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/© 2016 Elsevier Inc. All rights reserved.

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.

hierarchically organized such that the activity of every layer in the cortical hierarchy predicts the activity in the adjacent layer immediately below it.

Perception, action, and cognition are thus said to be produced by the same kind of process, *viz*. by the interplay between downward-flowing predictions and upward-flowing sensory signals in the cortical hierarchy. At each stage, inputs from the previous stage are processed as degree of deviation from expected features, and only unexpected features are signaled to the next stage. Applied iteratively in hierarchically organized neural networks, this compact processing scheme leads to bidirectional processing, where feed-forward connections convey information about the difference between what was expected and what actually obtained—i.e., prediction error—while feedback connections convey predictions from higher processing stages to suppress prediction errors at lower levels. So, processing at each stage signals difference between expected and actual features to the next stage up the hierarchy; and each stage sends back to the one below it the expected features.

This signal-processing strategy was originally invoked in neuroscience to explain extra-classical receptive-field effects of neurons in primary visual cortex and beyond (Lee & Mumford, 2003; Rao & Ballard, 1999). Hierarchical predictive coding provides for a model of the functional asymmetry of inter-regional visual cortical connections, where forward connections run from lower to higher layers, driving neural responses, and where backward connections run from higher to lower layers, playing a modulatory role. According to this model, expectations about the causal structure of local environments are encoded in the backward connections, while forward connections provide feedback by transmitting sensory prediction-error up to higher levels.

In hierarchical architectures, under restrictive (Gaussian) assumptions, the relative influence of bottom-up prediction errors and top-down predictions is controlled by *precision*, which, in a statistical context, is defined as the reciprocal of the variance of a distribution. Precision can be operationalized as a measure of uncertainty of the data due to noise or random fluctuations. In the context of PTB, precision modulates the amplitude of prediction errors. A more precise prediction error will have more weight on updating the system's models of the world. Precision-weighting on prediction errors allows brains to tip the balance between sensory input and prior expectations at different spatial and temporal levels in the processing hierarchy, in a way that is context- and task-sensitive.

3. PTB as a grand unifying theory?

In Section 1, we observed that its advocates intuit that PTB is a grand unifying theory, and then articulated that theory's basic constructs in Section 2. Unfortunately, evaluating their GUT intuition is difficult. Advocates of PTB have left unspecified the conditions under which they would take their assertion that PTB is a grand unifying theory to be true. While this lack of detail makes it hard to know what they have in mind (for a related worry see Rasmussen & Eliasmith, 2013), a curious aspect of the literature is that advocates are perfectly sanguine about broaching traditional and well-characterized topics in philosophy of science, such as unification or the nature of inter- and intratheoretical relationships; but when called to elaborate and justify their claims about such relations, they seek shelter in a dark room.

Fortunately, these terms and concepts have precise characterizations and long-standing analyses. So, before turning to the details of our test case in Section 4, we remedy this situation by drawing on the relevant literature in philosophy of science to characterize the main conditions under which advocates' GUT intuition would be satisfied.

3.1. Unification, monism, reductionism

For those who have it, the GUT intuition runs deep. For example, Friston and Stephan (2007: 418) asserted, '[t]he payoff for adopting [PTB] is that many apparently diverse aspects of the brain's structure and function can be understood in terms of one simple principle'. Friston, Daunizeau, Kilner and Kiebel (2010: 255) asserted that the free energy principle is a 'unifying approach to perception and action', which enjoys a simple and biologically plausible implementation. Clark (2013: 242) re-asserted the point, claiming that '[PTB] is a deeply unified theory of perception, cognition, and action'; likewise, Hohwy (2014: 146) claimed that PTB has maximal explanatory scope' and is 'a unified approach to mental function [tout court]'. Hohwy (2014: 146) also asserted that ' [PTB] is the only theory that can really make inroads on the problem of perception (or the problem of content)' and has the advantage of either 'having no clear alternative' or having as its alternative the 'conjunction of all theories of particular aspects of mental life'. More recently, Hohwy (2015: 8–9) added that the theory is not only unifying and singularly fundamental, but also reductionist in at least two senses: 'this [i.e., PTB] is all extremely reductionist, in the unificatory sense, since it leaves no other job for the brain to do than minimize free energy so that everything mental must come down to this principle. It is also reductionist in the metaphysical sense, since it means that other types of descriptions of mental processes must all come down to the way neurons manage to slow sensory input'.[3]

