Brain and Cognition 112 (2017) 3–12

Contents lists available at ScienceDirect

Brain and Cognition

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

Explanatory pluralism: An unrewarding prediction error for free energy theorists

Matteo Colombo [a,*], Cory Wright [b]

[a] Tilburg Center for Logic, Ethics & Philosophy of Science, Tilburg University, PO Box 90153, 5000 LE Tilburg, The Netherlands
[b] Department of Philosophy, McIntosh Humanities Building 917, California State University, Long Beach, 1250 Bellflower Boulevard, Long Beach, CA 90840-2408, USA

ARTICLE INFO

Article history:
Received 25 May 2015
Revised 12 February 2016
Accepted 13 February 2016
Available online 20 February 2016

Keywords:
Anhedonia
Dopamine
Explanation
Explanatory pluralism
Free energy
Incentive salience
Predictive processing
Reduction
Reward prediction error
Unification

ABSTRACT

Courtesy of its free energy formulation, the hierarchical predictive processing theory of the brain (PTB) is often claimed to be a grand unifying theory. To test this claim, we examine a central case: activity of mesocorticolimbic dopaminergic (DA) systems. After reviewing the three most prominent hypotheses of DA activity—the anhedonia, incentive salience, and reward prediction error hypotheses—we conclude that the evidence currently vindicates explanatory pluralism. This vindication implies that the grand unifying claims of advocates of PTB are unwarranted. More generally, we suggest that the form of scientific progress in the cognitive sciences is unlikely to be a single overarching grand unifying theory.

1. Introduction

The hierarchical predictive processing theory of the brain (PTB) claims that brains are homeostatic prediction-testing mechanisms, which function to minimize the errors of their predictions about the sensory data they receive from their local environment. The mechanistic function of minimizing prediction error is constituted by various monitoring- and manipulation-operations on hierarchical, dynamic models of the causal structure of the world within a bidirectional cascade of cortical processing.

The least generic (and arguably most interesting) formulation of PTB currently available is the *free energy* formulation, which names the thesis that any self-organizing system—not just brains—must act to minimize differences between the ways it predicts the world as being, and the way the world actually is, i.e., must act to minimize prediction error.[1] Central to the free-energy formulation of PTB is the *free energy principle*, which claims that biological, self-organizing systems must act to minimize their long-term average free energy (Friston, 2010: 127), where *free energy* refers to an information-theoretic measure that bounds the negative log probability of sampling some data given a model of how those data are generated.

Advocates of PTB are enthusiastic about the expected payoffs of their theory. In Friston's words, 'if one looks at the brain as implementing this scheme [i.e., free-energy minimization], nearly every aspect of its anatomy and physiology starts to make sense' (2009: 293). Dehaene agrees: '[m]ost other models, including mine, are just models of one small aspect of the brain, very limited in their scope. [PTB] falls much closer to a grand theory' (quoted in Huang, 2008: 33). PTB is said to offer 'a deeply unified theory of perception, cognition, and action' (Clark, 2013a: 186), and even to acquire 'maximal explanatory scope' (Hohwy, 2013: 242). Over time, this enthusiasm has given way to unbridled confidence, where PTB is said to 'offer a unified approach to mental function' (Hohwy, 2014: 146) and to 'explain everything about the mind' (Hohwy, 2015: 1), and to have 'the shape of a fundamental and unified science of the embodied mind' (Clark, 2015a: 16). Others have suggested that PTB is so powerful that even partial fulfillment of these expected payoffs would radically alter the course of cognitive science (Gładziejewski, 2016).

* Corresponding author.
E-mail addresses: m.colombo@uvt.nl (M. Colombo), cory.wright@zoho.com (C. Wright).

[1] Henceforth, we shall use 'PTB' and 'free-energy formulation of PTB' interchangeably.

http://dx.doi.org/10.1016/j.bandc.2016.02.003
0278-2626/

Rather than chalking up this language to rhetorical posturing, we begin—as a measure of interpretive charity—by taking these authors at their word. So, let us call the idea that PTB is maximally explanatory, deeply unifying, and in some sense singularly fundamental—i.e., that it has the shape a so-called *grand unifying theory* (GUT)—the GUT intuition of advocates of PTB (cf. Anderson & Chemero, 2013). Since it is an open empirical question whether, and how, PTB relates to other theories and hypotheses, this question should be answered on case-by-case grounds in light of both precise explications of concepts like UNIFICATION, REDUCTION, and EXPLANATION, as well as actual scientific practice. Consequently, this paper evaluates advocates' GUT intuition via examination of a central case: activity of mesocorticolimbic dopaminergic (DA) systems. We argue for two interrelated conclusions: first, that several current hypotheses of DA are mature, competitively successful alternatives in a pluralism of explanatory resources, and second, that the explanatory pluralism vindicated by these hypotheses is inconsistent with advocates' GUT intuition.

