Consciousness by Degree

Yifeng Chen[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. The authors have previously proposed that, with agents ranging from humans and other animals through cells to organisations and software, (e.g. AIs), a theory is possible which accounts in principle for agent consciousness. That theory has been previously developed from Booleans to numerical weights, hinting at degrees of awareness and consciousness.

In this paper, an agent's degree of awareness at any time is taken to reflect its freedom of choice amongst its possible behaviours. It is expressed as the number of actions which are enabled as a next behavioural step at that time and over which the agent has at least partial control. An agent is conscious of things which enable a fresh choice of action there, an enumeration of which provides its degree of consciousness. Those notions of degree are shown to provide a satisfactory account of realistic examples and to provide sensible elementary laws.

Valiant has shown that, in our terms, a living agent adapting daily to survive in its habitat as well its evolving in the very much longer term, can in both senses be expressed ecorithmically as learning. That approach is used here to consider the roles played by awareness and consciousness in the adaptation of an agent and a species.

1 Introduction

We assume that agents range from animals (humans and others both domesticated and wild) through cells to organisations and software (like AIs), and promote the view that different types of agent may exhibit different degrees of consciousness, quite possibly zero. The study of laws satisfied by agent consciousness is pertinent because of current popular and professional interest in the question of whether or not an AI, like a Large Language Model, can be conscious. Without some criteria, how are we to decide?

Agents exist in some context which we model as a system. We continue from our previous paper [6] to adopt the standard view that a model of any system is constrained by the interrelated criteria of breadth (or extent) and depth (or level of detail).

An agent's context is called its *habitat*, whose details depend on the domain and purpose of the model. Typically it includes other agents, features external

J. P. Bowen et al. (Eds.): *Theories of Programming and Formal Methods*, LNCS 14080, pp. 87–109, 2023.
https://doi.org/10.1007/978-3-031-40436-8_4

and internal to the agent, and a 'catch-all' category called[1] the *backdrop*. The backdrop is deemed to be a default agent and the distinction between physical and backdrop features will be determined by emphasis of the model, taking account of the control of its dynamics. Examples of these ideas appear in the next section.

A system's dynamics includes the behaviour of its agents, one step at a time. Each step we call an *action*. An action may lie under the control of more than one agent including the backdrop. Indeed an agent is a system component characterised by having control (perhaps partial) of at least one action. A system is expressed as a data type, so that ongoing actions normally described by safety and liveness are expressed instead by their individual steps.

Any *scientific* approach to agent awareness and consciousness must be phrased in falsifiable terms. Thus we eschew an animal's 'state of mind' which is not falsifiable (at least with current neuroscience). Thus it is a matter of belief, not science, that a dog is happy when it wags its tail. An agent's actions we thus take to be observable only if they are falsifiable, which requires hypothesis testing if the actions occur probabilistically. Henceforth by 'observable' we mean falsifiably so.

Since Descartes and Locke, if not before, human consciousness has been thought of in terms of the means by which a person becomes aware of features in its habitat. A contemporary rendering is given by Bernard Baars's Global Workspace Theory, GWT, [1], which has inspired a dozen architectures purporting to account for consciousness of a feature at a time; see Sect. 7.

Our approach departs from those architectures in our insistence on falsifiability. The alternative taken has been the standard mathematical one: of offering laws satisfied by awareness and consciousness in the hope that eventually sufficiently many will accrue to characterise it. In case of shortfall there may still be sufficiently many laws to falsify consciousness of some agents. Also, in the absence of a *definition* it is still helpful to have *heuristics* for awareness and consciousness, which are strong enough to show consistency of the laws when there is any doubt.

In this paper we concentrate on the underlying model which is inspired by but simplifies those we have considered previously [5,6].

An agent is deemed heuristically to be *aware* of a feature (external or internal) at a given time which enables some action (in the sense of establishing its precondition) which is at least partially within the agent's control. The action need not occur, but it is a candidate for the agent's next behavioural step. In terms of a scheduling protocol $\mathcal{P}$ for the agent's next action, the agent is cognizant of the domain of $\mathcal{P}$, the actions from which it chooses, even though the protocol itself is unknown.

An agent is deemed heuristically to be *conscious* of something which causes a fresh choice of action, even though the protocol for making the choice still remains unknown.

[1] Called 'the environment' in our previous work [5,6].

The paper is structured as follows. Features are introduced with a light touch in Sect. 3 and used to express the awareness heuristically in both Boolean and numerical terms. The models are simple because of the restricted use made of features and of time. They are used to give corresponding new models of consciousness in Sect. 4.

To test the formalism, the case study of a simplified cell is presented in Sect. 5 and its degrees of awareness and consciousness computed. Adaptation of living agents, and the roles played by awareness, consciousness and the protocol $\mathcal{P}$, are considered in Sect. 6 using Valiant's concept of ecorithm. The paper ends with Related and further work, and a Conclusion.

But we start with an uncharacteristically anthropomorphic example which exemplifies the ideas mentioned above and motivates the heuristics used.

2 Cameo

It is an autumn afternoon. Two parents are feeding their 3-month-old daughter in response to her cries, whilst their 2-year-old son builds a tower with blocks in his bedroom and their pet Golden Retriever naps in its bed in a corner of the laundry beside its water bowl.

The parents are being guided by intuition with the upbringing of their son and so are now more experienced and relaxed with their daughter. They are alert to her needs and often anticipate them, burping her after feeding and checking her nappy if she seems discontented. Their son is becoming autonomous, beginning to assert himself and often able to play by himself, though at 2 still needs support and supervision. The dog (and the parents) have been well trained at the local *Canine Academy* and it is treated as one of the family. Suddenly it rouses to bark protectively after sensing a passing pedestrian outside, unheard by the parents.

Apparently having fed enough, the baby falls asleep. One parent goes to the kitchen to prepare dinner whilst the other takes the dog's lead off its peg in the hall. The dog rushes to the front door, tail wagging, in anticipation of its daily walk to the park. The son, hearing activity and knowing the schedule, emerges from his bedroom. As his parent puts the leash on the dog the son requests 'Me too' to join the trip to the park. Wanting to walk like his parent, he refuses to be seated in his stroller; for now anyway. The parents call farewell to each other and the walk begins.

On the way to the local park the dog, on the extensible leash, enthusiastically engages in its usual routine with every tree and lamp post whilst the boy, clasping his parent's hand, looks around curiously. The parent is idly contemplating some thoughts about work, when they come to a crossroad. Becoming instantly alert, the parent ensures that the dog is by their side, and begins to teach the boy the time-honoured algorithm involving looking each way before crossing the road. Suddenly a car approaches, much noisier and faster than usual. The dog watches it and growls, and the parent pauses to check their safety, then resumes the lesson, using the car to stress the danger of roads. The car fades into the distance and they cross.

At the park the dog, free of the leash, fruitlessly chases a bird searching for worms and insects in the grass. The boy roams free, and decides to collect acorns in a pile in his stroller. The parent keeps a watchful eye on both from a park bench whilst ruminating on what wine to open with dinner.

