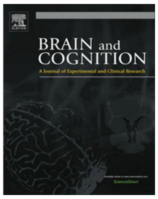
Brain and Cognition 112 (2017) 13–24

Contents lists available at ScienceDirect

Brain and Cognition

journal homepage: www.elsevier.com/locate/b&c



Predictive processing simplified: The infotropic machine

Chris Thornton

Centre for Research in Cognitive Science, University of Sussex, Brighton BN1 9QJ, UK

ARTICLE INFO

Article history:
Received 16 May 2015
Revised 18 March 2016
Accepted 21 March 2016
Available online 18 April 2016

Keywords:
Predictive processing
Predictive coding
Hierarchical prediction machine
Bayesian brain
Information theory
Cognitive informatics

ABSTRACT

On a traditional view of cognition, we see the agent acquiring stimuli, interpreting these in some way, and producing behavior in response. An increasingly popular alternative is the predictive processing framework. This sees the agent as continually generating predictions about the world, and responding productively to any errors made. Partly because of its heritage in the Bayesian brain theory, predictive processing has generally been seen as an inherently Bayesian process. The 'hierarchical prediction machine' which mediates it is envisaged to be a specifically Bayesian device. But as this paper shows, a specification for this machine can also be derived directly from information theory, using the metric of predictive payoff as an organizing concept. Hierarchical prediction machines can be built along purely information-theoretic lines, without referencing Bayesian theory in any way; this simplifies the account to some degree. The present paper describes what is involved and presents a series of working models. An experiment involving the conversion of a Braitenberg vehicle to use a controller of this type is also described.

1. Introduction

On a traditional view of cognition, we see the agent acquiring stimuli, interpreting these in some way, and producing behavior in response. An increasingly popular alternative is the predictive processing (also known as predictive coding) framework. This sees the agent as continually generating predictions about the world, and responding productively to any errors made (Brown, Friston, & Bestamnn, 2011; Clark, 2016; Friston, 2005; Friston, 2010; Hohwy, Roepstorff, & Friston, 2008; Huang & Rao, 2011; Jehee & Ballard, 2009; Knill & Pouget, 2004; Lee & Mumford, 2003; Rao & Ballard, 1999; Rao & Ballard, 2004). Clark characterizes this as 'the emerging unifying vision of the brain as an organ of prediction using a hierarchy of generative models' (Clark, 2013a, p. 185).[1] Granting that we can view actions as predictions put into a behavioral form, the proposal has the effect of unifying interpretive and behavioral functionality (Brown et al., 2011; Friston, Daunizeau, & Kiebel, 2009).[2] The model is also well positioned to use information theory (Shannon & Weaver, 1949; Shannon, 1948) as a way of explaining what is achieved. By improving performance in prediction, the agent renders the world less surprising, effectively gaining information (Cover and Thomas, 2006; Friston et al., 2012). The process can be seen as characteristically infotropic in this way (Thornton, 2014).

Partly because of its heritage in the Bayesian brain theory (Doya, 2007), predictive processing has generally been seen as an inherently Bayesian process. The 'hierarchical prediction machine' that mediates it is seen to be a specifically Bayesian mechanism. Processing is considered to be accomplished by inferential calculations. Backwards inference (i.e., application of Bayes' rule) is seen to be the means by which probabilities travel 'up' hierarchical structures, and forwards inference is the means by which they travel 'down.' Out of this bi-directional process, all functionalities of the brain are assumed to grow,[3] with the predictions of the machine being encapsulated in the conditional probabilities that connect one level of the hierarchy to another.

What the present paper draws attention to is an alternative way of specifying a machine of this type. In addition to the Bayesian formulation, there is an information-theoretic model, which is simpler in some respects. Key to this alternative is the metric of predictive payoff. Using basic principles of information theory, it is possible to measure the informational value of a prediction, provided

E-mail address: c.thornton@sussex.ac.uk

[1] The claim is part of a tradition emphasizing the role of prediction in perception and cognition, however (e.g. James, 1890/1950; Lashley, 1951; Mackay, 1956; Tolman, 1948; Yu & Dayan, 2005).

[2] The assumption underlying this is that 'the best ways of interpreting incoming information via perception, are deeply the same as the best ways of controlling outgoing information via motor action' (Eliasmith, 2007, p. 7).

[3] The 'pulling down' of priors is considered particularly significant (Hohwy, 2013, p. 33). As Clark comments, 'The beauty of the bidirectional hierarchical structure is that it allows the system to infer its own priors (the prior beliefs essential to the guessing routines) as it goes along. It does this by using its best current model—at one level—as the source of the priors for the level below' (Clark, 2013a, p. 3).

http://dx.doi.org/10.1016/j.bandc.2016.03.004
0278-2626/© 2016 Elsevier Inc. All rights reserved.

we know the value of the outcome predicted and whether or nor it occurs. We can measure the informational 'payoff' with respect to an event of known value. This metric then gives rise to a way of building prediction machines. Any network of inter-predicting outcomes in which evaluations are kept up-to-date propagates information between outcomes in a machine-like way. The general effect is that the machine transitions towards informational value. The network behaves infotropically, in a way that replicates the inferential activities of a Bayesian hierarchical prediction machine. The idea of predictive processing can thus be framed in a purely information-theoretic way, without using Bayesian theory.

The remainder of the paper sets out this alternative formulation in detail. Section 2 introduces the metric of predictive payoff, and examines its relationship to other measures from the Shannon framework. Section 3 shows how the metric provides the basis for building an information-theoretic version of the hierarchical prediction machine. Section 4 then demonstrates the behavior of some sample machines, including one deployed as the control system for a Braitenberg robot. Section 6 discusses neurophysiological issues, and Section 7 offers some concluding remarks.