To summarize, the free energy formulation of PTB is grounded in 'one simple principle', having a 'simple and biologically plausible implementation'; PTB affords a 'unified' and 'reductionist' 'theory of perception, cognition, and action', with 'maximal explanatory scope', whose only 'alternative is the conjunction of all theories of mental [phenomena]'. So, with Hohwy, Clark, Friston, and others, let T be a grand unifying theory only if T entails explanatory unificationism, monism, and reductionism. *Unificationism* names the thesis that explanations are derivations that unify as many descriptions of target phenomena to be explained, $\varphi_1, \ldots, \varphi_n$, from as few stringent argument patterns as possible (Kitcher, 1989; see Colombo & Hartmann, 2015 for an assessment of the idea that Bayesian modeling unifies cognitive science). *Monism* names the thesis that any given φ will always have exactly one adequate explanation based on a single set of fundamental principles. Closely related to the idea of explanatory unificationism and monism is the third thesis, *reductionism*, elaborated in detail in the next subsection.

For now, the general point is that each thesis is an individually necessary condition on a theory T satisfying advocates' GUT intuition, which can be made precise by drawing on the well-established literature in philosophy of science. A theory failing to unify descriptions of $\varphi_1, \ldots, \varphi_n$ would not then be a unifying theory; a theory requiring other theories to do so would not be monistic; and a theory that was not reductively more basic than the descriptions so unified could not be a grand unifying theory in advocates' intended sense.

3.2. Epistemic reductionism

To make the case for PTB being a single unifying and reductionist brain theory, Friston (2009, 2010) discusses the free energy formulation of PTB in relation to several principles, hypotheses, and

[3] Hohwy categorizes the first sense as 'a kind of theory reduction' and 'explanatory unification' (2015: 2). Immediately following this conflation of reductionist and unificationist theses is a confusion of functional decomposition for ontological reduction, and then a second conflation of ontological reduction and reductive explanation (Hohwy, 2015).

theories: e.g., the Bayesian brain hypothesis, infomax and the theory of efficient coding, the Hebbian cell assembly and correlation theory, neural Darwinism and value learning, and optimal control theory and game theory. But what 'relation' is that, exactly? Friston doesn't say.

Ontological reduction, which is a relation between posited entity types and tokens (whether objects, properties, states, etc.), is one answer. But the question, as suggested by Friston's discussion, concerns relations between theories. So, a better answer is epistemic reduction, which concerns how the concepts, explanations, or bodies of knowledge pertaining to one scientific domain are related to the concepts, explanations or bodies of knowledge pertaining to another scientific domain.

Approaches to epistemic reduction are normally grouped into two basic (non-exclusive) categories: *theory reduction* and *reductive explanation* (see, e.g., Brigandt & Love, 2012: Section 3). Theory reductions specify how pairwise theories will comport with one another after the fact of maturity and development. Reductive explanation is a kind of explanation, where fragments of a theory, generalizations of varying scope, or mechanisms or phenomena described in a higher-level vocabulary are explained by appeal to lower-level theories, concepts, principles, or mechanisms. With this distinction, advocates of PTB can plausibly either lay claim to reductive explanations of various neurobiological and psychological phenomena, or else to reducing higher-level theories of them. Let us clarify each in turn.

