Original Article

Adaptive Behavior

Adaptive Behavior
2017, Vol. 25(2) 72–82
© The Author(s) 2017
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1059712317700501
journals.sagepub.com/home/adb
SAGE

The Bayesian stance: Equations for 'as-if' sensorimotor agency

Simon M^c^Gregor

Abstract

The verb 'to do' plays a vital part in our understanding of the world, and it goes hand-in-hand with words such as active, action and agent. But the physical sciences describe only mechanical happenings, not acts. Their theoretical language is, in essence, a strict mathematical formalism applied to the description of variables (usually quantitative ones) that can – at least in principle – be measured by mechanical instruments. In such a language, what is the definition of an agent? Of an act? In contrast to previous approaches, which attempt to discriminate between agent and non-agent systems, we pursue a more Dennettian approach that attempts only to characterise the explanatory logic of intentional (agentive) interpretations of a physical system; we wish to do so purely in terms of the formal relations that hold between variables in a dynamical system or stochastic process. Our approach is straightforward: we use Pearl's causal formalism to identify physical variables at the causal boundary between 'agent' and 'environment', and identify these with variables in Bayesian decision theory; this provides a rigorous bridge between mathematical models of physics and mathematical models of rational decision-making.

Keywords

Agency, embodiment, 'as-if' theory, intentional stance, sensorimotor contingencies, Bayesian decision theory

1 Introduction

In place of traditional cognitivist mainstays such as representation and information processing, embodied cognitive theory relies, as noted by Barandiaran, Di Paolo, and Rohde (2009), on the concept of an *agent*, and more generally on the concept of *agency*. Agency is a notion that plays a major role in how neurotypical humans intuitively understand the world. We draw a basic distinction between two different types of occurrence:

- an *act*, which is *done* by an *agent*;
- a mere *happening*, which occurs as a result of *purely mechanical* causes.

Previously, the most explicit attempt to define this distinction within the artificial life literature has been Barandiaran et al. (2009), which like the less mathematically precise work by Weber and Varela (2002) is concerned with distinguishing *intrinsically* purposive behaviour from behaviour that merely appears 'as if' purposive to an observer.

Since philosophical debates on the nature of purpose are unlikely to be resolved soon, we prefer to remain methodologically agnostic as to whether there is any difference between intrinsic and extrinsic (observer-imposed) purposiveness. We hence propose a new topic for theoretical research: a 'thin', 'as if' notion of agency. Here, we broaden the meaning of the term 'as if' so that it neither assumes nor excludes 'intrinsic' purpose; we also use the term 'thin' to indicate that we do not require the action to be (as if) consciously willed.

To say that something behaves 'as if' it has property X usually implies that it does not, in fact, have property X. However, there is clearly a sense in which a system possessing property X must also behave as if it had property X; it is in this, less restrictive, sense that we intend the phrase 'as if'. In other words, we classify both the regulation of temperature by a thermostat, and also the pursuit of prey by an eagle, as 'as if' agency.

A convincing 'as if' theory should be of interest both to those who believe in a distinction between intrinsic and extrinsic purposes, and those who do not: if such a distinction is meaningful, the 'as if' theory describes the

University of Sussex, UK

Corresponding author:
Simon McGregor, University of Sussex, Sussex House, Falmer, Brighton, Sussex BN1 9RH, UK.
Email: s.mcgregor@sussex.ac.uk

common properties shared by intrinsic and extrinsic agency, and hence describes necessary logical conditions on intrinsic agency; if no such distinction is meaningful, then 'as if' agency is all there is.

The 'intentional stance' perspective advocated by Dennett (1987) suggests how the problem might be addressed indirectly, by shifting attention to how our *explanations* of systems work. As explained by Dennett (2009):

> [t]he intentional stance is the strategy of interpreting the behavior of an entity (person, animal, artifact, whatever) by treating it as if it were a rational agent who governed its 'choice' of 'action' by a 'consideration' of its 'beliefs' and 'desires.'

Note that the intentional stance, like the 'physical stance' (the mechanical causal perspective), is defined as an explanatory strategy rather than an intrinsic property of a system (such as possessing 'internal representations', or alternatively being 'self-sustaining'). The intentional stance is the choice to use the language of agency to describe a system, rather than the language of physical properties.

Dennett's choice of *rationality* as the core defining feature of the intentional stance makes it perhaps uniquely suited to formalisation, since ideal (normative) judgement and decision-making processes are topics of intense mathematical interest. According to Dennett, 'intentional systems theory [is] envisaged as a close kin of, and overlapping with, decision theory and game theory.' Dennett 1987 (, p. 58). The current article merely follows this reasoning to its logical end, within the context of embodied systems that are engaged in continuous interaction with their environment.

2 Theoretical objectives

We would like to be able to formulate a theory of a 'thin', 'as-if' agency that meets the following criteria.

1. It should provide an account of how some physical phenomena constitute as-if *active behaviour* that appears to be performed (not necessarily in any conscious or volitional sense) by:
   (a) a particular *agent*;
   (b) in pursuit of particular *goals* or *values*;

in a way that identifies the particular behaviour, agent and reasons for action.

2. The theory should be *semantically appropriate*:
   (a) it should correctly classify as many as possible of the phenomena we ordinarily view as as-if active behaviour, or provide a new insight that persuades us the ordinary view is wrong;
   (b) if it classifies a phenomenon as as-if active behaviour, and we would ordinarily view the phenomenon as purely mechanical, the theory should describe a credible perspective from which the phenomenon appears to be active behaviour.
3. It should be formalisable as a *mathematical theory*, and one that is commensurate with the mathematical language of the physical sciences. In particular:
   - it should be expressible purely in terms of the formal relations that hold between variables in a dynamical system or stochastic process (this criterion is also described in Biehl, Ikegami, & Polani, 2016).

