A Modal Approach to Consciousness of Agents

Chen Yifeng[1] and J. W. Sanders[2(✉)]

[1] Peking University, Beijing, China
cyf@pku.edu.cn
[2] AIMS South Africa, Cape Town, Republic of South Africa
jsanders@aims.ac.za

Abstract. An agent's awareness is modelled as a modal operator in such a way that awareness can be iterated and consciousness formalised as awareness of awareness. Agents are not necessarily human and may *a priori* be animals, organisations or software, in which setting awareness is expected to exist in degrees and so is modelled with nonnegative reals rather than just Booleans. The formalism thus expresses the degree to which an agent exhibits awareness (and so consciousness).

The context is an adaptive multi-agent system in which agents control actions, individually or in groups, and adapt ecorithmically (in the sense of Valiant) by adjusting behaviour in the short term and evolving in the very much longer term. Laws and inequalities are given and shown to be sound, but the intuition is that *awareness* 'enables' actions to form the agent's next behavioural step whilst *consciousness* provides the agent with an opportunity to adapt that behaviour.

1 Introduction

Consciousness has for long been considered beyond scientific explanation (*i.e.* not to be explicable by reduction) and instead to be an emergent property of that complex system the human brain. The fraught problem of understanding consciousness has been made no simpler by that concept cutting across neurophysiology, philosophy of the mind, physics, computer science, data science and more recently mathematics. But recent decades have heralded a fresh approach: the proposal of architectural models to account for consciousness.[1] Together with the success of machine learning (providing a candidate for artificial free will?), that has led to renewed interest in both the popular press and academic journals in the contentious question of whether or not an artificial agent can be sentient or conscious. Without a definition or even agreed properties of consciousness, how can that be answered? This work addresses that deficiency, in a modal setting.

A treatment of consciousness which does not *a priori* rule out the possibility of its application to non-humans must be general enough to embrace organisms

[1] A dozen such models are cited at the *Oxford Mathematics of Consciousness and Applications Network* site [19].

T. Margaria and B. Steffen (Eds.): ISoLA 2022, LNCS 13703, pp. 127–141, 2022.
https://doi.org/10.1007/978-3-031-19759-8_9

(like cells, plants and animals), organisations and artificial agents yet not be too weak when restricted to humans. For such entities we use the generic term 'agent'. We propose a Boolean notion (an agent is conscious of a feature or not), and a numerical notion (the strength of that consciousness).

We follow the usual approach when confronted with a complex concept and resort to identifying properties, or laws if possible, in place of a definition. Of course a model is still needed to show consistency of the laws, and we use as simple a model as possible. The aim of that approach is eventually to identify sufficiently many laws to characterise the concept. In the case of consciousness, where no definition seems forthcoming, it offers an enticing avenue for progress.

Our choice of laws is guided by the following intuition. An agent is aware of something that 'enables' or 'makes executable' actions under its control for use in its next step in behaviour. For instance a bird flying to its nest is aware of winds if they cause it to adjust its flight. On the other hand a person whose senses are not augmented by an appropriate receiver is unaware of radio waves since their presence 'enables' no actions within its control.

If an agent is aware of something then in some cases, identified here as those in which the agent is conscious of the thing, it uses that awareness to adapt its actions. Thus consciousness requires awareness but provides more: an opportunity for adapting the way in which the next step in behaviour is chosen. Unfortunately a definition of consciousness in those terms directly would not be observable without insight into the agent's 'mind'. So we resort to defining it in terms of iterated awareness. Thus: an agent is conscious if it is aware of its awareness (The Stanford Encyclopedia, [24]: Sect. 9.2, Reflexive Theories).

For instance a bird is conscious of fledgelings in the nest because it does not return directly to the nest, as usual, but adapts by landing first on a nearby branch. It is conscious of the flock, because it adapts its trajectory by averaging the velocities of its neighbours in the flock [5]. Thus it is not merely aware but conscious in both cases.

That intuition extends to agents many popular treatments of human consciousness. We refer to just one, by Dehaene [6], which takes human consciousness of something to mean 'the ability to report on it'. In our terms, reporting requires awareness of the thing to enable the ability to report on it, but moreover in choosing what to report the person demonstrates consciousness of it.

In this approach consciousness is necessary for adaptation, for which we follow the *ecorithm* approach of Valiant who makes a convincing case that ecorithms embody:

> ... mechanisms ... of two kinds, those that operate in individuals interacting with their environment, and those that operate via genetic changes over many generations. ... ecorithms are also construed broadly enough to encompass any process of interaction with an environment.
>
> Valiant [26], page 27.

In our context short-term adaptivity can be seen as adjustment by the agent to its environment, and long-term as evolution. As the system evolves, changes occur to the set of agents, the actions and their control by agents.