Contemporary models of theory reduction take reduction to be an indirect relation between a reducing theory T_{B} and a corrected analogue T^*_{R} of a theory T_{R}, specified within the framework of T_{B}. T^*_{R} is derived from T_{B}. The strength of the analogical mapping between T^*_{R} and T_{R} is associated with a space of theory-relations ranging from 'perfectly smooth' to 'bumpy' reductions. Where T^*_{R} is a perfectly equipotent isomorphic image of T_{R}, we have cases of 'perfectly smooth' reductions. Where the relationship between T^*_{R} and T_{R} is poorly analogous, we have 'bumpy' reductions. Unlike 'bumpy' reductions, which involve the replacement of the reduced theory T_{R} with the reducing theory T_{B}, 'perfectly smooth' reductions retain the reduced theory T_{R} since T_{B} and T_{R} are believed to characterize the exact same kinds of entities or properties, albeit with different concepts (Bickle, 1998; Bickle, 2003; Endicott, 1998; McCauley, 1986; McCauley, 1996; Wright, 2000).

According to what is arguably the most prominent version of reductive explanation, higher-level explanations in psychology only play a heuristic role in developing lower-level explanations in cellular and molecular neuroscience and are inevitably abandoned once lower-level explanations obtain: 'there is no need to evoke psychological causal explanations, and in fact scientists stop evoking and developing them, once real neurobiological explanations are on offer' (Bickle, 2003: 110; Kaiser, 2015). Hence, in the context of reductive explanation, higher-level explanations are (eventually) extinguished (Wright, 2007). This is unlike the context of theory reduction, where cases falling toward the retentive end of the inter-theoretic reductive spectrum do not extinguish the reduced theory T_{R} so much as they vindicate and preserve it.

3.3. Initial conceptual difficulties

Friston's (2009, 2010) remarks that the free energy principle is related to several theories—to optimal control theory and game theory, to the Bayesian brain 'hypothesis', to infomax and the theory of efficient coding, to the Hebbian cell assembly and correlation theory, and to neural Darwinism and the theory of value learning—suggest that several theories in cognitive neuroscience may be derived from PTB. However, even if PTB can serve as a basis for theory reduction, it is important to clarify that the reduction of one theory to another does not entail a commitment to reductive explanation. As reductionists themselves acknowledge, 'one can predict an intertheoretic reduction without tying one's methodological practices to reductive explanations. An intertheoretic reductionist can agree wholeheartedly with this methodological point. He need have no commitment to the exclusive use of reductive explanation' (Bickle, 1998: 153–4). Hence, nothing about PTB's being a reducing theory T_{B} rules out the reliance on numerous, diverse kinds of epistemic and explanatory tools that are not tied all together into one single fundamental principle.

The logical independence of theory reduction and reductive explanation is important because it allows advocates to maintain that PTB reduces other theories, and that the intended form of explanation afforded by PTB is mechanistic rather than reductionistic. Indeed, Hohwy states that 'one of the great attractions of the [prediction error] scheme is that it lends itself to a very mechanistic approach' (2013: 8).

The difference is subtle but important: mechanistic explanation proceeds using not only reductive explanatory strategies like decomposition and localization, but also anti-reductive strategies like composition and contextualization (Bechtel & Wright, 2009; Wright, 2007). For example, attempts to explain an activity or function at one hierarchical level in terms of the orchestrated operations of component parts at lower levels sometimes runs aground; for mechanists, explanatory success may come from reconstructing a given decomposable higher-level activity as a lower-level operation that instead composes a mechanistic activity at a higher-level of description and analysis. That is, explanatory success sometimes comes, not from trying to reduce it to some set of organized components operating at increasingly lower-levels, but instead situating it as a component operating in the context of some higher-level activity or function.

The mechanistic aims of PTB—coupled with logical independence of theory reduction and reductive explanation—implies that the epistemic reductionism inherent in advocates' GUT intuition is best understood in terms of theory reduction rather than reductive explanation. And it is here that PTB currently falls short of making its case; for it is not enough that PTB enjoys a mathematical formulation and 'relates' to other theories. Also necessary is an exact formulation in the relevant idiom of the reduced and reducing theories—for example, in terms of sets of models, with reduction and replacement defined in terms of empirical base sets, blurs and other set-theoretic relations, and 'homogenous' or 'heterogeneous ontological reductive links' between members (for further details, see Bickle, 1998). Minimally, this exact formulation involves the laborious fivefold task of (i) reconstructing PTB as a base theory T_{B}, (ii) reformulating various reduced theories $T_{\mathrm{R}1}$, ..., $T_{\mathrm{R}n}$, (iii) constructing and correcting analogues $T^*_{\mathrm{R}1}$, ..., $T_{\mathrm{R}n}{}^*$, (iv) demonstrating the derivations of $T^*_{\mathrm{R}1}$, ..., $T_{\mathrm{R}n}{}^*$ from within PTB, and then (v) demonstrating the mappings from $T^*_{\mathrm{R}1}$, ..., $T_{\mathrm{R}n}{}^*$ to $T_{\mathrm{R}1}$, ..., $T_{\mathrm{R}n}$.