2.1 Discussion

The agents in the model underlying that Cameo include a family of four, their pet dog, a car driver, birds and perhaps worms and insects at the park, depending on the breadth of the model. For example He Jifeng does not happen to be included. The dog's external features include its bed, lead, trees and lamp posts along the way to the park; but the position of the planets is not included. Its internal features include its nature and nurture, with remembered locations and events; but biometrics are not included. Its backdrop includes the passage of the sun across the sky; again, its details depend on the depth and breadth of the model.

The dog's behaviour, by its nature (and species in particular) lives up to its epithet as man's best friend. As a result its actions often indicate surprising awareness of and attentiveness to the family's needs. Considering an agent to be aware of things which enable an action at least partially within its control, the dog is aware of food (which enables its eating), the family and other dogs (which alter its behaviour), its daily routine (which it anticipates), opportunities to play and for human attention.

The dog behaves differently at different times of the day, due to its awareness of the position of the sun overhead and ambient animal noises. The sun is an external feature lying beyond the control of any component of the system and so belongs to the backdrop.

A rock at the park undergoes dynamics, due to erosion by the elements. But since those lie beyond its control, the rock is not an agent.

By comparison, the nature and nurture of the parents means they coordinate closely with each other as guardians and providers. Other internal features include their aspirations and social expectations. They are aware of idle thoughts which enable their ability to relate them. But they are not aware, for instance, of current popular TV series. Their backdrop includes the domestic water supply and movement of the sun.

The baby is aware of far fewer features than the son who is aware of fewer than the parents. The baby is just becoming aware of the appearance and noise of the dog, which attracts her attention but enables no further action. The son is in addition keen to play with the dog as are the parents who also act to ensure its health and safety.

Counting the number of actions under each agent's control which are enabled by the dog, the baby has fewer than the son who has fewer than the parents.

The dog's awareness of its lead being taken from the peg enables a fresh action at that moment: its walk outside. On the walk to the park the parent is conscious only initially of taking steps, because care is required in descending the front doorsteps, and then in matching pace to that of the boy and the dog.

But then the footsteps becomes routine and the parent is no longer conscious of them. But they return immediately to consciousness if a fresh action becomes enabled; like recovering from tripping over a misaligned paving stone.

3 Features

We suppose that any system contains a set $\mathcal{F}$ of *features* (from our earlier work [5,6]) which are time dependent and influence agent and system behaviour. Features are compounded from a set *Basic* of (domain-dependent) features under Boolean combinators corresponding to 'non occurrence', 'joint occurrence', 'conditional occurrence', 'eventual conditional occurrence' and 'awareness', provided the result is observable as discussed in the Introduction.

Awareness is included as a feature because it plays an important role in an agent's choice of next behavioural step. For instance the dog's behaviour depends on its awareness of its lead being taken from the peg, and the parent's behaviour then depends on its awareness of the dog's awareness.

In the Cameo features include: 'the passing pedestrian'; 'the lead being taken from the peg' which leads to 'the daily walk', and so on. Features do not include 'radio waves' unless the system also includes an appropriate receiver, without which the waves are not observable.

Definition 1 (Features). *At any time the features of a system are either Basic, or defined in terms of the combinators:*

$$\mathcal{F} \;:=\; \mathit{Basic} \mid \neg\mathcal{F} \mid \mathcal{F} \wedge \mathcal{F} \mid \mathcal{F} \Rightarrow \mathcal{F} \mid \mathcal{F} +\!\!\Rightarrow \mathcal{F} \mid A_a\, \mathcal{F}$$

Since a combination belongs to $\mathcal{F}$ only if it is observable, even if f is a feature the inconsistent conjunction $f \wedge \neg f = \mathit{false}$ is not.

A feature's time dependence we treat modally, making the time variable explicit only when necessary. The notation A_a for agent a's awareness is chosen because we regard awareness as a modal operator and that notation resembles that used in epistemic and doxastic logic. We begin by giving the semantics behind the syntax A_a after which we deal with the logical combinators.

3.1 Awareness

The Cameo motivates a Boolean notion and a numerical one of agent awareness and consciousness.

An agent is deemed to be aware of something at time t which enables an action, at least partially within its control, then. The action is therefore a candidate for the agent's next action at t. The number of such actions is its degree of awareness at t.

To express that, a little notation is helpful.

(a) The set of actions which are at least partially within agent a's control at time t is called a's *ambit* and denoted $\mathcal{A}m(a, t)$ (from our earlier work [5,14]). For instance the dog's walk to the park lies in both its ambit and that of its owner. But the weather at the park belongs to the ambit of neither.

(b) For action α its precondition, pre α, holds at just those states s and inputs in from which α is defined and terminates. Writing α as a predicate in the four free variables s (state before), in (input), s' (state after) and out (output):

$$(\text{pre } \alpha)(s, in) \ := \exists\, s', out \cdot \alpha(s, in, s', out).$$

For instance the precondition for the dog to eat from its bowl is that it be by the bowl which contains acceptable food.

Time, as used in (a), is replaced (following tradition) by state in (b). The two are reconciled by replacing state in (b) by either time from $\mathbb{T}$ or in both cases by a trace of actions which have occurred, in order of occurrence.

Awareness of a feature f at time t, that f enables some action in a's ambit at t, makes sense only if $f(t)$ holds (which is why negation of features is essential). Simplifying the formalization:

$$\exists\, \alpha : \mathcal{A}m(a, t) \cdot f \wedge (f \Rightarrow \text{pre } \alpha)$$

leads to the definition:

$$A_a(f, t) \ := f(t) \wedge \exists\, \alpha : \mathcal{A}m(a, t) \cdot \text{pre } \alpha. \tag{1}$$

For instance the dog is aware of the passing pedestrian which enables its bark. Until then the humans are not aware of it, having more limited hearing. But then the parents' curiosity is aroused so the action of looking out the window is enabled by the dog's bark. An enabled action need not occur, so the parents may choose instead to continue what they are doing, perhaps because it is common for the dog to bark at pedestrians, or what they are doing is more important.

It is convenient to set:

$$\mathcal{S}(a, f, t) \ := \ \{\alpha : \mathcal{A}m(a, t) \mid f(t) \wedge \text{pre } \alpha\} \tag{2}$$

so that (1) becomes:

$$A_a(f, t) \ = \ \mathcal{S}(a, f, t) \neq \varnothing.$$

That leads to a definition of degree of awareness:

$$|\, A_a(f, t)\,| := \ \#\mathcal{S}(a, f, t). \tag{3}$$

That numerical measure is defined only for awareness $A_a(f, t)$ and not for features in general, as in our earlier work. The result is a simpler model requiring less commitment to unnecessary detail in examples.

3.2 Features Resumed

We can now return to the semantics of the Boolean combinators on features. Provided the result is observable, they are given pointwise on the time variable:

$$\begin{aligned}(\neg f)(t) &:= \neg f(t)\\ (f \wedge g)(t) &:= f(t) \wedge g(t)\\ (f \Rightarrow g)(t) &:= f(t) \Rightarrow g(t)\\ (f +\!\!\Rightarrow g)(t) &:= f(t) \Rightarrow \exists\, u \geq t \cdot g(u)\\ A_a(f,t) &:= \text{Definition (1).}\end{aligned}$$

Evidently a compound scenario within a system can be described by a combination of features. For instance in the Cameo the lead being taken from its peg leads to the dog's walk and so on.