2. Measuring predictive payoff

The theoretical foundation for the present proposal is Shannon information theory (Shannon & Weaver, 1949; Shannon, 1948). At the heart of this framework is the observation that certain events are well-behaved from the informational point of view. Given a strict choice of outcomes (i.e., a set of events out of which precisely one occurs), the informational value of the outcome that does occur can be defined as

$$-\log p(x)$$

where x is the outcome in question, and $p(x)$ is its probability. As Shannon notes, measuring the value in this way can be justified on a number of grounds. For one thing, it ensures that more improbable outcomes have higher informational value, as intuition suggests they must. For another, the value then corresponds to the quantity of data needed to signal the outcome. If we take logs to base 2 and round the value up to an integer, it is also the number of binary digits needed to signal what occurs.[4] For this reason, the value is often said to be measured in 'bits' (a contraction of BInary digiTS).[5] More formally, the quantity is termed the *surprisal* of the outcome (Tribus, 1961). Weather events are a convenient way to illustrate use of the measure. If everyday it rains with probability 0.25, but is fine otherwise, the informational value of the outcome of rain is $-\log_2 0.25 = 2$ bits.

Given this way of measuring the informational value of individual outcomes, it is straightforward to derive an average. Assuming we know the probability for all outcomes within the choice, the average information gained from discovering the result is

$$-\sum_x p(x)\log_2 p(x)$$

This formula defines the information gained on average from discovering the outcome. We can also see it as the information that is *expected* to be gained from discovering the outcome. More generally, we can see the quantity as the uncertainty that exists with respect to the choice. Shannon notes this average plays an important role in statistical mechanics, where it is termed entropy. Accordingly, Shannon uses the term entropy as a description. Average information may thus be termed entropy, expected surprisal, average surprisal, expected information or uncertainty (Cover & Thomas, 2006; Mackay, 2003).[6] The weather illustration can be extended to show how entropy measurement is applied: if everyday it rains with probability 0.2, snows with probability 0.1, and is fine otherwise, the average informational value of an outcome is

$$-(0.2\log_2 0.2 + 0.1\log_2 0.1 + 0.7\log_2 0.7) \approx 1.15 \text{ bits}$$

One difficulty with the framework is the status of the probabilities taken into account. Whether they are objective (defined by the world), or subjective (defined by a model possessed by the observer) is not specified.[7] In practice, either interpretation can be applied, and theorists tend to adopt whichever is appropriate for their purposes. Where entropy is seen as quantifying uncertainty, probabilities are likely to be seen as subjective. Where the formula is seen as quantifying generated information, they are likely to be seen as objective.[8]

Problems then arise if there is any difference between the two distributions. To give a concrete example, imagine that every day it rains with probability 0.2, but that an observer predicts rain with probability 0.4. The observer's prediction gives rain a higher probability than it really has. Plugging the objective probability into the formula, we find that the outcome generates a little over 0.7 bits of information. Using the subjective probability, the figure is nearly 1 bit. Without a distinction being made between subjective and objective probabilities, the evaluation is ambiguous.

One way of dealing with this situation is simply to disallow it. The position can be taken that the Shannon framework does not accommodate any deviation between subjective and objective probabilities. More productively, we can view the subjective distribution as a predictive model. On this basis, the predictions that arise can be seen (and evaluated) as ways of acquiring the informational value of an outcome *before* it occurs. The calculation is made as follows. A predictive model must give rise to particular predictions. Given the informational value of a correct prediction must be the informational value of the correctly predicted outcome, we can calculate the expected informational value of predictions with respect to an outcome that does occur. We can find out, in other words, how much of the outcome's informational value is obtained in advance, by application of the predictive model.

Consider the following case. Imagine we are dealing with a choice of two outcomes, α and β. Let α' denote a prediction of outcome α, and β' a prediction of β. If the two events are objectively equiprobable, the informational value of each is $-\log_2 \frac{1}{2} = 1$ bit. If the predictive model gives rise to α' alone, and α is the outcome, we then have

$$I(\alpha') = I(\alpha) = 1 \text{ bit}$$

The value of the predictive model is 1 bit. Similarly, if the model gives rise to β' and β is the outcome, we have

$$I(\beta') = I(\beta) = 1 \text{ bit}$$

Again the model is worth 1 bit. If the model gives rise to both predictions together, its informational value is zero by definition. Predicting both outcomes is equivalent to making no prediction at all—the prediction merely recapitulates the choice. Thus

[4] For example, if event x has probability 0.25, we expect it to be drawn from a choice of $\frac{1}{0.25} = 4$ alternatives, for which we will need $-\log_2 0.25 = 2$ binary digits to signal the outcome.

[5] The term is original due to John Tukey.

[6] In developing the framework, Shannon was particularly concerned with problems of telecommunication (Shannon, 1956). Events are conceptualized as messages sent from a sender to a receiver by means of a communication channel. Theoretical results of the framework then relate to fundamental limits on channel capacity, and the way statistical noise can be eliminated by introduction of redundancy.

[7] The present paper makes no distinction between a subjective probability and a Bayesian 'degree of belief'; whether there is a valid distinction to be made is unclear (cf. Ramsay, 1990).

[8] For example, for purposes of analyzing perceptual organization, von Helmholtz (1860/1962) takes probabilities to be inherently objective. For purposes of analyzing musical creativity, Temperley (2007) takes them to be inherently subjective.