Our aims are primarily methodological, and hence we are concerned neither with being faithful to the views of any individual philosopher, nor with defending any particular philosophical perspective. Moreover, while the specific mathematical model of ideal rationality that we use (recursive Bayesian filtering combined with expected utility theory and exponential discounting) is ubiquitous in the literature, we consider that further developments in this area are required to do embodied rationality proper justice.

3 'Thick' and 'thin' agency

Authors such as Juarrero (2000), who defines agency as 'the difference between a wink and a blink', clearly appeal to a philosophically 'thick' notion of agency, that entails all the complexities of conscious volition. However, there is clearly an intuitive distinction between actions and mere happenings, even in cases where the agent in question is not consciously aware of the action they are taking, or perhaps even entirely unconscious (for instance, breathing while asleep, versus being sustained by a ventilator); we also allow that 'locked-in' patients may possess phenomenal consciousness during a period of time in which they are unable to act, i.e. do not function as agents.

These considerations suggest that it is worth considering a 'thin' notion of agency, that does not presuppose consciousness. In the sense that it achieves some goal or value for the agent, a blink – unlike, say, a Parkinson's tremor (Barandiaran et al., 2009) – is still an action, albeit an involuntary one.

Such an approach has the benefit of avoiding the notoriously 'hard problem' (Chalmers, 1995) of consciousness. Accounting for the subjective experience of bats (Nagel, 1974) is extraordinarily ambitious, let alone the subjective experience of simpler biological agents such as single-celled organisms. See the work by Boly et al. (2013) for a discussion of the challenges in scientifically studying consciousness even in non-mammalian animals with brains.

4 Related Work

4.1 Definitions of agency

Precise formal definitions of agency are in short supply within artificial life and artificial intelligence (see Barandiaran et al., 2009; Biehl et al., 2016, for a discussion). As Barandiaran et al. (2009) point out, previous definitions of agency within cognitive science (e.g. Franklin & Graesser, 1996) have merely replaced the problem of defining agency with the problem of defining other cognitive concepts such as perception, action, or goals.

The definition proposed by Barandiaran et al. (2009) unfortunately does not meet our theoretical needs: it is insufficiently precise, too coarse-grained, and rules out examples we wish to include; we discuss these issues in more detail in Appendix 1.

Biehl et al. (2016) propose an information-theoretic measure of coherence within spatiotemporal patterns as a 'foundation for agent representation', while their proposed definition is methodologically rigorous, and interesting as a definition of coherent structure, they are not so ambitious as to propose it as a definition of agency itself.

4.2 Definitions of related concepts

While they do not directly address the concept of agency, there have been attempts to define the related notion of autonomy using statistical (information-theoretic) measures. The authors Bertschinger, Olbrich, Ay, and Jost (2008) 'tentatively' propose two information-theoretic quantities, and Seth (2010) proposes to measure autonomy using Granger causality, which is known to be closely related to transfer entropy (Barnett, Barrett, & Seth, 2009).

4.3 Rationality

The use of idealised formalisms of rationality (and, in particular, the Bayesian formalism) to model purposive embodied action is not novel: for instance, it is used both within the free energy framework (Friston, 2010, 2012) on which the predictive processing perspective by clark (2013) is founded, and for the practical purposes of inferring human intentions in robot learning (Baker et al., 2009). However, in both these cases it is used to model purposive action in systems *already assumed* to be intrinsically agentive; to the best of our knowledge, we are the first to propose that it can be used to *define* ('as-if') purposive action in arbitrary (non-ergodic) physical systems.

5 Intentional explanations, not intentional systems

For Dennett, the difference between a boulder rolling down a hill and a tortoise walking down the same hill is merely that the boulder invites a physical stance explanation, while the tortoise invites an intentional stance explanation.

In this article we will not attempt to engage with such distinctions: we merely wish to answer the question 'what does an intentional stance explanation look like in reductive physical terms', not 'for what sort of a system is the intentional stance sensible'. In this sense, we will aim to precisely characterise even those seemingly contrived senses in which the boulder behaves as if it were an agent with beliefs, desires and so forth.

This approach has an extraordinarily panpsychist flavour, in that it is in principle willing to extend 'as-if' agency to every conceivable macroscopic or microscopic object (including peculiar ones such as 'the left half of that badger's nose plus the planet Saturn'), providing that there is some consistent sense in which the object behaves as if it were an agent.

We do not see this as a weakness of the theory, but rather as a strength. There will indeed be some logically consistent perspective from which every system appears as if it were rational. The interesting question, to us, is not a simple binary classification into cognitive/non-cognitive, but a nuanced understanding of the precise senses in which a particular system, construed in some particular manner, behaves as though it were cognitive. After all, most interpretations will not fit a given system: for instance, a thermostat does not behave as if it were trying to destabilise the temperature (assuming realistic expectations of the effects of its actions).

The interpretations according to which a half-nose-plus-Saturn is an 'as-if' agent will surely be different in kind from (and less intuitively satisfying than) the interpretations according to which a biological organism is one. Our eventual hope is that these differences can be captured in quantifiable mathematical relations. This ambition is impossible without a concrete mathematical language in which the question can be investigated; the current article attempts to spell out such a language as a necessary first step.

6 Formalising 'as-if' agency

6.1 The general approach

We wish here to indicate how a thin, 'as if' notion of agency could be explicitly constructed in the language of the physical sciences. Essentially, if the causal effects exerted by some system X on its environment are roughly rational decisions under some formal theory of rationality R parameterised by some cognitive parameters C, we will say that X is an 'as-if' agent with respect to (R, C).