Because advocates have never articulated this formulation or attempted this fivefold task, and only allude to it if at all, assessing whether PTB in fact reduces any other higher-level theories is not yet possible. Of course, this is not yet to say that it cannot be done. But until advocates do the work necessary to demonstrate genuine intertheoretic reductions, rather than just suggestively assert them, their GUT intuition is unwarranted. A detailed examination of the case of dopamine function in Sections 4 and 5 will substantiate this assessment.

3.4. Explanatory pluralism

Opposed to explanatory monism is *explanatory pluralism*, which denies that for any phenomenon there will always be exactly one single, complete, comprehensive explanation based on a single set of fundamental principles. As Looren de Jong characterized it,

> [explanatory pluralism] holds that theories at different levels of description, like psychology and neuroscience, can co-evolve, and mutually influence each other, without the higher-level theory being replaced by, or reduced to, the lower-level one. [. . .] Explanatory pluralism thus recognizes that various interesting interlevel relations can exist beyond reduction and elimination (2001: 731–732).

Similarly, Van Bouwel et al. write,

> '[e]xplanatory pluralism consists in the [two] claims that the best form (and level) of explanation depends on the kind of question one is willing to answer by the explanation, and that in order to answer all explanation-seeking questions in the best way possible we will need more than one form (and level) of explanation (2011: 36).

What these and other formulations have in common is what they imply: rather than a grand unifying theory that fits all explanatory interests and purposes, for a great many phenomena, there are multiple adequate explanations or models that are differentially assessed according to different norms of assessment without the requirement that they all tie into some fundamental principle or norm of assessment (Kellert et al., 2006; Wright, 2002). Underlying each unique explanation or model are different vocabularies that create and expand new ways of conceptualizing phenomena, and additional conceptualizations invite theoretical competition. For the explanatory pluralist, this kind of multidirectional selection pressure on scientific practice can be exerted only with the simultaneous pursuit of different kinds of explanations of a given phenomenon at multiple levels, in different domains or fields, using a variety of techniques and methods of interpreting evidence; and it is precisely this competition and selection pressure that is essential for scientific progress. Thus, plurality of explanation constitutes not a deficiency to be overcome by unification, but a patchwork of explanations whose unification would incur significant loss of content and inhibit scientific progress.

Like reductionism, explanatory pluralism is often justified by appeal to actual scientific practice. For examples, McCauley and Bechtel (2001) detailed research on visual processing at different levels of explanation to show how it productively stimulated increasingly-sophisticated and empirically-testable psychoneural claims. Bechtel and Wright (2009) show that explanatory monism misdescribes the psychological sciences, while Dale, Dietrich, and Chemero (2009) remind us that cognitive science, by its multidisciplinary nature, generates explanations that are inherently pluralistic. Brigandt (2010) observes that, in evolutionary developmental biology, whether the various relations of reduction, integration, unification, synthesis, etc. serve as a regulative ideal or scientific aim varies with the problem investigated. So, following this standard justificatory strategy of examining actual scientific practice, we turn to a central empirical test case: the mechanistic activity of the mesocorticolimbic system, sometimes referred to as *brain reward function*.

This case is ideal for testing the GUT intuitions of advocates of PTB and for contrasting them with explanatory pluralism, because—as we will see—there are several distinct, apparently competing, mature models of dopaminergic activity. Indeed, advocates of PTB have recently put forward a model of dopaminergic activity too. According to their model, 'dopamine may have a singular mechanism of action and computational function' (Friston et al., 2012: 14). This function is to control 'the precision or salience of (external or internal) cues that engender action' (Friston, Shiner, et al., 2012: 1). By associating dopamine with 'precision or salience,' PTB is said to 'absorb' alternative models associating dopamine with *reward prediction error*, and to 'connect to' other models in psychiatry and psychology positing that dopamine regulates *incentive salience* and *anhedonia*.