Laws of consciousness for agents must be decided without recourse to the concept of 'an agent's state of mind', and more generally eschewing anthropomorphism. Our solution is to resort, as much as possible (though not entirely), to externally observable behaviour. For instance we acknowledge that a pet dog is aware of its lead being taken from the peg in preparation for its daily walk because we observe that it wags its tail and rushes to the door. Naturally we refrain from postulating 'the dog is happy' (its state of mind).

The paper begins with our context of agents, actions, features and adaptive multi-agent systems. It motivates properties of awareness and expresses them in both Boolean and numerical forms. Then it formalises awareness and consciousness and proves soundness of the laws. After analysing properties of our adaptive mulit-agent systems, it discusses related work and draws conclusions.

2 Conscious Agents

This section provides the background to our general view of agents, the actions they perform and the features of which they may be aware. It then discusses the adaptive multi-agent systems they inhabit.

2.1 Agents

The agents we consider range from humans, other animals, plants, cells and organisations to software. They are considered not in isolation but as part of some habitat[2] which may be inhabited by various agents, but has an external environment. For instance in the local gardens we may consider birds and the things which affect them (like trees and worms). Birds exhibit a strong circadian rhythm which they exploit when deciding how to behave, but the sun and its movement which affect bird behaviour are external to the garden system. The external environment is treated as a default agent.

System actions are typically controlled by agents, individually or in groups. Care of fledgelings lies under the control of their parents whilst flocking is controlled by a group. Sunrise is controlled by the default environmental agent.

An agent is an entity in such a system having control over at least one action. The agent may be a sunflower which when growing exhibits heliotropic behaviour by tracking the sun during the day and then reorienting overnight to face east. Its movement is the combined effect of internal and external actions which result in the head tilting due to cells growing faster on the side of the stem facing the sun. Actions under the sunflower's control include the hormonal and circadian actions controlled by the plant but not solar movement; see Atamian *et al.* [1].

A rock in the garden erodes at a rate which depends on its location and composition, but as a result of action by the elements. Erosion is thus the result of environmental actions, and none under control of the rock, which is therefore not an agent.

[2] The term 'environment' is more commonly used to mean something external to an agent, but we are about to give 'external environment' another meaning.

Due to the generality of an agent, we cannot assume that it displays the kind of rationality assumed in logic and in particular dynamic epistemic logic. We cannot justify for instance the law that if an agent is aware of p which is stronger than q then it is also aware of q. However something of that kind holds, but with correlation instead. The pet dog is evidently conscious, from its behaviour already considered, that when its lead is taken from the peg then it is daily-walk time. We capture properties of consciousness with laws, but they are far weaker than the familiar logical laws.

2.2 Actions

The actions performed by agents either terminate or on-going and typically reactive. The former are described by postcondition with state-before, input, state-after and output. Then the precondition is defined to hold at a state-before and input if there is a state-after and output satisfying the postcondition. The latter types of action are described by safety and liveness. We try not to distinguish the two styles, thinking of an ongoing action as the iteration of a terminating action, perhaps forever.

Our descriptions of actions are not necessarily algorithmic nor even computable, but they are all state based. That allows inclusion of the view by Penrose & Hameroff (see the review by Hameroff [12]) that quantum reduction is primitive in any appropriate ecorithmic language for humans.

We use the following notation concerning an agent's control of actions. Left informal here, it has been formalised for software [21]. Suppose as given the sets *Agents* and *Actions*.

Notation 1 (Ambit)**.** *The* ambit *of an action act* : *Actions is the set of agents involved in its activation:*

$$ambit(act) := \{a : Agents \mid a \text{ has some control in } act\}.$$

The set of actions in which a : *Agents has control is denoted*

$$\mathcal{A}_a := \{act : Actions \mid a \in ambit(act)\}.$$

For instance the ambit of a bird's return flight to its nest contains itself, and weather conditions. The ambit of its flight when flocking contains its nearest neighbours in the flock.

2.3 Features

The things in an agent's habitat which may affect its behaviour we call *features*.

For instance features in a human's habitat may include memory of birthdays past, a vision of a unicorn, a (remembered) dream, social conventions, radio waves, climate change and interactions with its pets and other humans.

In general, the definition of feature relies on domain-specific knowledge. For instance the visual range of birds extends to much higher frequencies than our

own, as does the audio range of dogs. Features allow us to express concepts in terms of observable behaviour (rather than state of mind). The features of a human system may include 'the pet dog', 'its lead being taken from the peg', and 'daily walk'. A bird's features may include the state of the weather, its partner, number of fledgelings in its nest, the local flock and dawn chorus.