Simple laws of awareness involving those combinators appear in Fig. 1.

$$A_a(f,t) \Rightarrow f(t) \tag{4}$$

$$\begin{pmatrix} A_a(f,t) \\ A_a(g,t) \end{pmatrix} = A_a(f \wedge g, t) \tag{5}$$

$$\begin{pmatrix} A_a(f,t) \\ A_a(f \Rightarrow g, t) \end{pmatrix} \Rightarrow A_a(g,t) \tag{6}$$

$$A_a(f,t) \Rightarrow \nabla_a(f,t) \tag{7}$$

Fig. 1. Simple laws for awareness of agent a in the Boolean model, subject to the qualifications in Theorem 1. The dual modal operator is defined as usual, pointwise on t, by $\nabla_a(f) := \neg A_a(\neg f)$.

Theorem 1 (Laws for awareness). *The laws of Fig. 1 hold, Expressions (5) and (6) provided $\mathcal{A}m(a,t)$ is closed under the demonic choice of actions. Furthermore the implications are strict.*

Proof. Law (4) follows immediately since f is a conjunct in Definition 1 of $A_a(f,t)$. The converse clearly fails; for instance in the Cameo, the parents are not aware of the passing pedestrian when the dog is.

For Law (5) we reason that if f, g both hold at t they are consistent so $f \wedge g$ is also a feature and hence:

$$\begin{pmatrix} A_a(f,t) \\ A_a(g,t) \end{pmatrix}$$

$\equiv$ Definition (1) of awareness

$$\begin{pmatrix} \exists\, \alpha : \mathcal{A}m(a,t) \cdot f(t) \wedge \text{pre}\ \alpha \\ \exists\, \beta : \mathcal{A}m(a,t) \cdot g(t) \wedge \text{pre}\ \beta \end{pmatrix}$$

$\equiv$ logic

$$\exists\, \alpha, \beta : \mathcal{A}m(a,t) \cdot f(t) \wedge g(t) \wedge \text{pre}\ \alpha \wedge \text{pre}\ \beta$$

$$
\begin{aligned}
&\equiv && \text{pre } \alpha \wedge \text{pre } \beta = \text{pre } (\alpha \sqcap \beta)\\
&\exists\, \alpha, \beta : \mathcal{A}m(a,t) \cdot f(t) \wedge g(t) \wedge \text{pre } (\alpha \sqcap \beta)\\
&\equiv && \gamma := \alpha \sqcap \beta;\ \Leftarrow \text{straightforward}\\
&\exists\, \gamma : \mathcal{A}m(a,t) \cdot f(t) \wedge g(t) \wedge \text{pre } \gamma\\
&\equiv && \text{Definition 1 again}\\
&A_a(f \wedge g, t).
\end{aligned}
$$

The proof of Law (6) is similar using angelic choice $\alpha \sqcup \beta$ instead of demonic choice.

Law (7) requires simple propositional reasoning:

$$
\begin{aligned}
&A_a(f,t)\\
&\equiv && \text{definition}\\
&f(t) \wedge \exists\, \alpha : \mathcal{A}m(a,t) \cdot \text{pre } \alpha\\
&\Rrightarrow && \text{logic, for any } X\\
&f(t) \vee X\\
&\equiv && \text{logic, with } X := \neg \exists\, \beta \ldots\\
&\neg(\neg f(t) \wedge \exists\, \beta : \mathcal{A}m(a,t) \cdot \text{pre } \beta)\\
&\equiv && \text{definition}\\
&\neg A_a(\neg f, t)\\
&\equiv && \text{definition}\\
&\nabla_a(f,t).
\end{aligned}
$$

Evidently the implication is strict. For example in the Cameo the dog may not be aware of the lack of water in its bowl because it is on the walk; so it is not aware of the presence of water. □

A probabilistic choice between two actions is a special case of their demonic choice. By comparison the existence of the angelic combination of consistent actions is a strong assumption, leading to actions which backtrack and so on.

4 Consciousness

We now make the assumption that time is linear and discrete. If initialization is important to the model, the time domain $\mathbb{T}$ is often assumed to be an initial segment of $\mathbb{T} := \mathbb{N}$. In other words it is $\mathbb{N}$ or, if finite, the interval $[0, n]$ of integers. But if initialisation is unimportant and time infinite, a more convenient choice may be $\mathbb{T} := \mathbb{Z}$.

Either way we assume that each non-initial time $t : \mathbb{T}$ has a unique predecessor t^- and each non-final time has a unique successor t^+.

We regard an agent a as conscious of a feature f at time t if a is aware of f at t *via* a *fresh* action: one which was not enabled at t^-.

We define a modal operator C_a for consciousness by expanding A_a to incorporate freshness:

$$C_a(f,t) \;:=\; \exists\, \alpha : \mathcal{A}m(a,t) \cdot \begin{pmatrix} f(t) \\ (\text{pre}\ \alpha)(t) \\ \neg(\text{pre}\ \alpha)(t^-) \end{pmatrix}. \tag{8}$$

As always that existence does not mean the fresh action need be taken.

As with awareness, that Boolean notion extends to degrees by enumerating the fresh actions:

$$\mid C_a(f,t) \mid \;:=\; \#\{\alpha : \mathcal{A}m(a,t) \cdot \begin{pmatrix} f(t) \\ (\text{pre}\ \alpha)(t) \\ \neg(\text{pre}\ \alpha)(t^-) \end{pmatrix}\}. \tag{9}$$

And, as with the relationship between the Boolean and numerical models of awareness,

$$C_a(f,t) \;=\; \mid C_a(f,t) \mid > 0.$$

$$C_a(f,t) \Rightarrow A_a(f,t) \tag{10}$$

$$\begin{pmatrix} C_a(f,t) \\ C_a(g,t) \end{pmatrix} \;=\; C_a(f \wedge g, t) \tag{11}$$

$$\begin{pmatrix} C_a(f,t) \\ C_a(f \Rightarrow g, t) \end{pmatrix} \Rightarrow C_a(g,t) \tag{12}$$

$$C_a(f,t) \Rightarrow \nabla_a(f,t) \tag{13}$$

Fig. 2. Laws for consciousness corresponding to those of Fig. 1, subject to the qualifications of Theorem 2. The modal operator dual to C_a is defined by decorating the dual of A_a: $\nabla_a(f,t) := \neg C_a(\neg f, t)$.

Laws for consciousness that correspond to those of Fig. 1 are given in Fig. 2. Their correctness follows from both the content and method of Theorem 1.

Theorem 2 (Laws for consciousness). *The laws of Fig. 2 hold, (5) provided $\mathcal{A}m(a,t)$ is closed under demonic choice and (6) provided it is closed under angelic choice of consistent actions. Furthermore the implications are strict.*

5 Case Study: A Cell

In this section we give an example of a system and an agent which is simple enough for its features to be identified more completely than in the Cameo and for the agent's awareness to be determined.