Examining the relationship between PTB and these alternative models allows us to evaluate the GUT intuition in the light of actual scientific practice. After reviewing some aspects of the dopaminergic system in Section 4, we show that the case of dopamine exemplifies how actual scientific practice vindicates explanatory pluralism, and of how PTB is not the grand unifying theory it is too often said to be.

4. Dopaminergic operations in brain reward function

Dopamine (DA) is a catecholaminergic neurotransmitter released by DA neurons, which are phylogenetically old—found in all mammals, birds, reptiles, and insects, and so primarily located in evolutionarily older parts of the brains, particularly in two nuclei of the midbrain: the ventral tegmental area (VTA) and substantia nigra pars compacta (SNc).

Anatomically, the axons of DA neurons project to numerous cortical and subcortical areas. One of these is the nigrostriatal pathway. It links the SNc with the striatum, which is the largest nucleus of the basal ganglia in the forebrain and which has two components: the putamen and the caudate nucleus (CN). The CN, in particular, has the highest concentration of DA of any neural substructure. Another pathway is the mesolimbic, which links the VTA to the nucleus accumbens (NAc) and other structures in the forebrain, external to the basal ganglia, such as the amygdala and prefrontal cortex (PFC). Approximately 85% of the mesolimbic pathway connecting the VTA and NAc is composed of DA neurons.

Electrophysiologically, DA neurons show two patterns of firing activity—tonic and phasic—that modulate levels of extracellular DA. Tonic activity consists of regular firing patterns of ≈1–6 Hz that maintain a slowly-changing, extracellular, base-level of extracellular DA in afferent brain structures. Phasic activity consists of a sudden change in the firing rate of DA neurons, which can increase up to ≈20 Hz and cause transient increases in extracellular DA concentrations.

DA-specific receptors control neural signaling in the targets of DA neurons. There are at least five receptor subtypes—DA_1 and DA_2 being the most important—grouped into several families. Each has different biophysical and functional properties that affect many aspects of cognition and behavior, including motor control, learning, attention, motivation, decision-making, and mood regulation.

DA neurons are crucial components of the mesolimbic and nigrostriatal systems, which generally yoke the directive and hedonic capacities of motivation and pleasure to motor abilities for ascertainment behavior. Around 80% of DA neurons are synchronically activated in mechanisms producing reward (Schultz, 1998), and pharmacological blockade with DA antagonists induces impairments in reward functionality. DA also has been implicated in various pathologies: e.g., Parkinson's disease, major depressive disorder, schizophrenia, attention-deficit hyperactivity disorder, and addiction.

Since the 1950s, several specific hypotheses have been advanced about the function of DA neurons. Of the major competing hypotheses, three are currently prominent: the anhedonia (HED), incentive salience (IS), and reward-prediction error (RPE) hypotheses.

4.1. Anhedonia (HED)

In his *Varieties*, James characterized *anhedonia* as 'melancholy in the sense of incapacity of joyous feeling' (1902: 147). More recently, the term has been used within psychiatry to denote a degraded capacity to experience pleasure; so following James, anhedonic individuals are described as exhibiting a lack of enjoyment from, engagement in, or energy for life's experiences' and 'deficits in the capacity to feel pleasure and take interest in things' (DSM-V: 817).

While early work in functional neuroanatomy and electrophysiology promoted the idea of 'pleasure centers' (Olds, 1956), further pharmacological and neuroimaging studies failed to provide telling evidence for a positive and direct causal contribution of DA to the capacity to experience subjective, conscious pleasure (Berridge & Robinson, 1998; Salamone, Cousins, & Snyder, 1997; Wise, 2008). Yet, recent advances have shown that hedonic 'hotspots' are localized in the VP and rostromedial shell of the NAc, and that brains' ability to process pleasure implicate a complex interaction between DA operations and opioid systems (Peciña & Berridge, 2005). So, DA's role in pleasure-processing is probably indirect and modulatory.