Notation 2 (Feature)**.** *A feature is something which can affect system behaviour. Features are of diverse type and depend on domain knowledge, as the examples above show. As an example, the space* $\mathfrak{F}$ *of all features for a system of humans may be defined syntactically:*

$$\begin{aligned} Basic &::= Habitat \mid Remembered \mid Imagined \mid Dreamt \\ \mathfrak{F} &::= Basic \mid {\sim}\mathfrak{F} \mid \mathfrak{F} \,\&\, \mathfrak{F} \mid \mathfrak{F} \rightsquigarrow \mathfrak{F} \mid \mathfrak{F} \,\dot{\rightsquigarrow}\, \mathfrak{F} \mid A_a\mathfrak{F} \end{aligned}$$

The proposition 'feature f *occurs at time* t*' is written* $f{\downarrow}t$*. Then the Boolean operations above are defined:*

$$\begin{array}{rll} ({\sim}f){\downarrow}t := \neg(f{\downarrow}t) & & f \textit{ doesn't occur at } t \\ (f\&g){\downarrow}t := (f{\downarrow}t) \wedge (g{\downarrow}t) & & f, g \textit{ both occur at } t \\ (f \rightsquigarrow g){\downarrow}t := (f{\downarrow}t) \Rightarrow (g{\downarrow}t) & & g \textit{ occurs at } t \textit{ if } f \textit{ does} \\ (f \,\dot{\rightsquigarrow}\, g){\downarrow}t := (f{\downarrow}t) \Rightarrow (\exists u{\geq}t \cdot g{\downarrow}u) & & g \textit{ occurs with or after } f \textit{ at } t. \end{array}$$

Of course the implications $\rightsquigarrow$ and $\dot{\rightsquigarrow}$ hold if their antecedents fail.

The absence of the absence of f is the same as the occurrence of f: $\sim$ is an involution. However an agent may be aware of neither $f{\downarrow}t$ nor $({\sim}f){\downarrow}t$. $\&$ is commutative, associative and idempotent. $\rightsquigarrow$ and $\dot{\rightsquigarrow}$ are transitive. As usual duality (de Morgan's Law) may be used to define the analogue of disjunction as ${\sim}(({\sim}f)\,\&\,({\sim}g))$, representing occurrence of at least one of f and g.

Not all features are relevant to an agent at a particular time and those which are have different levels of relevance. For instance you react immediately if your peripheral vision registers an approaching lion. For us features sensed from the habitat seem dominant, usually justified in terms of survival. But we, and many other animals, are also strongly aware of social conventions and experience, which we classify under 'Remembered'. Evidently different animals have quite different strengths of social sense.

A feature is said to 'enable' any action whose precondition it establishes. Our systems also require a more general version, eventual enabledness, in which the precondition is established eventually. For example, having fledgelings in the nest enables the parental action of feeding them and eventually enables the various parental actions of mentoring/overseeing their leaving the nest and flight.

Definition 1 (Enables)**.** *Assume act is an action. We say that a feature* $f : \mathfrak{F}$ enables *act if it establishes the precondition of act. Pointwise by time,*

$$f \,\mathbf{En}\, act := f \rightsquigarrow \mathrm{pre}(act).$$

More generally f eventually enables *an action which is enabled some time in future:*

$$f \,\mathbf{En}^{+} act := f \,\dot{\rightsquigarrow}\, \mathrm{pre}(act).$$

2.4 Multiagent Systems

The systems of agents we consider are adaptable multi-agent transition systems but with the notion of control and ambit of an action as basic. Agents are distinguished in our systems, as we have observed, by belonging to the ambit of at least one action. Recall that the habitat's environment is expressed as a default agent.

Definition 2 (System)**.** *A* system $S := (Agents, Actions)$ *is composed of a set Agents of* agents*, one representing the* environment*, and a set Actions of* actions*, each having an ambit. Agents have disjoint state spaces and interact by actions. So the state of S is the disjoint union of the states of the agents and each act* $\in$ *Actions has locus of control ambit*$(act) \subseteq$ *Agents and type, on each interaction in general,*

$$act : (States \times Input) \leftrightarrow (States \times Output) \ .$$

An agent responds to features in its habitat by behaving in some way. We take agent behaviour to be observable, although its causes may not be. Indeed human behaviour results from survival, pleasure, social pressure and 'free will'. Cell behaviour supports homeostasis. Government behaviour concerns running the country in response to its electorate, whilst dictatorship does not take into account the electorate! All have observable steps in behaviour.

A feature may enable many of an agent's actions. But at any time the agent may perform only some of them. Typically the choice is routine or even, we'd say, subconscious. For instance one can drive under normal conditions on 'auto pilot' and be aware of changing gear only if something untoward occurs in which case one needs to react spontaneously.

Thus an agent chooses actions routinely if aware of the features which enable them. But in special conditions, of 'deep awareness' which we identify with consciousness, the agent is aware that it is aware of certain features and must adapt its choice of action. Thus we identify consciousness of a feature with awareness of awareness of it, and consider that to result in the agent's adapting its choice of action.

Our systems adapt at two levels. At the system level that results from 'long term' changes; for example of a bird to climate change in its habitat. At the agent level that is due not only to incremental response to long-term changes but also to inter-agent, social, interactions. We return to this in Sect. 5.

3 Appreciating Awareness

We consider Boolean laws and numerical inequalities for the awareness, $(A_a f)\downarrow t$, by agent a of feature f at time t. Throughout we consider just a single agent and seek laws reducing awareness of a compound feature to awareness of simpler features, taking into account the consequences, under correspondence theory of modal logic, for the semantics.