We take the minimal requirements for a theory of agency to be a distinction between the agent and its environment, a specification of what variables are taken to be under the control of the agent, and some notion of purpose or normativity (c.f. Barandiaran et al., 2009; Biehl et al.,2016). We propose that these can be captured along the following lines:

1. Formally define the agent, the environment, their sensorimotor interface, and a normative standard:
   (a) (Arbitrarily) distinguish a subsystem X within a Universe Ω, and define X's environment Y as everything in Ω that is not in X.
   (b) Define the system's *actuators* A and *sensors* S as the variables mediating the causal interaction between X and Y over time (in the formal sense from Pearl, 2000).
   (c) Define some theory of *embodied rationality* R that characterises how the actuators of an ideally rational agent should vary, in the context of a particular sensorimotor history, given certain notional 'cognitive' parameters C such as desires and beliefs.
2. Define a sub-trajectory A_T of A over a time interval T as being *as-if active behaviour* by X, **with respect to** (R, C), if and only if A_T approximately matches the prescriptions of (R, C).

Definition 1b is a first approximation that blurs the distinction between 'true' sensorimotor variables and mere interactional variables; this issue does not seem insuperable in principle and can be addressed in future work.

The important notion of *causation* may be cashed out formally using the well-known statistical, counterfactual model in Pearl (2000). In brief, this model introduces an 'intervention' operator into Bayesian graph theory that is distinct from an 'observation'; this operator is implicit in the equations that describe a physical system, but does not have a counterpart in standard statistical treatments.

7 Expected utility theory

The dominant normative framework in decision theory is *expected utility theory* (EUT), which deals with decision-making under conditions where outcomes z are not 100% subjectively certain; we take this to be the default condition of embodied agents. According to EUT, a rational action a is one that maximises the expected (mean) value of some 'utility' function $f(z)$ according to a subjective probability model $\mathbb{P}(Z \mid A)$ that relates actions A to outcomes Z. This mathematical form is not arbitrary; EUT's proponents argue that it can be derived from simple and plausible constraints (see, e.g. the von Neumann–Morgenstern representation theorem by Von Neumann & Morgenstern, 1945). Note that expected utility theory in this form abstracts decisions into nothing but actions and their possible consequences; the consideration of possible relevant information is ignored.

We can identify the utility function with a quantification of an agent's preferences, and the probability model with its (implicit) 'beliefs' or 'anticipations'. Intuitively, it makes sense that these inextricably co-determine which action is rational: without some notion of preferences, there would be no motivation to act; and without some anticipated link between action and outcome, there would be no justification for choosing a particular action.

7.1 Example: Thermostat

In principle, we would like to be able to reduce our definitions to the explicit level of microscopic physics, but this presents formal challenges that are well beyond the scope of the current article. Here, we will simply make the presumption that we may describe a physical system unproblematically in terms of macroscopic variables.

Consider a room containing a bimetallic thermostat, which connects an electrical heating circuit when the temperature falls below a certain value τ. We will assume that we can draw some arbitrary boundary around the thermostat, whereby the main external variable affecting the thermostat is the local temperature S, and the main internal variable affecting the environment is the thermostat's electrical output A.

We will imagine that the values taken by A might be (roughly) those values that would be rational decisions if A was being chosen so as to optimise the expectation of some utility function f (say, the squared deviation of the future global room temperature Y from some reference value τ). Then X would be an 'as-if' agent according to $(R, (f, \mathbb{P}))$ where R is EUT and $\mathbb{P}$ is a physically realistic model of the likely effects of A on Y given the local temperature S.

The point is not that this way of describing the thermostat is necessarily a useful one; after all, bimetallic thermostats are simple enough that we would typically prefer a 'physical stance' (mechanistic) explanation. Rather, we wish to indicate how the statement 'the thermostat behaves as if it were seeking to regulate the temperature' might be given at least one entirely unambiguous mathematical interpretation.

7.2 Example: Bacterium

Consider a microscopic environment containing a bacterium, which follows the local gradient of some nutrient chemical in its movements by altering the direction of rotation of its flagella.

We'll assume that we can draw a boundary in some way around the bacterium's external membrane and flagella. The agent-environment system's state can then be decomposed into the internal state of the bacterium X, and the external state Y which consists of the position and identity of every external molecule, for formal convenience considered *relative to the bacterium and its* *direction of motion* (this move collapses both the environment's state, and the bacterium's position and orientation in it, into a single variable Y).

The 'sensory' and 'actuator' variables S and A will then consist of the local physical state on either side of our (partially arbitrary) boundary. Of course, the causal effects in each direction will be extraordinarily complex, but for the sake of argument we'll ignore everything except the effects of the bacterium's flagellar rotation direction A on its position and orientation (considered egocentrically as motion of the environment Y relative to the bacterium), and the effects of the (average) local nutrient concentration S on the bacterium's internal chemistry.

The idea is that the direction of flagellar rotation A will correspond, roughly, to the rational decisions that the bacterium ought to take, if it only had the history of the average local nutrient concentration S (and its own actions A) to base its decisions on, and wanted to move to a peak area of nutrient concentration. Our aim is to indicate – in principle – how this intuition could be given precise shape in terms of Bayesian decision theory, although the details in practice are likely to be mathematically complex. By doing so, some important challenges become more sharply defined.

8 Formal complications

For simplicity, let us pretend that our example systems operate in discrete rather than continuous time. We can then talk of our variables S, A and Z as time series $(S_0, \ldots, S_n)$; $(A_0, \ldots, A_n)$; $(Z_0, \ldots, Z_n)$.

In our examples, in order to identify the 'agent' system's actuators A with rational decisions under EUT, we need to attribute the system both some preferences over possible outcomes Z (representing global temperature, for the thermostat, or closeness to high nutrient areas, for the bacterium), expressed as a utility function $f(z)$, and some systematic basis for linking decisions to outcomes, expressed as a subjective conditional probability distribution $\mathbb{P}(Z \mid A)$ (representing a 'model' or 'beliefs' about the effects of decisions, which we take for convenience to be 'realistic', subject to limited information, in both example systems).

Several conceptual issues come immediately to light here; we will address them in the next sections.