Literature reviews (e.g., Gaillard, Gourion, & Llorca, 2013; Salamone et al., 1997) have not supported the simple hypothesis that anhedonia is robustly correlated with diminished gratification or behavioral reactions to pleasurable stimuli. This outcome has led some to argue that the very concept ANHEDONIA should be reconceptualized as a complex and disjunctive concept of 'diminished capacities to experience, pursue, and/or learn about pleasure' (Rømer-Thomsen, Whybrow, & Kringelbach, 2015). Such proposals, if they do not simply conflate what was previously distinct (Berridge & Robinson, 2003), might be understood as a form of 'conceptual re-engineering' that resurrects the explanatory power of HED by increasing its construct validity and scope, and decreasing reliance on traditional self-report measures as its evidential basis.

Either way, the thought that abnormal mesocorticolimbic DA operations are casually relevant to negative changes in the subjective pleasure associated with, or devaluation of, rewarding stimuli is only one lesser aspect of HED. In addition to impairments in hedonic experience, anhedonic states also involve motivational impairments. HED implies that normal levels of DA in these circuits are causally relevant to normal motivation, as motivational mechanisms are constituted by mesolimbic DA circuits and their projections to prefrontal areas (Der-Avakian & Markou, 2012). More regimented formulations of HED have focused on DA's relationship to selective attenuation of 'goal-directed' or motivational arousal and positive reinforcement (Wise, 1982, 2008).

HED has also helped explain DA's role in a constellation of psychopathological deficits, clinical symptoms, and psychotic disorders—notably, major depressive disorder and schizophrenia. Impairments in mesocortical DA circuits in patients with these disorders are specifically associated with the motivational deficits in anhedonia (Horan, Kring, & Blanchard, 2006; Howes & Kapur, 2009; Treadway & Zald, 2011). In patients with major depressive disorder, quantitative measures of anhedonia severity are negatively correlated with the response magnitude in the ventral striatum to pleasant stimuli, and positively correlated with the magnitude of activity in the ventromedial PFc (Gaillard et al., 2013; Keedwell, Andrew, Williams, Brammer, & Phillips, 2005). In patients with schizophrenia, regions in the right ventral striatum and left putamen show reduced responses to pleasant stimuli, and higher anhedonia scores are associated with reduced activation to positive versus negative stimuli in the amygdala and right ventral striatum (Dowd & Barch, 2010).

In summary, HED states that DA function consists in regulating motivation, arousal, and hedonic responses. Abnormal DA activities in mesolimbic and prefrontal circuits—particularly in the ventral striatum, NAc, and CP—are casually relevant factors in the motivational deficits observed in anhedonic patients (DSM-V 2013). These deficits are explained in terms of lower response activations in DA pathways and lower volume of specific DA circuits.

4.2. Incentive salience (IS)

The IS hypothesis states that afferent DA release by mesencephalic structures like the VTA encodes 'incentive' value to objects or events (Berridge & Robinson, 2003; Berridge, 2007). It relates patterns of DA activations to a complex psychological property called *incentive salience*, which is an attractive, 'magnet-like' property conferred on internal representations of external stimuli that make those stimuli appear more salient or 'attention-grabbing', and more likely to be wanted, approached, or consumed. Attribution of incentive salience to stimuli that predict rewards make both the stimuli and rewards 'wanted' (Berridge & Robinson, 1998). Because incentive salience attributions need not be conscious or involve feelings of pleasure, explanations of DA function in terms of incentive salience and of anhedonia are distinct.

In the late 1980s and 1990s, IS was offered to explain the differential effects on 'liking' (i.e., subpersonal states of pleasure or hedonic impact) and 'wanting' (i.e., incentive salience) of pharmacological manipulations of DA in rats during taste-reactivity tasks (Berridge, Venier, & Robinson, 1989). Subsequently, IS has been invoked to explain results from electrophysiological and pharmacological experiments that manipulated DA activity in mesocorticolimbic areas of rats performing Pavlovian or instrumental conditioning tasks (Berridge & Robinson, 1998; Robinson, Sandstrom, Denenberg, & Palmiter, 2005). Further, increasing DA concentrations appears to change neural firing for signals that encode maximal incentive salience, but not maximal prediction (Tindell, Berridge, Zhang, Peciña, & Aldridge, 2005).