For example in the Boolean model, awareness of f at t should imply that the proposition $f{\downarrow}t$ holds:

$$(A_a f){\downarrow}t \implies f{\downarrow}t \ . \tag{1}$$

Naturally the converse fails: an agent is aware of only certain features from its habitat.

Law (1), modal logic's Law T, implies by correspondence theory that the accessibility relation in the Kripke semantics is reflexive. It also implies that no time lag is required for a to become aware of $f{\downarrow}t$.

Numerically, if a is aware of $f{\downarrow}t$ then the strength of that awareness should equal the strength of $f{\downarrow}t$ (at the same time). We write that:

$$|(A_a f){\downarrow}t| = |f{\downarrow}t| \ . \tag{2}$$

Concurrency. Recall that $(f\&g){\downarrow}t = (f{\downarrow}t) \wedge (g{\downarrow}t)$. So is awareness of $(f\&g){\downarrow}t$ equivalent to awareness of $f{\downarrow}t$ and $g{\downarrow}t$ independently?

The former holds if $(f\&g)\downarrow t$ enables some action. But the latter holds if individually each feature enables an action, which is not necessarily the same due to the usual difference between pointwise and uniform behaviour gained by interchanging quantifiers. So we expect only implication to hold:

$$(A_a f){\downarrow}t \wedge (A_a g){\downarrow}t \Rightarrow (A_a(f\&g)){\downarrow}t \ . \tag{3}$$

A slightly contrived counterexample to the converse is provided by an agent which requires *two-factor authentication* from users before giving them access to some information. It enables user access if presented with the feature consisting of an ID plus two passwords. But if presented with an ID and a single password it does nothing. So the converse of (3) fails.

By comparison if only *single-factor authentication* were required then of course (3) would hold. But in neither case would the agent necessarily be conscious; it responds, so is aware, but with a strict strategy. A firewall, which requires two-factor authentication and which 'attacks' users submitting single passwords, would be conscious if its attack were developed *ad hom*, indicating flexibility of response.

Intuitively, the strength of awareness of a concurrent combination should be bounded above by the stronger of the strengths of f and g, and below by the weaker. Using $\sqcap$ and $\sqcup$ for min and max of numbers rerspectively:

$$|(A_a f){\downarrow}t| \sqcap |(A_a g){\downarrow}t| \ \leq \ |(A_a(f\&g)){\downarrow}t| \ \leq \ |(A_a f){\downarrow}t| \sqcup |(A_a g){\downarrow}t|. \tag{4}$$

Consequence. The fundamental Law K of Modal Logic is

$$\Box f \wedge \Box(f \implies g) \implies \Box g.$$

In terms of A_a that relies on an agent to appreciate when one feature is stronger than another which, as already discussed, is unrealistic for agents in general. But replacing the first occurrence of $\implies$ by $\rightsquigarrow$ leads to a plausible Boolean law:

$$(A_a f){\downarrow}t \wedge (A_a(f \rightsquigarrow g)){\downarrow}t \implies (A_a g){\downarrow}t. \tag{5}$$

By comparison with Law (3), we expect awareness of $f{\downarrow}t$ and $g{\downarrow}t$ to imply awareness of $(f{\rightsquigarrow}g){\downarrow}t$:

$$(A_a f){\downarrow}t \wedge (A_a g){\downarrow}t \implies (A_a(f{\rightsquigarrow}g)){\downarrow}t \ . \tag{6}$$

For strength, reasoning as for concurrency,

$$|(A_a{\sim}f){\downarrow}t| \sqcap |(A_a{\sim}g){\downarrow}t| \ \leq \ |(A_a(f{\rightsquigarrow}g)){\downarrow}t| \ \leq \ |(A_a f){\downarrow}t| \sqcup |(A_a g){\downarrow}t|. \tag{7}$$

Absence and Dual. Features f and ${\sim}f$ cannot occur simultaneously, by the meaning of ${\sim}$. So in the Boolean model an agent can not be aware of both $f{\downarrow}t$ and $({\sim}f){\downarrow}t$:

$$(A_a f){\downarrow}t \implies \neg((A_a{\sim}f){\downarrow}t). \tag{8}$$

The modal dual of A_a we write ∇_a.

Definition 3 (Dual)**.** *If at time t agent a is not aware of the absence of a feature, then the feature is considered to be* feasible *from a's point of view:*

$$(\nabla_a f){\downarrow}t := ({\sim}(A_a{\sim}f){\downarrow}t){\downarrow}t \ .$$

Our Boolean version of modal logic's Law D follows from Law (8):

$$(A_a f){\downarrow}t \implies (\nabla_a f){\downarrow}t. \tag{9}$$

By correspondence theory accessibility in a Kripke semantics is serial.