We choose to model an idealised typical cell and find it, not surprisingly, to be an agent which is aware but not conscious. No specialized biological knowledge[2] is assumed. We use the Z notation, mostly[3] as covered by Spivey [17].

The cell is distinguished from its environmental periplasm by a semipermeable membrane containing the cell's cytoplasm. For homeostasis, temperature and various concentrations like pH within the cell must remain within certain bounds. Temperature is determined by the environment but regulation of various concentrations in the cytoplasm is achieved by transpiration through the membrane, sometimes requiring energy from the cell. We abstract the various concentrations, but include as a fundamental feature $alive : \mathbb{B}$, whether or not the cell is alive. We suppose that for $t_0, t_1 : \mathbb{R}$ and temperature *temp* in centigrade,

$$alive \;\Rightarrow\; t_0 \leq temp \leq t_1.$$

Transpiration is achieved by 'channels' which import nutrients (like sugars and amino acids) and which export the byproducts of metabolism (like sodium ions, or volatile compounds). A channel may be:

(a) *passive*, not requiring energy but working with the gradient by osmosis or diffusion or being 'facilitated'; or
(b) *active* requiring cell energy to work against a concentration gradient using one of several methods.

Energy is produced by the break down of ATP, *Adenosine Triphosphate*, with water to give ADP, *Adenosine Diphosphate*, and phosphorus; see `Wikipedia`, [20]: Adenosine triphosphate. ATP is produced and stored in the cell's mitochondria by the TCA (Citric Acid or Krebs Cycle); see the survey by Massimo Bonora *et al.*, [4]. We abstract that mechanism entirely, and instead consider just the amount of *energy* available in the cell; see Garrett Heinrich [9].

We begin by modelling a cell's active importing channel as follows.

5.1 Cell Importer

An *importer* is an active channel which imports certain kinds of molecule, of type $\mathbb{M}ol$, to the cell. It is formed from two *domains*, one atop the other, as shown in Fig. 3.

[2] A helpful reference for further relevant details is `Wikipedia`, [20], for instance: Cell membrane; Active transport; Facilitated diffusion; Ion channel.
[3] The definition of operation *Release* in two steps is nonstandard but hopefully clear.

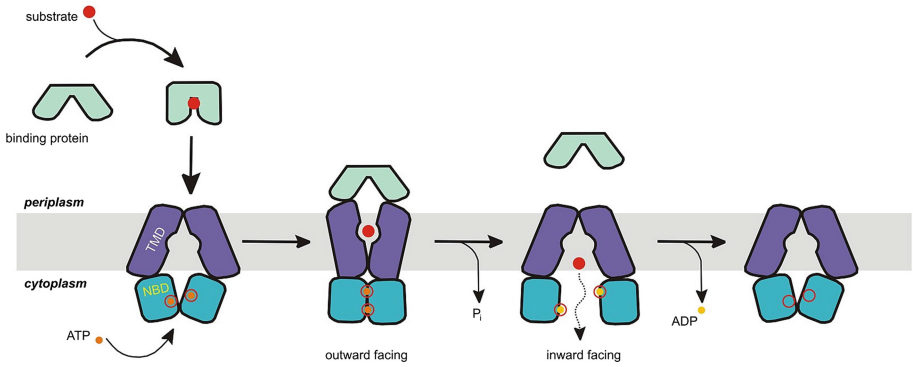

Fig. 3. Cross section of an importer through the cell membrane. Image from Wikipedia, [20]: ATP-binding cassette transporter (captured February 2023).

The domain contacting the periplasm, the *transmembrane domain*, has an outer gate, *tout*, to the periplasm and an inner gate, *tin*. Each is either closed or open, $tout, tin : \mathbb{B}$, with *false* representing closed. When the cell is live, if one of the gates is open the other is closed: $tin = \neg tout$.

Beneath is the *nucleotide-binding domain* whose outer gate *nout* contacts the cytoplasm and inner gate, *tin*, connects to the inner gate *nin*, so $tin = nin$. However now both gates of the nucleotide-binding domain may be closed, both may be open, or *nout* may be closed whilst *nin* is open. In summary $\neg(nout \wedge \neg nin)$, or $nout \Rightarrow nin$.

The two domains join in a cavity capable of holding a molecule of type $\mathbb{M}$ol (which depends on the kind of channel). Combining the four gate observables with the state of the cavity, the temperature and whether or not the cell is alive gives the importer's *State*. (Temperature is included to record the influence of the environment on the state of the cell).

Initially, both outer gates are closed, the inner gates are open and the cavity is empty. We describe initialisation as an operation which starts from an arbitrary state and terminates in an initial state, so we can use it later to reinitialise the state. As usual $\perp$ denotes the undefined state and $X_\perp$ denotes the type X augmented with $\perp$.

State

$$
\begin{array}{l}
alive : \mathbb{B} \\
temp : \mathbb{R} \\
energy : \mathbb{R}^{\geq 0} \\
cavity : \mathbb{M}\mathrm{ol}_{\perp} \\
tin, tout : \mathbb{B} \\
nin, nout : \mathbb{B} \\
\hline
alive \Rightarrow \begin{pmatrix} t_0 \leq temp \leq t_1 \\ energy > 0 \\ tin = \neg tout \\ tin = nin \\ nout \Rightarrow nin \end{pmatrix}
\end{array}
$$

Initialise

$$
\begin{array}{l}
\Delta State \\
\hline
alive' \\
cavity' = \perp \\
\neg tout' \wedge \neg nout'
\end{array}
$$

The implication in the state invariant is not an equivalence because:

(a) the cell may die for other reasons; and
(b) if the cell dies after attaining unsafe levels, it remains dead even if they subsequently return to normal.

Importing a molecule to the cell *via* the importer channel is done in three stages: docking, *Dock*; followed by release, *Release*; then reinitialisation, *Initialise*:

$$Import := Dock \fatsemi Release \fatsemi Initialise.$$

Dock inputs a molecule $m : \mathbb{M}\mathrm{ol}$ bound to a *binding protein* $bp(\ ) : \mathbb{P}\mathrm{tn}$ from the periplasm, in the form $bp(m)$. Formally, that is defined by feature combination: $bp(m)(t) := (bp(\ )\&m)(t)$. *Dock* also inputs a quantum of energy, en_0, from the cell, as discussed above.

Dock requires both outer gates to be closed initially (from which it follows by the state invariant that both inner gates are open) and the *cavity* to be empty (its content equals $\perp$). Afterwards it ensures that *tout* is open, *nout* remains closed (so by the state invariant *tin* and *nin* are both closed), *nout* is closed, *bp* is empty, the cavity contains molecule m, and energy has been consumed.