Incentive salience has also helped explain phenomena observed in addiction and Parkinson's disease (O'Sullivan et al., 2011; Robinson & Berridge, 2008). Substance abuse, addiction, and compulsive behavior are hypothesized to be caused by over-attribution of incentive salience to drug rewards and their cues in mesocortical DA projections, due to hypersensitivity or *sensitization*, which refers to increases in drug effects caused by repeated drug administration. Sensitized DA systems would then cause pathological craving for drugs or other stimuli.

In summary, IS claims that 'DA mediates only a 'wanting' component, by mediating the dynamic attribution of incentive salience to reward-related stimuli, causing them and their associated reward to become motivationally 'wanted" (Berridge, 2007: 408). Specifically, the IS hypothesis states that DA function consists in the attribution of a subpersonal psychological property to stimuli or behavior, i.e., incentive salience. Abnormal attributions of incentive salience to stimuli or behavior are underlain by abnormal DA activity in mesocorticolimbic mechanisms and are causally relevant to addiction and compulsive behavior.

4.3. Reward prediction error (RPE)

The reward prediction error (RPE) hypothesis states that phasic firing of DA neurons in the VTA and SNc encodes reward-prediction errors (Montague, Dayan, & Sejnowski, 1996). It relates patterns of DA activation to a computational signal called *reward prediction error*, which indicates differences between the expected and actual experienced magnitudes of reward and drives decision-formation and learning for different families of reinforcement-learning algorithms (Sutton & Barto, 1998).

RPE states that DA neurons are sensitive to the expected and actual experienced magnitudes of rewards, and also to the precise temporal relationships between occurrences of both reward-predictors and actual rewards. This latter aspect connects a specific reinforcement-learning algorithm, *temporal difference* (TD), with the patterns of phasic activity of DA neurons recorded in the VTA and SNc (e.g., of awake monkeys while performing instrumental or Pavlovian conditioning tasks; see Schultz, Dayan, & Montague, 1997).

TD-learning algorithms are driven by differences between temporally successive estimates (or predictions) of a certain quantity—e.g., the total amount of reward expected over the future. At particular time steps, estimates of this quantity are updated to conform to estimates at the next time step. The TD-learning algorithm outputs predictions about future values, and then compares them with actual values. If the prediction is wrong, the difference between predicted and actual value is used to learn.

RPE fits many neurobiological results in learning and decision-making tasks (Niv, 2009; Colombo, 2014; Glimcher, 2011). If correct, then neurocomputational mechanisms—partially constituted by phasic operations of midbrain DA neurons—execute the task of learning what to do when faced with expected rewards and punishments, generating decisions accordingly. DA would then play a causal role in signaling reward prediction errors and selecting actions to increase reward.

In summary, RPE posits 'a particular relationship between the causes and effects of mesencephalic dopaminergic output on learning and behavioral control' (Montague et al., 1996: 1944). RPE states that the function of DA in the VTA and SNc during reward-based learning and decision-making consists in computing reward prediction errors.

5. Pluralism vindicated

5.1. Three hypotheses

The HED, IS, and RPE hypotheses advance inequivalent claims—each with different implications—regarding the causal and functional profile of DA operations. While each hypothesis has been partially corroborated by an array of different kinds of evidence from humans and other animals, none provide a comprehensive explanation of DA complexities based on a single set of fundamental principles. Instead, each one provides a partial model of DA, and different scientific communities rely on these different models of DA for different explanatory purposes; as we shall show, none can be intertheoretically reduced or 'absorbed' into PTB without explanatory loss.

