence, pp. 42–51. Springer, Cham (2023)

[39] Bennett, M.T.: Emergent causality and the foundation of consciousness. In: Hammer, P., Alirezaie, M., Strannegård, C. (eds.) Artificial General Intelligence, pp. 52–61. Springer, Cham (2023)

[40] Bennett, M.T.: Technical Appendices. https://doi.org/10.5281/zenodo.7641741 . https://github.com/ViscousLemming/Technical-Appendices Accessed 2024-12-10

[41] Bennett, M.T.: Is complexity an illusion? In: Thórisson, K.R., Isaev, P., Sheikhlar, A. (eds.) Artificial General Intelligence, pp. 11–21. Springer, Cham (2024)

[42] Howard, W.A.: The Formulae-as-Types Notion of Construction. In: Seldin, J.P., Hindley, J.R. (eds.) To H.B. Curry: Essays on Combinatory Logic, Lambda Calculus and Formalism, pp. 479–490. Academic Press, Cambrdige MA (1980)

[43] Bennett, M.T.: Are Biological Systems More Intelligent Than Artificial Intelligence? Manuscript under review (2024)

[44] Hutter, M.: Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability. Springer, Berlin, Heidelberg (2010)

[45] Leike, J., Hutter, M.: Bad universal priors and notions of optimality. Proceedings of The 28th Conference on Learning Theory, in Proceedings of Machine Learning Research, 1244–1259 (2015)

[46] Leike, J., Hutter, M.: On the computability of solomonoff induction and aixi. Theoretical Computer Science **716**, 28–49 (2018) https://doi.org/10.1016/j.tcs.2017.11.020 . Special Issue on ALT 2015

[47] Rolla, G., Figueiredo, N.: Bringing forth a world, literally. Phenomenology and the Cognitive Sciences, 1–23 (2021) https://doi.org/10.1007/s11097-021-09760-z

[48] Vervaeke, J., Lillicrap, T., Richards, B.: Relevance realization and the emerging framework in cognitive science. J. Log. Comput. **22**, 79–99 (2012) https://doi.org/10.1093/logcom/exp067

[49] Vervaeke, J., Ferraro, L.: In: Ferrari, M., Weststrate, N.M. (eds.) Relevance, Meaning and the Cognitive Science of Wisdom, pp. 21–51. Springer, Dordrecht (2013)

[50] Vervaeke, J., Ferraro, L.: Relevance realization and the neurodynamics and neuroconnectivity of general intelligence. In: Harvey, I., Cavoukian, A., Tomko, G., Borrett, D., Kwan, H., Hatzinakos, D. (eds.) SmartData, pp. 57–68. Springer, New York, NY (2013)

[51] Jaeger, J., Riedl, A., Djedovic, A., Vervaeke, J., Walsh, D.: Naturalizing relevance realization: Why agency and cognition are fundamentally not computational. Frontiers in Psychology **15** (2024)

[52] Savage, L.J.: The Foundations of Statistics. John Wiley & Sons, NY, USA (1954)

[53] Friston, K., Da Costa, L., Sakthivadivel, D.A.R., Heins, C., Pavliotis, G.A., Ramstead, M., Parr, T.: Path integrals, particular kinds, and strange things. Physics of Life Reviews **47**, 35–62 (2023) https://doi.org/10.1016/j.plrev.2023.08.016

[54] Solms, M.: The Hidden Spring. Profile Books, London (2021)

[55] Dreyfus, H.L.: Why heideggerian ai failed and how fixing it would require making it more heideggerian. Philosophical Psychology **20**(2), 247–268 (2007) https://doi.org/10.1080/09515080701239510

[56] Bennett, M.T.: On the computation of meaning, language models and incomprehensible horrors. In: Hammer, P., Alirezaie, M., Strannegård, C. (eds.) Artificial General Intelligence, pp. 32–41. Springer, Cham (2023)

[57] Floridi, L., Chiriatti, M.: Gpt-3: Its nature, scope, limits, and consequences. Minds and Machines, 1–14 (2020)

[58] Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach, 4th Edition. Prentice Hall, Hoboken (2020)

[59] Bishop, C.M.: Pattern Recognition and Machine Learning, pp. 1122–1128. Springer, NY (2006)

[60] Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. MIT press, MA (2018)

[61] Goertzel, B.: Artificial general intelligence: Concept, state of the art. Journal of Artificial General Intelligence **5**(1), 1–48 (2014)

[62] Pearl, J., Mackenzie, D.: The Book of Why: The New Science of Cause and Effect, 1st edn. Basic Books, Inc., New York (2018)

[63] Richens, J., Everitt, T.: Robust agents learn causal world models. In: The Twelfth International Conference on Learning Representations (2024). https://openreview.net/forum?id=pOoKI3ouv1

[64] Pearl, J.: Causality, 2nd edn. Cambridge Uni. Press, United Kingdom (2009)

[65] Thompson, E.: Précis of waking, dreaming, being: Self and consciousness in neuroscience, meditation, and philosophy. Philosophy East and West **66**(3), 927–933 (2016) https://doi.org/10.1353/pew.2016.0059

[66] Dawid, A.P.: Influence diagrams for causal modelling and inference. International Statistical Review / Revue Internationale de Statistique **70**(2), 161–189 (2002). Accessed 2024-02-22

[67] Bennett, M.T., Maruyama, Y.: Philosophical specification of empathetic ethical artificial intelligence. IEEE Transactions on Cognitive and Developmental Systems **14**(2), 292–300 (2022)

[68] Atkin, A.: Peirce's Theory of Signs. In: Zalta, E.N., Nodelman, U. (eds.) The Stanford Encyclopedia of Philosophy, Spring 2023 edn. Metaphysics Research Lab, Stanford University, ??? (2023)