Dock

$$
\begin{array}{l}
\Delta State[cavity, tout, tin, nout, \\
\qquad nin, energy, alive] \\
bp(m)? : \mathbb{P}\mathrm{tn} \times \mathbb{M}\mathrm{ol} \\
\hline
alive \\
energy' = energy - en_0 \\
\neg tout \wedge tout' \\
\neg nout \wedge \neg nout' \\
cavity = \perp \\
cavity' = m?
\end{array}
$$

pre *Dock*

$$
\begin{array}{l}
State \\
bp(m)? : \mathbb{P}\mathrm{tn} \times \mathbb{M}\mathrm{ol} \\
energy? : \mathbb{R}^{+} \\
\hline
alive \\
\neg tout \wedge \neg nout \\
cavity = \perp \\
C.energy \geq en_0
\end{array}
$$

The precondition of *Dock* is that *tout* and *nout* are closed, the cavity is empty and the cell has sufficient energy.

Release assumes the conditions established by *Dock*. It outputs the empty binding protein, $bp(\ )$, closes *tout* and opens *tin*, *nin* and *nout* so that the molecule m in the cavity can enter the cytoplasm. We describe *Release* as two steps in sequence. In the first step, from *State* to *State*′, the upper outer gate closes whilst the lower outer gate remains closed, and $bp(\ )$ is output. In the second step, from *State*′ to *State*″, the upper outer gate remains closed whilst the other gates open and m is output to the cytoplasm.

Release

$\Delta^2 State$
$bp(\ )! : \mathbb{P}\mathrm{tn}$
$m! : \mathbb{M}\mathrm{ol}$

$alive \wedge alive' \wedge alive''$
$temp = temp' = temp''$
$tout \wedge \neg tout' \wedge \neg tout''$
$\neg tin \wedge tin' \wedge tin''$
$\neg nout \wedge \neg nout' \wedge nout''$
$cavity = cavity' = m!$
$cavity'' = \bot$

pre *Release*

State

---

alive
tout
$\neg nout$
$cavity \neq \bot$

The precondition is that *tout* is open, *nout* is closed and the cavity is nonempty.

Finally the importer is reinitialised with operation *Initialise*, leaving it in a state satisfying pre *Dock* (except for the cell's energy level). Of course *Initialise* is total.

To be able to function against a concentration gradient, the system of gates must function like an airlock. We use that property to 'validate' the breadth and depth of our model of an importer; without some such validation we can have little confidence in the accuracy of our model.

Theorem 3 (Airlock). *The action Import at no time connects the periplasm and cytoplasm directly.*

Proof. Since *Import* is the sequential composition of three actions it suffices to show the claim for each, *Dock*, *Release* and *Initialise*. We show that at no time (not just at the end of each step) are all four gates open:

$$\neg(tout \wedge tin \wedge nin \wedge nout). \tag{14}$$

Initially both outer gates, *tout* and *nout*, are closed so (14) is established. We argue operationally, but in Hoare-logic style, that (14) is preserved during the animation of *Import*.

Dock keeps closed the lower gate *nout* whilst closing the inner gates *tin*, and *nin* and opening the upper outer gate *tout*. So (14) is maintained.

The first step of *Release* keeps closed the upper outer gate *tout* whilst the lower outer gate *nout* remains closed. The second step of *Release* keeps closed the upper outer gate *tout* whilst the lower three gates, *tin*, *nin* and *nout*, are opened. So (14) remains true.

Finally *Initialise* keeps closed the lower outer gate *nout* whilst the others, and *tout* in particular, are unchanged. We conclude that (14) is maintained throughout. □

5.2 Cell Awareness

At any time our idealised cell may engage in the production of more ATP, or the import or export of molecules (including the ingredients or byproducts of the TCA cycle) on its channels like that described above with the four action steps of *Import*. Energy production also involves importing and exporting through the membrane of the mitochondria. The number of mitochondria depends on the metabolism requirements of the cell. In all, many actions are involved (characterising the breadth of study), each composed of many steps (depending on the depth of study).

We suppose for simplicity that the cell has 15 mitochondria, 20 importers, 20 exporters and 25 residual mechanisms. Each of those involves an action composed of steps in sequence, like *Import*. They are interdependent when resources are low, which for simplicity we overlook. The cell's degree of awareness of its internal activity f at t equals the number of actions in $\mathcal{A}m(cell, t)$ which are enabled at t. Enumerating by the four kinds of mechanism, that might typically be:

$$\begin{aligned} A_{cell}(f,t) &= 15 + 20 + 20 + 25 \\ &= 80. \end{aligned}$$

However that awareness, although observable, does not enable any fresh action and so the cell is not conscious of f at t:

$$\begin{aligned} &\neg C_{cell}(f,t) \\ &\mid C_{cell}(f,t) \mid = 0. \end{aligned}$$

6 Adaptation

We view a species, subject to evolution, as an agent in control of its DNA. The control is partial because of epigenetic influences; but that is sufficient for it to satisfy our definition of agenthood. On one hand such an agent adapts to its environment day-by-day and its species adapts by evolving generation-by-generation.

In this section we combine those forms of adaptation following Valiant's concept of ecorithm, to understand the roles played by awareness and consciousness in adaptation.

In interacting day-by-day with its habitat (including other agents), and buffeted by its backdrop, an agent adapts to survive. The result is a change in both the habitat and the agent's *nurture*. Evolving generation-by-generation the species improves its fitness to survive, subject to epigenetics and genetic mutations of its DNA. The result is a change in its *nature*.

Both forms of adaptation have been explained with the concept of an *ecorithm* by Leslie Valiant, [19] (who has been able to clarify and formalize Darwinian evolution in terms of machine learning). The 'fitness function' (or 'performance function' in Valiant's terms) evolves and maintains improvement towards a limit given by some mathematical, ideal function.

> *Making the assumption already implicit in Darwin's work-that different choices of action have various levels of benefit for the evolving entity-we can define the performance and the target in terms of the notion I call an ideal function. For any species (or other evolving entity), at any instant, in any specific environment, this ideal function will specify in every possible situation the most beneficial course of action.*
>
> Leslie Valiant, [19] p. 111.

We begin by formalising the life of a living agent in terms of its nature (DNA) and nurture (learning from its habitat). A machine-learning system learns how to classify a given datum on the basis of experience. The simple example of binary classification by means, used by Bernhard Schölkopf & Alexander Smola, [15]: Section 1.2, is specified in Fig. 6 of the Appendix as a data type. We now use those ideas to describe a living agent, culminating in Fig. 4.

6.1 Living Agent

The state of a living agent we take to consist of: whether or not it is alive, *live* : $\mathbb{B}$; its DNA, *dna* : *DNA*; its behaviour, or history of actions, *data* : seq *Action* from its ambit; the behaviour *habs* : seq *Action* of its habitat beyond its control; the behaviour *envs* : seq *Action* controlled by its backdrop, *e*; and its (unknown) choice protocol, $\mathcal{P}$, as above. We overlook the agent's identity.

Initially: the agent is alive with some DNA, dna_0; empty *data*, *habs* and *envs*; and some protocol $\mathcal{P}_0$.

It is no more straightforward to classify a living agent's interactions with its habitat than it is to provide details of $\mathcal{P}$.

> *Since the actions we take in one circumstance may influence what is the most beneficial action in another, it is the combination of all the action functions that is evaluated. The ideal one is that which produces most benefit in that snapshot of an environment.*
>
> Leslie Valiant, [19]; page 112.