Numerically, from that we expect:

$$|(A_a f){\downarrow}t| \ \leq \ |(\nabla_a f){\downarrow}t|. \tag{10}$$

Consciousness. Consciousness implies awareness by definition, confirmed by Law (1):

$$(A_a(A_a f){\downarrow}t){\downarrow}t \implies (A_a f){\downarrow}t. \tag{11}$$

But not conversely, as for (1), since then there would be no difference between awareness and consciousness. By correspondence theory accessibility in a Kripke semantics is not transitive.

Numerically, from that we expect:

$$|(A_a(A_a f){\downarrow}t){\downarrow}t| \leq |(A_a f){\downarrow}t|. \tag{12}$$

Time. Awareness of a feature $f{\downarrow}t$ fades with time after t unless it is refreshed in some way. For instance driving home I am careful to select reverse gear to leave the parking lot and may be aware of the first couple of gear changes. But by the time I reach home I am unaware of having changed gear *en route* unless something untoward required me to pay particular attention.

Thus the strength of awareness of f in the future is at most its strength now, unless the awareness is refreshed. We expect the inequality: provided $\neg(A_a f){\downarrow}u$,

$$\forall u > t \ \cdot \ |(A_a f){\downarrow}u| \leq |(A_a f){\downarrow}t|. \tag{13}$$

If $(A_a f){\downarrow}u$ then the law holds only if $(A_a f){\downarrow}t$ too, in which case equality holds.

4 Soundness

In this section we continue with a single agent's perspective and define a simple model, define awareness with respect to it and show the foregoing laws to be sound. We write $\mathbb{T}$ for the time domain which we now assume is $\mathbb{N}$.

Recall that an agent is not aware of all features in its habitat, but for those of which it is aware, it is aware with a certain strength. For instance in Definition 2 the default strengths of the basic features for humans might be ranked

$$|Habitat| \ > \ |Remembered| \ > \ |Imagined| \ > \ |Dreamt|. \tag{14}$$

Indeed our survival depends on quick responses to threatening features in our habitat, but we are guided by memory in particular of social mores. For now we simplify and consider features to have the same strength, using feature strength to define strength of awareness.

The strength of a feature at time t is 1 if occurs at t and otherwise is inversely proportional to the length of time before t since it occurred.

Definition 4 (Strength)**.** *The length of time before or at t when feature f last occurred is a minimum of lengths of time:*

$$\tau(f,t) := \sqcap\{t{-}n \mid t \geq n,\ f{\downarrow}n,\ \forall m : (n,t] \cdot \neg(f{\downarrow}m)\}.$$

Thus it is zero if $f{\downarrow}t$. We adopt the convention that it is ∞ if f has not occurred up till t.

The strength $|f{\downarrow}t|$ *of feature f at time t is defined to be inversely proportional to the length of time $\tau(f,t)$:*

$$|f{\downarrow}t| := (1+\tau(f,t))^{-1},$$

where as usual $1+\infty=\infty$ and $\infty^{-1}=0$. Thus it is 1 if the feature occurs at t.

We also adopt a convention for successor and predecessor strengths, for use below. Suppose strength $d = (1+e)^{-1}$ where $e : \mathbb{N}^\infty$. Then the successor *is $d^+ := e^{-1}$ if $e > 0$ and undefined for $e = 0$. The* predecessor *is $d^- := (2+e)^{-1}$ for any $e : \mathbb{N}$.*

The strength of a combined feature is not readily expressed in terms of the individual strengths so the only bounds are simple:

Lemma 1 (Feature strength)**.** *The strength of a feature lies in $[0,1]$ and satisfies*

1. $(\sim)$ $\quad |(\sim f,t)| < 1 \ iff \ |(f,t)| = 1.$
2. $(\&)$ $\quad |(\sim f,t)|^+ \sqcap |(\sim g,t)|^+ \ \leq \ |(f\&g,t)| \ \leq \ |(f,t)| \sqcup |(g,t)|.$
3. $(\rightsquigarrow)$ $\quad |(\sim g,t)|^+ \ \leq \ |(f \rightsquigarrow g,t)| \ \leq \ |(g,t)|.$

Next awareness is formalised as follows. First we define when an agent is aware of a feature at a given time, and then in that case the strength of awareness.

Definition 5 (Awareness)**.** *Agent a is* aware *of feature $f : \mathfrak{F}$ at time $t : \mathbb{T}$ if at that time f enables some action at least partially within a's control:*

$$(A_a f)\downarrow t := \exists act : \mathcal{A}_a \cdot (f \,\mathbf{En}\, act)\downarrow t. \tag{15}$$

Using instead $\mathbf{En}^+$ gives the notion of eventual awareness*. The set of features of which a is aware at time t is denoted $\mathfrak{A}_a(t)$.*

The strength of awareness of feature $f : \mathfrak{A}_a(t)$ is defined to be the strength of f at time t (without delay):

$$|(A_a f)\downarrow t| := |f\downarrow t|. \tag{16}$$

Definition 6 (Consciousness)**.** *Agent a is* conscious *of feature f at time $t : \mathbb{T}$ if it aware of f at t and moreover immediately aware that it is aware:*

$$(C_a f)\downarrow t := (A_a((A_a f)\downarrow t))\downarrow t. \tag{17}$$

The strength at time u of consciousness is simply the strength of that awareness of 'awareness at time t':

$$|(C_a f)\downarrow t| := |(A_a((A_a f)\downarrow t))\downarrow t|. \tag{18}$$

The Boolean laws rely on the following result.