1. In both cases, the 'sensory' variable S affects the likely effect of decisions A on outcomes Z (and the actual values taken by the physical variable A). For the bacterium, this applies not only to the instantaneous value of S, but also to the past history of both A and S.
2. The outcomes Z we have specified are 'objective' global variables; we would prefer to be able to characterise the 'world (as if) for the system in terms only of the local variables at the causal boundary.
3. In general, we would like to be able to entertain notions of (as if) preference that describe the extended future, not only the immediate 'next step'.

8.1 Sensory dependence

Consider the thermostat. The likely room temperature outcome Z_{n+1} of different output currents A_n will depend on the immediate temperature Z_n; the local temperature S_n is correlated with Z_n and hence should affect the subjective 'model' of a rational agent.

The dominant formal model of how one should *update* subjective beliefs in response to evidence is, again, Bayesian: it prescribes the use of Bayes' rule, so that (in the decision-theoretic application) choosing decisions to maximise the expectation of $f(z)$ based on $\mathbb{P}(Z_{n+1} = z \mid A_n, S_n)$ rather than simply $\mathbb{P}(Z_{n+1} = z \mid A_n)$. Again, this formalism is not a matter of convenience; see, e.g. the work by Greaves (2013) for a principled justification.

The bacterium's flagellar rotation A is probably sensitive to the *temporal gradient* of local nutrient concentration S, not just S's instantaneous value. If the previous nutrient concentration S_{n-1} was higher than the current concentration S_n, then the bacterium is likely to be travelling against the nutrient gradient; if lower, it is likely to be travelling along the gradient.

The general history-sensitive case can be readily modelled in Bayesian decision theory by a subjective conditional distribution $\mathbb{P}(Z_{n+1} \mid A_{[0\ldots n]}, S_{[0\ldots n]})$ that takes the reasoner's entire sensorimotor history into account, in order to determine which action A_n is rational. This accommodates the thermostat's history-insensitive case as well; the thermostat may simply be taken to consider its history irrelevant (equivalently, to 'imagine itself' in a world with no hidden external 'state').

8.2 The 'as if' umwelt

So far, the possible outcomes Z that we have attributed our systems 'preferences' and 'expectations' regarding (room temperature, or the location of nutrient peaks) are not directly 'visible' to the systems, which only 'see' local variables.

We would like, if possible, to express the 'as if' beliefs and goals we attribute to a system in terms of only the physical 'sensorimotor' variables at its causal boundary. It is in fact possible to do so within the Bayesian framework, although a full exposition is beyond the scope of this article.

The main insight is that the *umwelt* (Ay & Löhr, ay2015; von Uexküll, 1934) or world-(as if)-for-the-system may be described entirely in terms of the *sensorimotor contingencies* (O' Regan & Noë, 2001) of the system, in contrast to the *umbegung* or 'objective' external world described by variables directly 'visible' only to the theorist. In Appendix II, we indicate how this idea can be formalised within Bayesian probability theory, using ideas found in the work by, e.g. Ay and Löhr (2015) and Orseau & Ring (2012), and introducing a novel 'trick' that considers agents to have beliefs and preferences regarding their future rational epistemic states. This trick resembles the approach described by Friston, Daunizeau, and Kiebel (2009).

8.3 Preferences for future states

For simplicity's sake, we have assumed that the behaviour of the thermostat and the bacterium can be mapped onto rational choices in pursuit of immediate goals (global temperature and position relative to a nutrient peak) expressed by utility functions over instantaneous outcomes.

However, in ongoing interaction scenarios, the use of utility functions requires some way of balancing future benefits against current ones. Constructing temporally extended measures of utility that converge to finite values can, for technical reasons, be tricky. A simple and commonly-used trick is to 'discount the future' by a discount factor $\gamma : 0<\gamma<1$, such that benefits i steps in the future are weighted by an exponentially-decreasing function γ^i, although alternative tricks have also been suggested, e.g. by Legg and Hutter (2006).

9 A concrete definition of 'as-if' agency

We will define a *sensorimotor policy* ('policy') π as a structure that provides a probabilistic prescription for taking actions, taking into account the trajectory of sensory stimuli and actuator engagements that have occurred in the meantime.

Any physical system, by the nature of its current state and its dynamics, in combination with its causal coupling with its environment, will implement a unique such policy; this includes arbitrarily complex systems, but also simple input-output systems like the thermostat (whose policies are 'memoryless' and deterministic).

Within expected utility theory (with exponential future-discounting), it is straightforward to define an optimal policy π^* (see, e.g. Orseau & Ring, 2012), relative to three parameters:

- a 'forward model' m taking the form of a conditional probability distribution $\mathbb{P}(Z_n, S_n \mid A_{n-1}, Z_{n-1})$;
- preferences encoded in a function $f : \mathcal{Z} \rightarrow \mathbb{R}$;
- and a future-discounting factor $\gamma \in \mathbb{R}$ that specifies how to balance future benefits at different times against one another.

This permits an explicit model of sensorimotor rationality. The set of (m, f, γ) triplets according to which particular system's action policy π is (f, γ)-optimal under m corresponds to the set of interpretations under which that system behaves as though it were a rational agent. While there may be multiple such solutions, most triplets (e.g. that the thermostat has a realistic forward model and wants temperature to be close to 200°C right away) are not solutions and will be ruled out by the system's behaviour, since the rational policy under such interpretations will differ from the system's dynamics.

Note that this definition does not solely consider the actual behaviour that the theorist observes in a particular instance of the system. In considering the policy π that the system implements, we give equal weight to counterfactual behaviour that the system could have exhibited if it had been exposed to different sensory stimuli.

10 Discussion

This article has presented a sketch of how, in a completely unambiguous mathematical sense, a coherent (albeit narrow, 'as-if') cognitive interpretation could in principle be imposed on factual physical dynamics. The 'glue' that allows the cognitive and physical domains to be joined in this way is the identification of certain physical variables as sensorimotor ones; the dynamics of these variables then have an (attributed) cognitive meaning as well as a physical one.

The position we present here is, in some sense, merely a conventional formalisation of an idea that will seem intuitively obvious to many. However, we are unaware of any previous attempt to spell it out in quite such unambiguous terms, including the following features.