Informed by *ML*, we describe the ways in which an agent's state can change. We include operations *Learn*, *Predict*, *FreeWill*, *Supervene*, *Vicinity*, *Beget* and

Die. Each requires the agent to be live. The operations *Learn* and *Predict* are as described by the general setting of the Appendix but using an unknown ideal function *FF* to update $\mathcal{P}$.

The details of *FreeWill* are concealed within the protocol $\mathcal{P}$. It is given as a separate operation because without it the classification of a living agent's interactions would seem incomplete. *Supervene* describes an action of the backdrop, either instantaneous by normal time scales like the eruption of Vesuvius, or incremental like an ice age. It may result in the agent's death, but anyway updates *envs*. *Vicinity* describes actions in the habitat which lie beyond the agent's control; it extends the trace *habs*. Like *Supervene*, its actions may well impact a's behaviour.

Beget is the only operation to output an agent, the offspring $b!$. It is specified with partner $a_0?$ and is asexual iff that equals the current agent. The function *Fme* describes how the offspring's DNA is formed from those of its parents, taking into account mutations and epigenetics. *Beget*'s precondition is that both the agent and its partner are alive.

After *Die* a living agent is no longer living (to state the obvious). Its control is difficult to specify because the operation may be internal to the agent, due to congenital malady or old age (Queen Elizabeth II), the result of actions of other agents (Julius Caesar) or of the environment (the population of Pompeii under erupting Vesuvius).

The specification of the type *Living agent* is naturally an extension of the type *ML*, which is the point of having considered it first. In spite of that we present it in Fig. 4 from scratch and for readability ignore various Z shortcuts.

In the spectrum of agents considered here, those which evolve are particularly important because they provide a way to understand the evolution of consciousness. It may be that there is an almost-Darwinian sense in which software evolves (did ChatGPT3.5 beget ChatGPT4?); certainly in the early years programming languages were classified by 'generation'. But for now we assume that evolution applies only to living agents.

6.2 Family Tree

Viewing a species as an agent we now extend the description of a single living agent to a species *via* its family tree, adapting ecologically. See Fig. 5.

The family tree consists of a sequence of finite sets of living agents, each an offspring of an agent in the previous generation. It also contains a fitness function (for simplicity we consider just one) which evolves by generation. In each generation all agents are live and share the same habitats and backdrops.

Initially the sequence is a singleton, the first generation consists of some nonempty set of agents and the fitness function is undefined, awaiting learning.

That concept of generation, represented by g in Fig. 5, requires comment. For the case of parthogenesis, asexual reproduction, it is well defined. Most human cultures have taboos against parent-child breeding, in which case g is also well defined. But otherwise, including for many animal species (see the interesting discussion of Victoria Pike *et al.*, [13], g needs to be defined as the length of the shortest path between the two sets of agents.

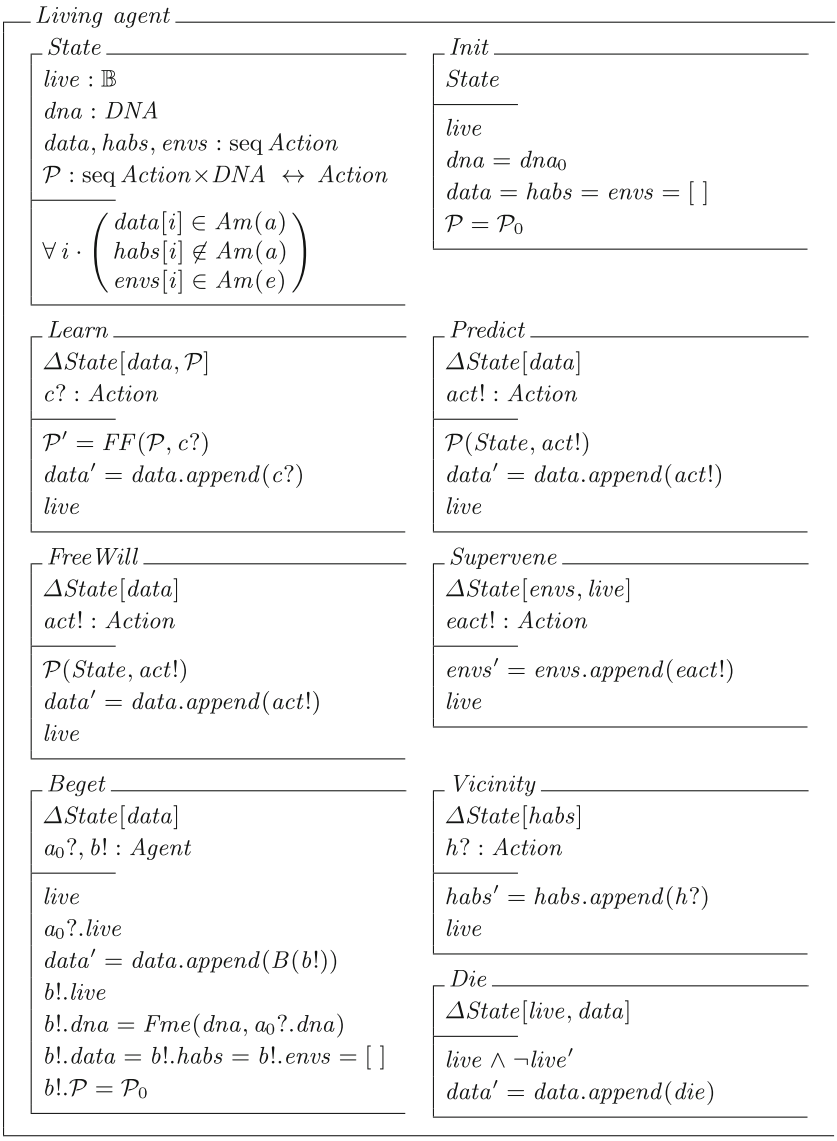

Fig. 4. The format of a living agent adapting day-to-day, in terms of machine-learning. Unknown procedures, like how the next action is the result of nature and nurture, or of free will, are abstracted in the next-step protocol $\mathcal{P}$ which is assumed to be updated in learning by *FF*. The agent's next action benefits from its nature and nurture by way of operation *Predict*.

Supervention by the backdrop may affect a whole generation as well as providing changes in the fitness function. It is total.