Lemma 2 (Closure)**.** *The space $\mathfrak{A}_a(t)$ of features of which a is aware at t is closed under $\&$, $\rightsquigarrow$ and $\barwedge$ but not $\sim$.*

Proof. For the typical case of & we reason:

$$\begin{array}{ll}
f, g \in \mathfrak{A}_a(t) & \\
\equiv & \text{definition of } \mathfrak{A}_a(t) \\
A_a(f,t) \wedge A_a(g,t) & \\
\equiv & \text{Definition 5 of awareness} \\
\exists F, G : \mathcal{A}_a \cdot (f \,\mathbf{En}\, F)\downarrow t \wedge (g \,\mathbf{En}\, G)\downarrow t & \\
\Rrightarrow & f\&g \in \mathfrak{F}, \text{ and } H \text{ discussed below} \\
\exists H : \mathcal{A}_a \cdot (f\&g \,\mathbf{En}\, H)\downarrow t & \\
\Rrightarrow & \text{Definition 5 again} \\
A_a(f\&g, t) & \\
\equiv & \text{definition of } \mathfrak{A}_a(t) \text{ again} \\
f\&g \in \mathfrak{A}_a(t). &
\end{array}$$

Since both f, g occur at t they are consistent so $f\&g \in \mathfrak{F}$. The action H may be taken to be any nondeterministic choice of the two actions which results in being either F or G, the choice being resolved at a lower level of detail.[3] Any such choice H satisfies

$$\mathrm{pre}(H) = \mathrm{pre}(F) \vee \mathrm{pre}(G)$$

[3] For instance a choice of probability p may be attributed to the environment and F chosen with probability p (and G with probability $1-p$).

so $f\&g$ enables H as required. Furthermore $H \in \mathcal{A}_a$ because

$$ambit(H) = ambit(F) \cup ambit(G)$$

and $a \in ambit(F) \cap ambit(G)$.

For $\sim$ we observe that if $f \in \mathfrak{A}_a(t)$ then by Definition 1 $f(s)\downarrow t$. By definition of $\sim$ and the assumption that at most one of f and $\sim f$ occur at any time, $(\sim f)\downarrow t$ cannot hold, so again the definition ensures $\sim f \notin \mathfrak{A}_a(t)$. □

The next result establishes soundness of both Boolean and numerical laws.

Theorem 1 (Correctness)**.** *The laws (1) to (13) from Sect. 3 hold.*

Proof. The Boolean laws in Sect. 3 not already established, (1), (3), (5) and (6) follow from simple arguments using the Closure Lemma.

For the numerical laws, the proof of Law (4) is typical. We reason:

$$\begin{array}{ll} |(A_a(f\&g))\downarrow t| & \\ = & \text{Definition 4} \\ |(f\&g)\downarrow t| & \\ = & \text{Definition 5 with appropriated } d : \mathbb{N} \\ 1/(1+d)\ . & \end{array}$$

Now $(f\&g)\downarrow t$ iff both $f\downarrow t$ and $g\downarrow t$ by Definition 2. So d, the time to the most recent occurrence of both f and g, is bounded above by the time to the more recent of f and g which is:

$$\begin{array}{rcl} |(f\&g)\downarrow t| & \leq & |f\downarrow t| \sqcup |g\downarrow t| \\ & & = |(A_a f)\downarrow t| \sqcup |(A_a g)\downarrow t|. \end{array}$$

It is bounded below by the first occurrence of either f or g which is one more than the most recent occurrence of either $\sim f$ or $\sim g$:

$$\begin{array}{rcl} |(f\&g)\downarrow t| & \geq & |(\sim f)\downarrow t|^+ \sqcap |(\sim g)\downarrow t|^+ \\ & & = |(A_a \sim f)\downarrow t|^+ \sqcap |(A_a \sim g)\downarrow t|^+. \end{array}$$

We infer Law (4). □

5 Adaptivity

In this section we reflect on the kinds of system agents inhabit.

Our agents adapt both in the short term and very much longer term and so fit, as already observed, squarely with Valiant's ecorithms [26]. Short-term, day-to-day, adaptations we regard as adjustments and long-term adaptations as evolutionary. But our approach supports both, without any need for an inverse limit which would imply some limit to evolution, which seems implausible.

In terms of multi-agent systems, adjustments can be incorporated in the description of the system because they are predictable and so state can be

expanded to include changes. That is analogous to an aware agent not needing to change its manner of choosing the next step in behaviour. However evolution is not predictable and so state must be expanded and actions updated. In retrospect at any time the current system can be seen as a more comprehensive but non-adaptive system using the Myhill-Nerode construction [18] to construct states as equivalence classes of sequences of actions.