[69] Goertzel, B.: The Hidden Pattern: A Patternist Philosophy of Mind. Brown-Walker Press, USA (2006)

[70] Ciaunica, A., Crucianelli, L.: Minimal self-awareness: From within a developmental perspective. Journal of Consciousness Studies **26**(3-4), 207–226 (2019)

[71] Merker, B.: Consciousness without a cerebral cortex: A challenge for neuroscience and medicine. Behavioral and Brain Sciences **30**(1), 63–81 (2007) https://doi.org/10.1017/S0140525X07000891

[72] Barron, A.B., Klein, C.: What insects can tell us about the origins of consciousness. Proceedings of the National Academy of Sciences **113**(18), 4900–4908 (2016) https://doi.org/10.1073/pnas.1520084113 https://www.pnas.org/doi/pdf/10.1073/pnas.1520084113

[73] Solé, R., Moses, M., Forrest, S.: Liquid brains, solid brains. Philosophical Transactions of the Royal Society B: Biological Sciences **374**(1774), 20190040 (2019) https://doi.org/10.1098/rstb.2019.0040 https://royalsocietypublishing.org/doi/pdf/10.1098/rstb.2019.0040

[74] Grice, P.: Meaning. The Philosophical Review **66**(3), 377–388 (1957)

[75] Grice, P.: Utterer's meaning and intention. The Philosophical Review **78**(2), 147–177 (1969)

[76] Morin, A.: Levels of consciousness and self-awareness: A comparison and integration of various neurocognitive views. Consciousness and Cognition **15**(2), 358–371 (2006) https://doi.org/10.1016/j.concog.2005.09.006

[77] Alexander, S.A., Castaneda, M., Compher, K., Martinez, O.: Extending environments to measure self-reflection in reinforcement learning. Journal of Artificial

General Intelligence **13**(1), 1–24 (2022)

[78] Fields, C., Levin, M.: Scale-free biology: Integrating evolutionary and developmental thinking. BioEssays **42** (2020)

[79] Bongard, J., Levin, M.: There's plenty of room right here: Biological systems as evolved, overloaded, multi-scale machines. Biomimetics **8**(1) (2023)

[80] Merker, B.: The liabilities of mobility: A selection pressure for the transition to consciousness in animal evolution. Consciousness and Cognition **14**(1), 89–114 (2005) https://doi.org/10.1016/S1053-8100(03)00002-3 . Neurobiology of Animal Consciousness

[81] Boltuc, P.: The engineering thesis in machine consciousness. Techné: Research in Philosophy and Technology **16**(2), 187–207 (2012)

[82] Shanahan, M.: The brain's connective core and its role in animal cognition. Philosophical transactions of the Royal Society of London. Series B, Biological sciences **367**, 2704–14 (2012) https://doi.org/10.1098/rstb.2012.0128

[83] Ball, P.: Materials with agency. Nature Materials **22**(3), 272 (2023)

[84] Levin, M.: Bioelectrical approaches to cancer as a problem of the scaling of the cellular self. Progress in Biophysics and Molecular Biology **165**, 102–113 (2021). Cancer and Evolution

[85] Heylighen, F.: The meaning and origin of goal-directedness: a dynamical systems perspective. Biological Journal of the Linnean Society **139**(4), 370–387 (2022)

[86] Bielecki, J., Dam Nielsen, S.K., Nachman, G., Garm, A.: Associative learning in the box jellyfish tripedalia cystophora. Current Biology (2023) https://doi.org/10.1016/j.cub.2023.08.056

[87] Yu, A.J., Rankin, C.H.: In: Krause, M.A., Hollis, K.L., Papini, M.R.E. (eds.) Learning and Memory in the Nematode Caenorhabditis elegans, pp. 15–32. Cambridge University Press, ??? (2022). https://doi.org/10.1017/9781108768450.004

[88] Willett, D.S., Alborn, H.T., Stelinski, L.L., Shapiro-Ilan, D.I.: Risk taking of educated nematodes. PLOS ONE **13**(10), 1–10 (2018) https://doi.org/10.1371/journal.pone.0205804

[89] Lüersen, K., Faust, U., Gottschling, D.-C., Döring, F.: Gait-specific adaptation of locomotor activity in response to dietary restriction in caenorhabditis elegans. Journal of Experimental Biology **217**(14), 2480–2488 (2014)

[90] Artyukhin, A.B., Yim, J.J., Cheong Cheong, M., Avery, L.: Starvation-induced collective behavior in c. elegans. Scientific reports **5**(1), 10647 (2015)

[91] Wehner, R.: Life as a cataglyphologist—and beyond. Annual review of entomology **58**(1), 1–18 (2013)

[92] Seeley, T.: The wisdom of the hive cambridge. MA: Harvard University Press [Google Scholar] (1995)

[93] Oades, R.D., Isaacson, R.L.: The development of food search behavior by rats: the effects of hippocampal damage and haloperidol. Behavioral biology **24**(3), 327–337 (1978)

[94] Bugnyar, T., Kotrschal, K.: Observational learning and the raiding of food caches in ravens, corvus corax: is it 'tactical' deception? Animal Behaviour **64**(2), 185–195 (2002) https://doi.org/10.1006/anbe.2002.3056

[95] Joel Crampton, D.A.P. Celine H. Frère: Australian magpies gymnorhina tibicen cooperate to remove tracking devices. Australian Field Ornithology **39**, 7–11 (2022) https://doi.org/10.20938/afo39007011

[96] Carruth, A.: Powerful qualities, zombies and inconceivability. Philosophical Quarterly **66**(262), 25–46 (2016) https://doi.org/10.1093/pq/pqv055