Each interaction of an agent with its habitat now consists of one of the operations *Learn*, *Predict*, *FreeWill*, *Vicinity*, *Beget* and *Die* from Fig. 4, the

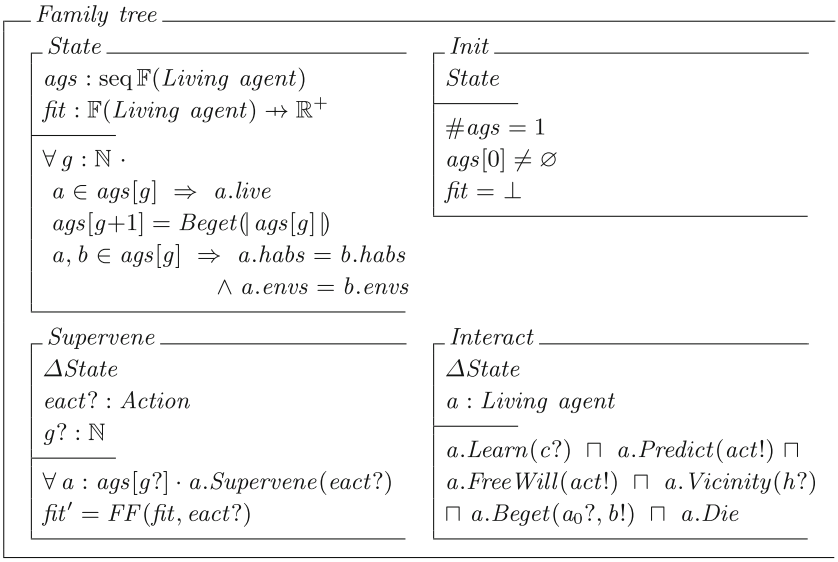

Fig. 5. The format for a family tree of living agents adapting and evolving in some habitat. Adaptation of a single agent to survive is as in Fig. 4, whilst evolution is described by supervention to change DNA between generations.

choice being made by its choice protocol. At this level we simply combine those actions nondeterministically.

6.3 Adaptation and Consciousness

What evolutionary advantage does consciousness confer?

Adapting day-to-day, a living agent is aware of features which enable some action for its protocol $\mathcal{P}$ to choose from. It is conscious of features which enable fresh actions and hence update the domain of $\mathcal{P}$. But whilst interacting and adapting, $\mathcal{P}$ changes incrementally as in machine learning.

Adapting in the long-term the species DNA is modified and as a result so is the protocol $\mathcal{P}$ (a change we allocate to *FF*). But now the change seems most unlikely to be incremental: arbitrary increments are likely. The only thing comparable in one generation would be a complete change of habitat, like animal migration or man walking on the moon.

Changes in $\mathcal{P}$ are thus either incremental, by generation, or possibly wild and unprecedented, in evolution. In terms of seeking a ground state in evolutionary phase space, presumably the goal of any species, that results in a well-known searching strategy which combines stepwise local search with jumps to avoid capture by local minima.

7 Related and Further Work

Russell and Whitehead and
Hegel and Kant,
Maybe I shall and maybe I shan't.
Maybe I shan't and maybe I shall.
Kant Russell Whitehead, Hegel et al.

Frederick Winsor, [21].

Our previous work (Chen & Sanders, [5,6]), used reflexivity to define consciousness as awareness of awareness and achieved reflexivity using the notion of 'feature' in an agent-based system. The model of feature strength was based on the observation that awareness and consciousness fade with time unless refreshed. A Boolean model was accompanied by a numerical one in which the strength of a feature was defined to be proportional to the inverse of the time since its occurrence.

When we came to use feature strength to give an account of examples, like those in the Cameo and the cell, we found that the assignment of weights to features, though elegant in theory, seemed too arbitrary in practice. Moreover, the justifying stability analysis seemed much too difficult.

In this paper we have considered an alternative, more restrictive, approach. It is still based on the number of possible behaviours under the control of the agent at any time and again supports agent consciousness by degree. But we have replaced consciousness as reflexive awareness with an approach which seems to work better on examples.

There is not much directly-related work, though of course the topic of consciousness is burgeoning. The Global Workspace Theory, GWT, of Bernard Baars [1] already mentioned has been very influential concerning human consciousness and its appreciation in terms of a means by which features are promoted to consciousness. Over the past two decades the idea of a global workspace has been refined by a dozen architectures, many explicitly computational like the Conscious Turing Machine [3] of Manuel & Lenore Blum.

Our work departs from the GWT architectural approach by insisting that agents be general and that as far possible concepts be falsifiable.

Stanislas Dehaene [7] has proposed that a human is conscious of any feature on which he or she can report. In our terms the form of the report may be predetermined (by the person's choice protocol) but its content is entirely feature dependent and so the report itself is fresh. That indicates consciousness in our terms, so our heuristic can be seen as generalising consciousness to arbitrary agents Dehaene's approach.

For further work, less related but nonetheless interesting and important, we refer to Section 6 of our earlier paper [5] which includes: Giulio Tononi's Information Integration Theory, IIT, [18]; Donald Hoffman's Computational Evolutionary Perception, CEP, [10]; Mark Solms & Karl Friston's use of the 'free-energy principle' in modelling homeostasis with the prospect of consciousness,

[16], Chapter 7; and the enticing evolutionary aspects of consciousness, briefly touched on here in Sect. 6.3: Simona Ginsburg & Eva Jablonska's [8].

We have not availed ourselves of Thomas Nagel's hugely influential view [12] that an agent is conscious iff *there is something it is like to be that agent*. Owen Holland [11] makes the point that Nagel's view is based on living agents and that for artificial agents it might instead be replaced by an approach founded in engineering, and he discusses the difference between physical and virtual AIs. Our approach, restricted to those cases, though different does not seem far away.

We have recently discovered the work of Yoshua Bengio, for instance [2], which also takes an entirely non-architectural approach to consciousness of AI (Large Language Models in particular) but at a lower level of abstraction. Nonetheless his priors have much in common with our features (when interpreted probabilistically) and may suggest a way forward with feature strength.

There is a desperate need for realistic case studies, particularly concerning the development of consciousness, which seems so far to lie in fiction.[4] Having identified a degree of consciousness, it would be interesting to consider the rate of change of consciousness during a living agent's lifetime.

This work has followed the classical view that time is linear and events, though they may be concurrent, are viewed in a sequential manner. For much of science that view is sufficient. It is fundamental to the global workspace metaphor and also provides the basis for the traces of concurrent computations. It may well be that a nonlinear time domain makes more sense in considering consciousness.

8 Conclusion

We conclude that agent awareness and consciousness may be explained by degree (without explicitly assigning strength to features) in a way which makes much sense in examples.

In the case study of a cell we have inferred that the cell is aware but not conscious, with a degree of awareness $| A_{cell}(f, t) | = 80$. Eighty? Eighty! **He Jifeng**, we offer salutations and congratulations on your 80^{th} *Festschrift* and look forward to your 90^{th}, and before then more of your hugely influential work from which we have benefitted directly and as part of the community.

A tiny step has been made towards a setting in which to study the ecological development and contribution of consciousness.

Acknowledgements. The authors acknowledge the support of Chinese grant 2021YFB0301100. They greatly appreciate the standard established by Jonathan Bowen, Xu Qiwen and Li Qin, under the auspices of Zhu Huibiao, in the difficult setting of a *Festschrift*. Jonathan in particular has gone far beyond the call of duty to ensure the event and publication a success.

They are grateful that the paper in this unfamiliar topic has benefitted considerably from five reviews with a range of backgrounds. We are extremely grateful to the referees for their patience and insights.

[4] *The Enigma of Kaspar Hauser*, Werner Herzog, originally *Jeder für sich und Gott gegen alle*, 1974.

Appendix: Machine Learning

A machine-learning system for binary classification of data in $\mathbb{R}^n$ may be specified as an abstract data type as follows, resulting in Fig. 6.