Explanatory pluralism enjoys several characterizations. What they all share is a commitment to denying that 'the ultimate aim of science is to establish a single, complete, and comprehensive account of the natural world (or the part of the world investigated by the science) based on a single set of fundamental principles' (Kellert, Longino, & Waters, 2006: x). In the case of DA activity, we argue that the GUT intuition shared by advocates of PTB is currently unwarranted. Our argument has the form of an abductive inference: if pluralism were correct, then the scientific investigation of DA activity would demand multiple, diverse epistemic tools without a requirement to collapse into a fundamental theory of how brains work. As this multiplicity and diversity are just what is observed in current scientific practice, pluralism is vindicated. Since explanatory pluralism is inconsistent with the reductive and monistic claims of free energy theorists, our argument calls into the status of PTB as a grand unifying theory.

In Sections 2 and 3, we rehearse several constructs central to PTB and articulate the conditions under which PTB would count as a grand unifying theory. We highlight three prominent hypotheses of DA in Section 4, and explain in Section 5 why current scientific practice supports more explanatory pluralism than the GUT intuitions of advocates of PTB. In Section 6, we conclude.

2. PTB: nuts and bolts

Although the general insight that brains perform predictions has a long and heterogeneous tradition, PTB is associated with recent work by Friston and Stephan (2007), Friston (2009), Friston (2010), Hohwy (2013), and Clark (2013a), Clark (2013b, Clark (2015b). While their respective formulations are inequivalent and have different consequences, advocates have converged on several basic commitments and a fixed stock of theoretical terms.[2] Two of these commitments are, firstly, that brains are prediction-testing mechanisms, and secondly, that brains produce psychological phenomena by constantly attempting to minimize prediction errors.

To articulate these commitments, several terms require clarification—foremost being *prediction*, which is understood as a (homonymous) technical term with no semantic relation to its ordinary sense. PTB defines *prediction* (or *expectation*) within the context of probability theory and statistics as the weighted mean of a random variable, which is a magnitude posited to be transmitted downwards as a driving signal by the neurons comprising pairwise levels in the cortical hierarchy.

The term *prediction error* refers to magnitudes of the discrepancies between predictions about the value of a certain variable and its observed value (Niv & Schoenbaum, 2008). In PTB, prediction errors quantify mismatches between expected and actual sensory data (or sensory input), as the brain putatively encodes probabilistic models of the world's causal structure in order to predict its sensory data. If predictions about sensory data are not met, then prediction errors are generated so as to tune brains' probabilistic models, and to reduce discrepancies between what was expected and what actually obtained.

In information theory, *entropy* refers to a measure of the uncertainty of random quantities. That a probability distribution (or a statistical model) has low entropy implies that data sampled from that distribution are relatively predictable. If probability distributions are used to describe all possible sensory states that an adaptive agent could instantiate, then the claim that adaptive agents must resist a tendency to disorder can be reconceived as the claim that the distributions of their sensory states should have low entropy. If probability distributions of the possible sensory states of adaptive agents have low entropy, those agents will occupy predictable states.

The term *predictable state* concerns the amount of surprisal associated with that state, which quantifies how much information it carries for a system. *Surprisal* refers to the negative log probability of an outcome, and, like entropy, is a measure relative to probability distributions (or statistical models). When applied to adaptive agents, entropy (or average surprisal) is construed as a function of the sensory data they receive and of their internal models of the environmental causes of that data.

Computationally-bounded agents, however, can only minimize surprisal indirectly by minimizing free energy. Given how many variables (and their possible values) can be associated with agents' sensory states, minimizing surprisal directly is intractable. Computationally-bounded agents are instead said to minimize surprisal indirectly by minimizing free energy. Free energy is an information-theoretic quantity that can be directly evaluated and minimized, and 'that bounds or limits (by being greater than) the surprisal on sampling some data given a generative model' (Friston, 2010: 127).

A *generative model* is a statistical model of how data are generated, which, in PTB, consists of prior distributions over the environmental causes of agents' sensory data and generative distributions (or likelihoods) of agents' sensory data given their environmental causes. By providing a bound on surprisal, minimizing free energy minimizes the probability that agents instantiate surprising states. Since agents' free energy depends only on their sensory data and on their internal models of the causes of their sensory data, computationally-bounded adaptive agents can avoid surprising states (and, presumably, live longer) by directly minimizing their free energy.

The free energy principle is said to logically entail other principles incorporated within PTB—namely, the so-called *Bayesian brain hypothesis* and principles of predictive coding (Friston, 2013: 213). For its part, the Bayesian brain hypothesis was motivated by the increased use and promise of Bayesian modeling to successfully answer questions about biological perception. 'One striking observation from this work is the myriad ways in which human observers behave as optimal Bayesian observers' (Knill & Pouget, 2004: 712). A fundamental implication for neuroscience is that 'the brain represents information probabilistically, by coding and computing with probability density functions or approximations to probability density functions' (Knill & Pouget, 2004: 713; Colombo & Seriès, 2012).

Predictive coding names an encoding strategy in signal processing, whereby expected features of an input signal are suppressed and only unexpected features are signaled. Hierarchical predictive coding adds to this strategy the assumption of a hierarchy of processing stages. By implication, PTB maintains that brains are

[2] We leave it open as to whether our argument applies to formulations that are not committed to the free-energy principle.