1. We explicitly target a class of systems that includes examples both of supposedly extrinsic purpose (such as thermostats) and supposedly intrinsic purpose (such as human beings), in order to avoid metaphysical commitments.
2. We appeal to the formal model of causation in Pearl (2000) to identify the physical variables constituting 'inputs' and 'outputs' of a system.
3. We do not attempt to define a distinction between intentional and non-intentional systems, but instead to define a set of consistent *interpretations* of arbitrarily-delineated physical systems.

It is only by attempting to spell out these intuitions in mathematical detail for the embodied case that certain issues come to the fore: for instance, the need to describe not only beliefs and desires (as assumed by Dennett) but also some principle for balancing desires at different time intervals against one another (such as the future-discounting factor γ, which is the most arbitrary part of the framework). Addressing these issues provides a gradual route for progress.

10.1 Open challenges

It is important to emphasise that, even though the framework proposed in this article addresses a relatively modest question (what does it mean to say that a physical system behaves as though it were a rational sensorimotor agent?), it still leaves a number of significant theoretical challenges unaddressed. We discuss a few below, although the list is by no means exhaustive!

10.1.1 Quantification. Our future hope for the 'as-if' theory is that it will be possible to identify relevant dimensions of 'as-if' agency, including quantitative measures, such that, along some measure of 'agentiness', half-a-badger's-nose-plus-Saturn will be ranked lower than the half-nose alone, while the entire badger should presumably be ranked many orders of magnitude higher. In this way, the 'as-if' theory might begin to fit with some of our intuitive notions about 'real' agency, which distinguish boulders and thermostats from badgers and humans.

The pursuit of quantifiable measures fits with the idea of a continuity, rather than a discontinuity, between simpler physical systems and biological ones (Mcgregor & Virgo, 2009). Such a notion is attractive because there must at some stage have been a transition between the two. Similarly, the possibility of simultaneous 'as-if' cognition in multiple physically non-disjoint systems is highly relevant to discussions such as whether humans have an 'extended mind' (Clark & Chalmers, 1998), whether there is collective cognition in social insects, or whether the immune system is cognitive (Hershberg & Efroni, 2001).

10.1.2 Non-dualism. The notion of sensorimotor rationality presented in this paper is likely inadequate for general theoretical purposes, since it fails to address issues involving the 'non-dualistic' (Orseau & Ring, 2012) nature of embodied rationality. By virtue of being embodied, fundamental thermodynamic limitations regarding information-processing (Landauer, 1961) apply, and interactions with the world are not restricted to conventional sensorimotor channels. Hence, physically embodied agents are (intrinsically) *boundedly rational* (Braun & Ortega, 2014) (e.g. they have limited memory capacity), and their reasoning can affect or be affected by the world in 'unintended' ways (e.g. a decision to drink alcohol can be made on the basis of how it will affect one's reasoning capacities; the intensity of a thought process affects consumption of available metabolic energy and hence the radiant heat one emits).

Our everyday notion of rationality sometimes prescribes clear answers about what decisions are, or are not, rational under such non-dualistic embodied constraints; presumably, these should also be captured by a proper theory of embodied sensorimotor rationality. Some promising indications in this direction are offered by research such as space-time embedded general intelligence (Orseau & Ring, 2012), information-theoretic bounded rationality (Braun & Ortega, 2014), the free-energy framework (Friston, 2010), and the application of Bayesian decision theory to utility-driven decisions about belief updates themselves (Greaves, 2013); all adopt probabilistic formalisms, like the approach proposed in this article.

As a consequence of the embodied limitations and costs of cognition, humans must continually judge how much time and attention to give to their decisions. We often find ourselves acting in ways where, if we had thought longer or harder about a decision, we would have chosen to act differently. In the intuitive sense, such actions seem *irrational*. However, it seems plausible that a theory of bounded rationality might disclose important senses in which such actions are indeed rational.

10.1.3 Interoception. By considering the causal boundary between the organism and its environment to define a sensorimotor interface, we have identified the embodied sensorimotor process as entirely *exteroceptive*. This is a serious deficiency, as preferences over *interoceptive* senses such as pain and hunger are clearly crucial in embodied cognition (Seth, 2013).

10.1.4 Identifying models and preferences. We have not addressed how, given the theory of rationality R, one might go about computing the cognitive parameters C, according to which some particular system is an 'as-if' agent. The framework of *inverse reinforcement learning* (Baker et al., 2009) describes, in principle, how one might go about doing this using ordinary Bayesian inference, and demonstrates some practical applications. At present, the method has been applied only to inferring goals; the problem of inferring goals and beliefs simultaneously is no more theoretically difficult, but seems likely to be less tractable in practice than inferring goals alone.

11 Conclusion

This article constitutes an attempt to write out an explicit version of a simplified intentional stance, in the language of stochastic processes over physical sensorimotor variables. While this may seem trivial (or even pointless), it is only by doing so that important subtleties and limitations come to light: sensorimotor contingency models can be treated as outcomes over which an agent has preferences; we lack a ‘non-dualistic’ theory of rationality; and so on.

If we are right, this article, which is relatively explicit in the mathematical details of its physical reductionism, is merely the beginning of an outline of a formal theory of embodied cognition; a great deal of painstaking work is still required to flesh out our framework and move beyond the ‘thinnest’ form of ‘as-if’ cognition.

Acknowledgments

Nathaniel Virgo, Pedro Martínez Mediano, and Paulo De Jesus provided invaluable feedback on early versions of this article. Discussions with Inman Harvey, and feedback from anonymous reviewers, have transformed the focus, structure and exposition of the text.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Notes

[1] Of course, in the long term, the toxin fatally disrupts the bacterium’s locomotive behaviour.

[2] If the utility function takes the form of a KL divergence from an ergodic sensorimotor density, we may expect to reconstruct something similar to the notion of active inference (Friston et al., 2009) within the free energy framework.