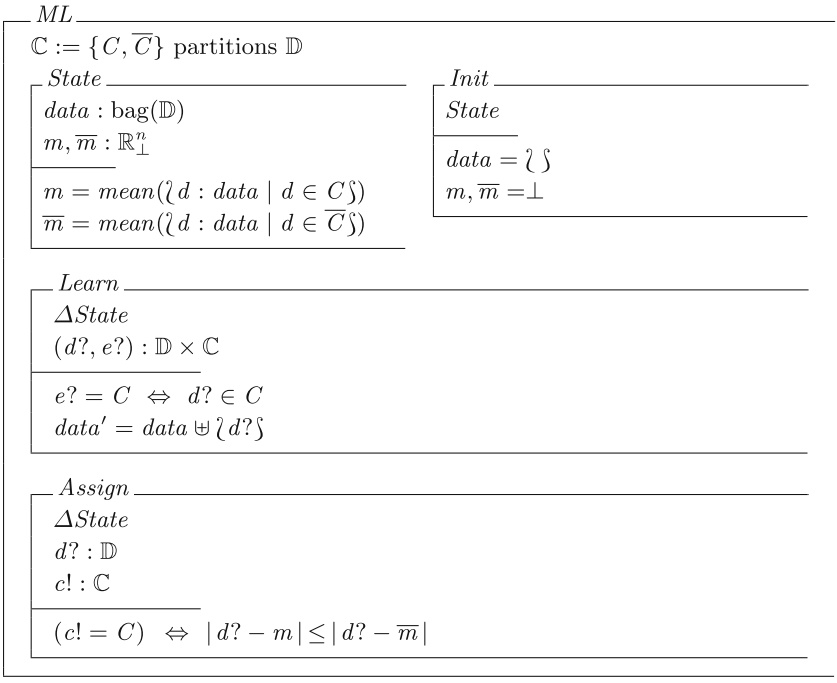

Fig. 6. A simple machine-learning system input, as described by Bernhard Schölkopf & Alexander Smola, [15]: Section 1.2. The system learns and assigns to an input datum 'the' class whose mean lies closer.

Assume a type $\mathbb{D}$ of data, already a subset of $\mathbb{R}^n$ for some $n > 0$, and a Boolean partition $\mathbb{D} = C \cup \overline{C}$ in which each datum is assigned to either C or its complement $\overline{C}$ (we use other notation for the mean), determined by its membership. We write $\mathbb{C}$ for the partition $\{C, \overline{C}\}$ and assume it remains constant.

The state of the machine-learning system consists of a bag, or multiset, *data*, of data seen so far, together with the means $m, \overline{m}$ of their assignments to $C, \overline{C}$ (respectively) so far. Initially the bag of data is empty, $data = \wr\, \wr$, with the means undefined $m, \overline{m} = \perp$.

Learning results from correctly-assigned input data. The training operator, *Learn*, takes a datum and its classification and adds the datum to *data* and updates the means. So *Learn* is total.

The assignment operator, *Assign*, assigns to an input datum the category to whose mean it is closer in $\mathbb{R}^n$. For a nonempty bag D of data we write its mean

as $mean(D) := (\#D)^{-1} \sum D$, where $\sum$ denotes bag summation. *Assign* is non-deterministic if the input datum is equidistant from both means. Its precondition is that the means are well defined: *data* contains data from each class.

In general machine learning, assignment of a general class $\mathbb{C}$ to an unseen input $d?$ is done by a protocol $\mathcal{P}$ which has been learnt in the same way that $m, \overline{m}$ are learnt there (facilitating output of a class with mean closer to $d?$). In general the protocol is a relation:

$$\mathcal{P} : \text{bag}\,(\mathbb{D} \times \mathbb{C}) \times \mathbb{D} \;\leftrightarrow\; \mathbb{C}.$$

In the case of Fig. 6 with binary assignment it has the simple form:

$$\begin{aligned} &\mathcal{P}(d?, m, \overline{m}, c!) := \\ &(c! = C) \;\Leftrightarrow\; |\, d? - m \,| \leq |\, d? - \overline{m} \,| \;. \end{aligned}$$

In general we assume $\mathcal{P}$ to be updated in learning by some function FF.

Some machine-learning systems first learn $\mathcal{P}$ and then use it to classifies input. Our description in Fig. 6 allows further learning at any stage, so the format more closely matches that of a living agent. For instance a young animal spends its early years learning whilst interacting with its habitat; a process which continues throughout its life.

References

1. Baars, B.J.: A Cognitive Theory of Consciousness. CUP, Cambridge (1998)
2. Bengio, Y.: The Consciousness Prior (2017). https://arxiv.org/abs/1709.08568
3. Blum, M., Blum, L.: A theoretical computer science perspective on consciousness (2020). https://arxiv.org/ftp/arxiv/papers/2011/2011.09850.pdf
4. Bonora, M., et al.: ATP synthesis and storage. Purinergic Signal **8**(3), 343–357 (2012)
5. Chen, Y., Sanders, J.W.: A modal approach to consciousness of agents. In: Margaria, T., Steffen, B. (eds.) ISoLA 2022. LNCS, vol. 13703, pp. 127–141. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19759-8_9
6. Chen, Y., Sanders, J.W.: A modal approach to conscious social agents. In: Steffen, B., Wirsing, M., Margaria, T. (eds.) Transactions on FoMaC. LNCS, Springer (2023, to appear)
7. Dehaene, S.: Consciousness and the Brain. Penguin, London (2014)
8. Ginsburg, S., Jablonka, E.: The Evolution of the Sensitive Soul: Learning and the Origins of Consciousness. MIT Press, Cambridge (2019)
9. Heinrich, G.: How to Measure the Energetic Status of Cells? Enzo Life Sciences, TechNotes (2022)
10. Hoffman, D.D., Singh, M.: Computational evolutionary perception. Perception **41**, 1073–1091 (2012)
11. Holland, O.: Forget the bat. J. Artif. Intell. Consciousness **7**(1), 83–93 (2020)
12. Nagel, T.: What is it like to be a bat? Philos. Rev. **83**(4), 435–450 (1974)
13. Pike, V.L., Cornwallis, C.K., Griffin, A.S.: Why don't all animals avoid inbreeding? Proc. R. Soc. B **288**, 20211045 (2021)
14. Sanders, J.W., Turilli, M.: Dynamics of control. UNU-IIST Report 353 (2007)

15. Schölkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge (2002)
16. Solms, M.: The Hidden Spring: A Journey to the Source of Consciousness. W. W. Norton & Co., New York (2021)
17. Spivey, J.M.: The Z Notation: A Reference Manual, 2nd edn. Prentice Hall International Series in Computer Science (1992)
18. Tononi, G.: An information integration theory of consciousness. BMC Neurosci. **5**, 22, Article no. 42 (2004)
19. Valiant, L.: Probably Approximately Correct: Nature's Algorithms for Learning and Prospering in a Complex World. Basic Books, New York (2013)
20. Wikipedia: Cell membrane; Active transport; Cotransporter; Ion transporter; ATP; etc
21. Winsor, F.: The Space Child's Mother Goose. Purple Press House (1956). (2001 edition)