Considering that representation of an adaptive system retrospectively as a (non-adaptive) system, the changes satisfy a 'causality' (or non-magic) invariant. In the space-time of Physics, an event x can affect only those events in its future *light cone* $C^+(x)$. Events in the past cone $C^-(x)$ require 'retro causality' and those in its future but outside $C^+(x)$ require 'superluminal' communication.

For adaptive systems, a realistic causality condition is more complicated because connectivity is not homogeneous and some communications are synchronous whilst others are asynchronous. Because the relation $\mathbf{En}^+$, of eventual enabledness, is transitive it can be used to define an analogue of light cones.

Definition 7 (Cones)**.** *If* $act : Actions$ *then the* future and past cones of *act consist respectively of all actions which it eventually enables, and which eventually enable it:*

$$C^+(act) := \{act' : Actions \mid pre(act)\, \mathbf{En}^+\, act'\}$$
$$C^-(act) := \{act' : Actions \mid pre(act')\, \mathbf{En}^+\, act\}.$$

An agent is stable *at some point in the evolution of an adaptive system if further interactions do not change it: subsequently its state space and the actions entirely under its control remain unchanged.*

Our adaptive systems satisfy the invariant that changes occur only as restricted by future cones.

6 Related Work and Progress

Boolean laws for awareness, and hence for consciousness (seen as awareness of awareness), have been proposed as have inequalities for its strength. They have been shown to be sound, in spite of the reflexivity required for consciousness, in a very simple model.

The driving intuition has been that awareness 'enables' actions to form an agent's next behavioural step whilst consciousness provides an opportunity for an agent to adapt its way of deciding that behaviour. It seems difficult to formulate those ideas in observable (*i.e.* falsifiable) terms which is why we have resorted to laws and inequalities.

We know of no similar work, either law-based or in terms of choice of next behavioural step. Recent work seems to concentrate on architectural models which exhibit consciousness, and mostly for humans [19]. Influential examples are the Global Workspace Theory, GWT, of Baars [2] and the related Conscious Turing Machine, CTM, [4] of the Blooms. Those base the selection from many

alternatives of one for consciousness by ranking whilst our approach is less specific: an agent is conscious if it needs or is offered the chance to adapt the protocol for its behaviour. An interesting alternative based entirely on network density is the work of Grindrod, [10,11].

Early computational-based work stems from Johnson-Laird's general analogy between mind and computer [16]. In those terms remembered and subconscious features may be thought of as being like random-access memory. When a feature is 'downloaded' afresh from memory, it enters 'local store', and so the agent's awareness. That provides a computational analogy with caching which has been made explicit in different ways by GWT and CTM.

The origins of the computational approach to system evolution go back to Barricelli's experiments [3] and Ray's *Tierra* [20], extremely early and restricted precursors of Valiant's ecorithms [26]. Of importance in the evolutionary setting will be Hoffman's work on Computational Evolutionary Perception, CEP, [13] which overturns the naive interpretation of 'what we see is what's out there', by considering its use to the observer. Similar ideas will apply to an arbitrary agent and all features, and be essential in quantifying our approach further.

Tononi's Information Integration Theory, IIT, [25], provides a measure of consciousness but in view of the computational complexity of its evaluation, current interest appears to be in its simulation. To be a model of what the brain does, it must be feasible computationally.

There is much work on awareness in adaptive system theory, from which the reader may like to compare [14,15].

The generality of our agents means that they are not necessarily rational, so we are unable to exploit work on dynamic epistemic logic. Relaxed notions of modal awareness, necessary for reasoning by logical agents who lack logical omniscience and have only bounded computational ability, have been introduced by Fagin & Halpern [7]. They refine Levesque's idea [17] of 'explicit' and 'implicit' belief (the latter being the logical closure of the former), and show how to achieve the result within a Kripke semantics adapted to include time.

Our approach can be thought of as formalising and extending to agents that of Dehaene [6]. There are many more recent popular books by experts on consciousness than we can refer to, as well as several fine youtube videos. The topic seems recently to have captured popular interest.

This work suffers several deficiencies. Features have been assumed to have the same strength, though it is simple to assign them weights when defining feature strength, depending on their basic constituents, subject to say (14). The definition of awareness of features has made no attempt to relate features, which seems likely in reality but would require currently unknown structure on $\mathfrak{F}$.

A single agent has been considered. Realistically different agents have different strengths which could be incorporated in the definition of strength of awareness by a. That agent weight would vary with evolution during which species 'search' anti-entropically for a niche with lower potential energy, but expending both energy and time in the process. Consciousness acts to break a barrier and initiate an entropy-increasing run of awareness. The connection of this with

'free energy' [22,23] is an enticing topic of further work which would incorporate recent advances in understanding the evolution of awareness (for instance Graziano [9]) and consciousness (for instance Ginsburg & Jablonka [8]).