References

Ay, N., & Löhr, W. (2015). The umwelt of an embodied agent—a measure-theoretic definition. *Theory in Biosciences*, *134*, 105–116.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*, 329–349.

Barandiaran, X. E., Di Paolo, E., & Rohde, M. (2009). Defining agency: Individuality, normativity, asymmetry, and spatio-temporality in action. *Adaptive Behavior*, *17*, 367–386.

Barandiaran, X. E., & Di Paolo, E. A. (2014). A genealogical map of the concept of habit. *Frontiers in Human Neuroscience*, *8*, 522.

Barnett, L., Barrett, A. B., & Seth, A. K. (2009). Granger causality and transfer entropy are equivalent for Gaussian variables. *Physical Review Letters*, *103*, 238701.

Bertschinger, N., Olbrich, E., Ay, N., & Jost, J. (2008). Autonomy: An information theoretic perspective. *Biosystems*, *91*, 331–345.

Biehl, M., Ikegami, T., & Polani, D. (2016). Towards information based spatiotemporal patterns as a foundation for agent representation in dynamical systems. *Paper presented at proceedings of artificial life 2016*. Cancún Mexico, 4–8 July 2016, pp. 722–730. Cambridge, MA: MIT Press.

Boly, M., Seth, A. K., Wilke, M., Ingmundson, P., Baars, B., Laureys, S., …Tsuchiya, N. (2013). Consciousness in humans and non-human animals: Recent advances and future directions. *Frontiers in Psychology*, *4*, 625.

Braun, D., & Ortega, P. (2014). Information-theoretic bounded rationality and ϵ-optimality. *Entropy*, *16*, 4662–4676.

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

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*, 181–204.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, *58*, 7–19.

Dennett, D. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

Dennett, D. (2009). Intentional systems theory. In A. Beckermann, B. P. McLaughlin, & S. Walter (Eds.), *The Oxford handbook of philosophy of mind* (pp. 339–350). Oxford, UK: Oxford University Press.

Egbert, M. D., & Barandiaran, X. E. (2014). Modeling habits as self-sustaining patterns of sensorimotor behavior. *Frontiers in Human Neuroscience*, *8*, 590.

Franklin, S., & Graesser, A. (1996). Is it an agent, or just a program? A taxonomy for autonomous agents. In: J. P. Müller, M. J. Wooldridge & N. R. Jennings (Eds.), *Intelligent agents III agent theories, architectures, and languages* (pp. 21–35). Berlin, Heidelberg: Springer.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*, 127–138.

Friston, K. (2012). A free energy principle for biological systems. *Entropy*, *14*, 2100–2121.

Friston, K. J., Daunizeau, J., & Kiebel, S. J. (2009). Reinforcement learning or active inference? *PloS ONE*, *4*(7), e6421.

Greaves, H. (2013). Epistemic decision theory. *Mind*, *122* (488), 915–952.

Hershberg, U., & Efroni, S. (2001). The immune system and other cognitive systems. *Complexity*, *6*(5), 14–21.

Juarrero, A. (2000). *Dynamics in action: Intentional behavior as a complex system*. Cambridge, MA: MIT Press.

Landauer, R. (1961). Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development*, *5*(3), 183–191.

Legg, S., & Hutter, M. (2006). A formal measure of machine intelligence. In: Y. Saeys, E. Tsiporkova, B. De Baets & Y. Van de Peer (Eds.), *Paper presented at proceedings of the 15th annual machine learning conference of Belgium and The Netherlands (Benelearn’06)*, Ghent, Belgium, 11–12 May, (pp. 73–80). Amsterdam.

McGregor, S., & Virgo, N. (2009). Life and its close relatives. In: K. György, K. István & S. Eörs (Eds.), *Paper presented at European conference on artificial life*, Budapest, Hungary, 13–16 September, (pp. 230–237). Berlin: Springer.

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, *83*, 435–450.

O’Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and brain sciences*, *24*, 939–973.

Orseau, L., & Ring, M. (2012). Space-time embedded intelligence. In: B. Joscha, G. Ben & I. Matthew (Eds.), *Paper presented at 5th international conference on artificial general intelligence AGI 2012*, Oxford, UK, 8–11 December 2012, (pp. 209–218). Berlin: Springer.

Pearl, J. (2000). *Causality: Models, reasoning and inference*. New York: Cambridge University Press.

Seth, A. K. (2010). Measuring autonomy and emergence via granger causality. *Artificial Life*, *16*, 179–196.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, *17*, 565–573.

Thompson, E. (2007). *Mind in life: Biology, phenomenology, and the sciences of mind*. Cambridge, MA: Harvard University Press.

Von Neumann, J., & Morgenstern, O. (1945). Theory of games and economic behavior. *Bulletin of the American Mathematical Society*, *51*, 498–504.

Von Uexküll, J. (1934). *Streifzüge durch die Umwelten von Tieren und Menschen*. Springer. Read in English trans. J.D. O'Neill, "A foray into the worlds of animals and men", 2010. Minneapolis, MN: University of Minnesota Press.

Weber, A., & Varela, F. J. (2002). Life after Kant: Natural purposes and the autopoietic foundations of biological individuality. *Phenomenology and the cognitive sciences*, *1*(2), 97–125.

Appendix I

A1.1 Barandiaran, Di Paolo and Rohde definition

In the work by BDR (Barandiaran, Di Paolo & Rohde, 2009), a completely different view on agency than the one we advocate here has been proposed. In terms of definitional precision, BDR set the bar for theories like ours; while we disagree with their approach, their level of explicit detail makes meaningful discussion possible. BDR's definition of agency appears to be intended not merely to account for biological agency, but to supersede the circular definitions offered by AI scientists, and hence is very ambitious in scope. It is as follows:

A system S is an agent for a particular coupling C with an environment E iff:

1. S is an *open autonomous* system in an environment E, meaning that:
   (a) among a set of processes a system S can be distinguished as a network of interdependent processes whereby every process belonging to the network depends on at least another process of the network and enables at least another one so that isolated from the network any component process would tend to run down or extinguish;
   (b) the set of processes (not belonging to S) that can affect S and are affected by S defines S's environment (E); and
   (c) S depends on certain conditions (specified by S) that in turn depend on E;
2. S *modulates* the coupling C in an *adaptive* manner.
   (a) where *modulation* indicates an alteration (dependent on S) in the set of constraints that determine the coupling between S and E;
   (b) *adaptive* means that the change in the coupling C contributes to the maintenance of some of the processes that constitute S.

We will argue that this definition is inadequate for three main reasons:

(a) it does not operate at a fine enough *level of granularity*;
(b) there are serious mathematical problems with the notions of *network of processes* and *modulation*, that will prevent the BDR theory from being commensurate with the mathematical sciences until suitable solutions are found;
(c) the use of a survival-based norm makes the intuitively meaningful notion of a *self-destructive act* logically contradictory.

A1.1.1 Level of granularity. BDR's definition does not operate at the granularity of acts; only of agents. Earlier in their article Barandiaran et al. (2009, p. 5), BDR explain that Parkinson's tremors are not actions because tremors do not achieve a normative end. Intuitively, this seems reasonable, but BDR go further in stipulating that '[a] human undergoing Parkinson's tremors' is '[not an agent]' because 'movements are not directed or responding to any internally generated norm.' (Table 1.)

There is a tension here: as BDR themselves acknowledge, 'the person is a well identifiable entity and the genuine source of her interactions with the environment' Barandiaran et al. (2009, p. 5). Surely, persons with Parkinson's disease can perform actions while experiencing a tremor at the same time; it is simply that the tremors themselves are not actions. BDR's definition can only classify the entire person (tremors, actions and all) as an agent or otherwise.

A1.1.2 Processes. The flagship mathematical methodology of enactivism, dynamical systems theory, formalises the abstract notion of a (single, holistic) process. A (non-arbitrary) network of sub-processes does not pop out directly from the dynamical systems formalism; proponents of the autopoietic view need to formally and rigorously define what such a thing even means. In this sense, BDR's definition simply trades questions about the individuality of agents for questions about the individuality of processes.

A1.1.3 Modulation. BDR are well aware that the notion of agency involves a distinction between doings and undergoings, and that, while humans intuitively see such a distinction as primitive and fundamental, physics does not. They attempt to resolve this inconsistency by introducing a notion of modulation as a special form of interaction. In our opinion, however, what distinguishes modulation from other forms of interaction relates to an agentive semantics, and hence defining agency in terms of modulation is bound to be circular. The onus is on BDR to show otherwise by presenting a precise and unambiguous mathematical definition.

BDR attempt to explain what they mean by a system *Smodulating* the interaction with its environment E using the following equations Barandiaran et al. (2009, p. 54)

$$dS/dt = F_Q(S, E) \tag{1}$$

$$dE/dt = G_Q(S, E) \tag{2}$$

Equations (1) and (2) describe the coupled dynamics of S and E, while in BDR's words:

> The parameter Q represents a set of conditions and constraints on the coupling, including constraints internal to each system.

This is a mathematically peculiar move: remember, E is taken to be everything that affects S that is not S itself. In dynamical systems terms, unless Q is intended to express discontinuous aspects of the system's dynamics (in which case, this point should be explicit and centre stage), equations (1) and (2) will be equivalent to

$$dS/dt = F(S, E) \tag{1a}$$

$$dE/dt = G(S, E) \tag{2a}$$

for some F and G that are not parameterised and do not change over time. These are the canonical equations describing the S,E dynamical system in question, and they do not involve the seemingly redundant parameter Q.

Even more confusingly, BDR proceed by offering the following equation

$$\Delta p = H_T(S) \qquad p \subset Q \tag{3}$$

where 'a subset of [Q] is described at a given time by the parameter p.' This stipulation that $p \subset Q$, is also very remarkable: it implies that the values Q can take are themselves sets; as such, there really needs to be an explicit definition of the set U for which $Q \in U$.

In order for equation (3) to constitute a predicate that can be true of some dynamical systems but not others, the variables p and T also need to be the subject of the existential quantifiers $\forall$ or $\exists$ (or be bound variables of some previously introduced expression).

BDR explain their equation in words as follows:

> Equation 3 describes the asymmetrical modulation of the coupling by the agent. It applies only for an interval of time T and not for all time.

Even leaving aside the mathematical imprecisions that threaten to make equation (3) vacuous, it does not induce the asymmetry it is supposed to account for. In particular, it does not rule out the possibility that

$$\Delta r = I_T(E) \qquad r \subset Q \tag{4}$$

(whatever this might mean) for some appropriate r and I_T, re-introducing the symmetry between S and E. Hence, it is impossible to understand in what sense it defines an asymmetrical modulation.

A1.1.4 Self-destructive acts. In the autopoietic account, the primary source of value is survival. There have been some attempts by autopoietic cognitive scientists to account for behaviours such as injecting heroin, which do not seem to be survival-normative, as stemming from the self-maintenance of some higher-level process (Barandiaran & Di Paolo, 2014; Egbert & Barandiaran, 2014). The claim is that survival-threatening behaviour maintains some sort of organisation or identity at a different level than the biological organism as a whole: for instance, a habit or a cultural institution, or some stable pattern in the organism's internal dynamics. While it may make sense to posit such non-organismal identities, they do not help to explain the fine-grained structure of agency at the organismal level.

Consider a bacterium which through a point mutation has a tendency to exhibit chemotaxis towards a toxic substance. The relations of toxin and nutrient to the bacterium's metabolism are different, but the relation to the bacterium's short-term locomotive behaviour may be the same for toxin and nutrient.[1]

Proponents of the autopoietic account need to do one of two things.