We have not considered higher-order awareness beyond the degree 2 to define consciousness. One of the benefits of our approach is the possibility of doing so to explain subconscious and anomalous behaviour like blindsight (The Stanford Encyclopedia, [24]: Higher-order Theories of Consciousness).

The incorporation of more general, nonlinear, time would make the theory more realistic, as would the inclusion of probability of actions and the observation that consciousness does not seem to be independent for each feature, but to be bunched by kind of feature. Finally, the theorem implied in Sect. 5 of an adaptive system represented as a system could be formalised and simulation criteria used to establish agent consciousness.

Acknowledgments. The authors thank colleagues Professors Ronnie Becker and Hans Georg Zimmermann for wise council in early presentations of this material, and the referees for identifying obscurities and encouraging us to extend related work.

References

1. Atamian, H.S., Creux, N.M., Brown, E.A., Garner, A.G., Blackman, B.K., Harmer, S.L.: Circadian regulation of sunflower heliotropism, oral orientation, and pollinator visits. Science **353**(6299), 587–590 (2016)
2. Baars, B.J.: A Cognitive Theory of Consciousness. CUP (1998)
3. Barricelli, N.A.: Work from the 2nd half of the 50s at Princeton IAS summarised in Chapter 7 of G. Darwin Among the Machines. Penguin, Dyson (1997)
4. Blum, M., Blum, L.: A theoretical computer science perspective on consciousness, 16 November 2020. arxiv.org/ftp/arxiv/papers/2011/2011.09850.pdf
5. Cucker, F., Smale, S.: Emergent behaviour in flocks. IEEE Trans. Autom. Control **52**(5), 852–862 (2007)
6. Dehaene, S.: Consciousness and the Brain. Penguin (2014)
7. Fagin, R., Halpern, J.Y.: Belief, awareness and limited reasoning. Artif. Intell. **34**, 39–76 (1988)
8. Ginsburg, S., Jablonka, E.: The Evolution of the Sensitive Soul: Learning and the Origins of Consciousness. MIT Press, Cambridge (2019)
9. Graziano, M.S.A.: Speculations on the evolution of awareness. J. Cogn. Neurosci. **26**(6), 1300–1304 (2014)
10. Grindrod, P.: On human consciousness: a mathematical perspective. Netw. Neurosci. **2**(1), 23–40 (2017)
11. Grindrod, P., Lester, C.: Cortex-like complex systems: what occurs within?, June 2020. www.researchgate.net/publication/341901669
12. Hameroff, S.: How quantum brain biology can rescue conscious free will. Front. Integr. Neurosci. **6**, 93 (2012)
13. Hoffman, D.D., Singh, M.: Computational evolutionary perception. Perception **41**, 1073–1091 (2012)
14. Hölzl, M., Wirsing, M.: Towards a system model for ensembles. In: Agha, G., Danvy, O., Meseguer, J. (eds.) Formal Modeling: Actors, Open Systems, Biological Systems. LNCS, vol. 7000, pp. 241–261. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-24933-4_12

15. Hölzl, M., Gabor, T.: Reasoning and learning for awareness and adaptation. In: Wirsing, M., Hölzl, M., Koch, N., Mayer, P. (eds.) Software Engineering for Collective Autonomic Systems. LNCS, vol. 8998, pp. 249–290. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16310-9_7
16. Johnson-Laird, P.N.: The Computer and the Mind. Harvard University Press (1988)
17. Levesque, H.J.: A logic of implicit and explicit belief. In: Proceedings AAA1-84, Austin, TX, pp. 198–202 (1984). Revised and expanded: Lab. Tech. Rept. FLAIR, #32, Palo Alto, CA, 1984
18. Lewis, H.R., Papadimitriou, C.H.: Elements of the Theory of Computation, 2nd edn. Prentice-Hall, Hoboken (1998)
19. Oxford Mathematics of Consciousness and Applications Network. omcan.web.ox.ac.uk
20. Ray, T.S.: An approach to the synthesis of life. In: Boden, M.A. (ed.) (Without Appendices) The Philosophy of Artificial Life. Oxford Readings in Philosophy, pp. 111–145. OUP (1996)
21. Sanders, J.W., Turilli, M.: Dynamics of Control. UNU-IIST Report 353, March 2007
22. Solms, M., Friston, K.: How and why consciousness arises: some considerations from physics and physiology. J. Conscious. Stud. **25**, 202–238 (2018)
23. Solms, M.: The Hidden Spring: A Journey to the Source of Consciousness. W. W. Norton & Co. (2021)
24. Stanford Encyclopedia of Philosophy. The Metaphysics Research Lab, Center for the Study of Language and Information (CSLI), Stanford University, January 2014
25. Tononi, G.: An information integration theory of consciousness. BMC Neurosci. **5**, 42 (2004). 22 pages
26. Valiant, L.: Probably Approximately Correct: Nature's Algorithms for Learning and Prospering in a Complex World. Basic Books (2013)