1. Account for why the mutant's chemotaxis constitutes active behaviour on the part of the very entity whose existence is threatened by that behaviour (the entire bacterium).
2. Deny that the mutant's chemotaxis constitutes active behaviour on the part of the bacterium itself (either by denying that it constitutes active behaviour of any sort, or by asserting that it constitutes active behaviour on the part of some other level of identity).

It is hard to see how the first can be done while still maintaining that the intentional nature of acts performed by an agent is grounded purely in the survival-normativity of that very same agent (this being what sets the autopoietic approach apart from other schools of cognitive science).

The second option is more internally consistent, but it seems to fly in the face of what we mean by active behaviour. In particular, it renders the notion of performing intentionally self-destructive actions literally contradictory: either the events are not actions at all, or they are the actions of some entity that is not destroyed by them.

Note that, according to our view, it is perfectly possible for some physical dynamics to be an ‘as-if’ self-destructive act, if they are roughly the decisions a rational agent bent on self-destruction would take. Of course, this would require the use of a suitable theory of rationality R in which a preference for self-destruction could be encoded in some cognitive parameter C; it is unclear whether this would be meaningful in our proposed sensorimotor decision theory, but there is no reason to suppose it is impossible in our more general framework.

A1.1.5 Skepticism regarding the autopoietic view. We have raised serious formal and semantic objections to the definition of agency in the work by Barandiaran et al. (2009), and by extension other autopoiesis-based definitions. It is possible that these objections will be overcome by future versions of the theory; if so, we welcome such developments.

Appendix 2

A2.1 Sensorimotor decision theory

This appendix sketches out, in the loosest mathematical detail, an approach for formalising embodied decision-making that puts *sensorimotor contingencies* (O’Regan & Noë, 2001) centre stage: contingency models correspond to the probabilistic models that are required in Bayesian decision theory; and contingency models themselves are taken to be the objects of the agent’s preferences.

To begin, we assume that an agent’s sensorimotor contingency ‘model’ at time i is a (subjective) stochastic process $\mathbb{P}_i$ over a variable S (‘sensors’) causally conditioned (Pearl, 2000) on another variable A (‘actuators’), inducing a well-defined interventionised probability function $\mathbb{P}_i(S_{[i...j]} \mid \widehat{A}_{[i...j-1]})$; for every $j \geq i$.

Since this ‘model’ will change from time step to time step, we will treat the instantaneous model itself as a random variable M in a stochastic process $\mathbb{P}$, that parameterises the agent’s subjective expectations at each time step:

$$\mathbb{P}_i(S_{[i...j]} \mid \widehat{A}_{[i...j-1]}) = \mathbb{P}(S_{[i...j]} \mid \widehat{A}_{[i...j-1]}, M_i = m_i)$$

where $\mathbb{P}(S_{[i...i+n]} \mid \widehat{A}_{[i...i+n-1]}, M_i = m_i)$ is independent of i. (We will not worry about interpreting expressions like $(S_{[i\cdots j]} \mid M_i, M_j)$ that condition on multiple models M_i, M_j.

There is a relation between different models of this sort: we can consider what would happen if we ‘updated’ a model m_i with a sensorimotor trajectory $(s_{[i...j-1]}, a_{[i...j-1]})$ using Bayesian conditioning, to produce a new sensorimotor model m_j operating from j onwards. We will write this new model $m_j = m_i \circ (s_{[i...j-1]}, a_{[i...j-1]})$, pronounced ‘$m_j$ is m_i updated with $(s_{[i...j-1]}, a_{[i...j-1]})$’, defined essentially as follows (some notational liberty has been taken for the sake of brevity)

$$\begin{aligned} m_j = m_i \circ (s_{[i...j-1]}, a_{[i...j-1]}) \Leftrightarrow \\ &\mathbb{P}(S_{[j...k]} \mid \widehat{A}_{[j...k-1]}, s_{[i...j-1]}, \widehat{a}_{[i...j-1]}, M_i = m_i) \\ &= \mathbb{P}(S_{[j...k]} \mid \widehat{A}_{[j...k-1]}, M_j = m_j) \end{aligned}$$

This allows us to define an action-contingent stochastic process over the M_n variables themselves

$$\begin{aligned} &\mathbb{P}(M_i = m_i \mid \widehat{a}_{[1...i]}, M_1 = m_1) = \\ &\sum_{s_{[1...i]} \in V} \mathbb{P}(S_{[1...i]} = s_{[1...i]} \mid \widehat{a}_{[1...i]}, M_1 = m_1) \end{aligned}$$

where $V = \{s_{[1...i]} \in \mathcal{S}^i : m_i = m_1 \circ (s_{[1...i]}, a_{[1...i]})\}$ is the set of all sensory trajectories such that m_i is m_1 updated with $(s_{[1...i]}, a_{[1...i]})$.

We have identified the objects of an agent’s ‘beliefs’M as consisting of nothing more than the potential effects of its own actions A on its own future sensations S, and indicated how such ‘beliefs’ can also be understood, self-referentially, as describing the likely consequences of actions A on beliefs M themselves.

Within this framework, we can introduce a utility function $f : \mathcal{M} \rightarrow \mathbb{R}$ that describes preferences over (logically entailed) beliefs M at future time steps.[2] A standard application of EUT then provides a model of purely sensorimotor rationality that dispenses with all references to an external world entirely. Relevant aspects of the external world may then be (re-)constructed in sensorimotor terms if desired.

About the Author

Simon McGregor is an interdisciplinary scientist whose research focuses on theoretical questions relating to the continuities between biological organisms, artificially intelligent machines and other physical systems. He received his doctorate in Artificial Intelligence from the University of Sussex, and has worked on research projects in robotics and computational biology. Dr McGregor is currently an Honorary Research Fellow at the University of Sussex, UK, where he is associated with COGS (Centre for Cognitive Science) and the CCNR (Centre for Computational Neuroscience and Robotics). He is a qualified member of the British Psychological Society.