HED makes general claims about relations between anhedonic symptoms and the disruption of DA signaling in limbic and prefrontal circuits, and draws much of its evidential base from qualitative models and experimental designs used to investigate psychiatric disorders. Psychiatry relies on ANHEDONIA for characterizing and diagnosing two widespread mental disorders—e.g., schizophrenia and depression—and cannot exchange this partially qualitative construct with either INCENTIVE SALIENCE or REWARD PREDICTION ERROR without thereby suffering decreases in explanatory power.

The IS hypothesis makes stronger claims than HED. It denies DA's regulatory role and impact in anhedonic symptomology. It is also at odds with RPE, and so is not easily integrated into either competitor: 'to say dopamine acts as a prediction error to cause new learning may be to make a causal mistake about dopamine's role in learning: it might [...] be called a 'dopamine prediction error'' (Berridge, 2007: 399). Like HED, the IS hypothesis is under-constrained in several ways: it has not localized the mechanistic componency of incentive salience attribution, and is uncommitted as to possible different roles of phasic and tonic dopaminergic signaling. Finally, it is not formalized by a single model that yields quantitative predictions. And yet, affective psychology and neuroscience have adopted incentive salience as helpful for marking a distinction between subdoxastic states of liking and wanting, and has helped clarify the role of DA operations in the NAc shell as well as helped explain drug addiction, changes in conative and motivational states, and eating disorders.

The RPE hypothesis is quantitatively and computationally more exacting, borrowing concepts like REWARD PREDICTION ERROR from reinforcement learning. As formulated by Montague et al. (1996), its scope is qualified: it concerns phasic VTA DA activity, and does not claim that all DA neurons encode only (or in all circumstances) reward prediction errors. Neither does it claim that prediction errors can only be computed by DA operations, nor that all learning and action selection is executed using reward prediction errors or is dependent on DA activity. Given these caveats, RPE, which is arguably a major success story of computational neuroscience (Colombo, 2014), may be reducible to PTB only insofar as DA operations other than encoding reward prediction errors are neglected. But what does PTB claim, exactly, about DA?

5.2. PTB, dopamine, and precision

Advocates of PTB intuit that their theory is a foundational base theory T_B that can intertheoretically reduce and unify the three previously mentioned DA hypotheses (and many others besides). This 'absorption' occurs in two main steps. First, PTB is said to explain away posits like REWARD and COST FUNCTION (Friston, Samothrakis, & Montague, 2012). Second, the diverse roles of DA are said to be fully explained by a single mechanism that neither computes cost functions nor represents value (Friston, Shiner, et al., 2012; Friston et al., 2014).

In order to demonstrate these two steps, Friston, Samothrakis, et al. (2012) construe decision-making under uncertainty as a partially observable Markov decision process (POMDP), where agents' tasks are to make optimal decisions in uncertain, dynamic environments. Typically, task solutions consist in finding optimal policies for agents, which maximize some cumulative function of the rewards received in different environments, and then in specifying which action agents will choose as a function of the environment they find themselves. Because task solutions need not involve rewards or cost functions—optimal policies are in principle replaceable by expectations about state transitions—Friston, Samothrakis, et al. (2012) attempt to demonstrate mathematically how optimal decisions in POMDP can be made. Basically, instead of maximizing expected reward, agents make optimal decisions by minimizing a free energy bound on the marginal likelihood of observed states. Their perception and action minimize the surprisal associated with their sensory input: perception minimizes exteroceptive prediction error, while action minimizes proprioceptive prediction error. To the extent that prediction error (or surprisal) is minimized, agents act to fulfill prior 'beliefs' about transitions among states of the world. By fulfilling prior beliefs about state transitions, agents avoid surprising exchanges with the environment that can disrupt their physiological and ethological homeostasis. Optimal decision-making would thus consist in 'fulfilling prior beliefs about exchanges with the world [... while] cost functions are replaced (or absorbed into) prior beliefs about state transitions' (Friston, Samothrakis, et al., 2012). In other words, Friston et al. contend that control problems associated with POMDP can be formulated as problems of Bayesian inference. Action aims at producing the most likely sensory data given 'beliefs' about state transitions, instead of producing valuable outcomes.

If theoretical terms like *reward* and *value* are eliminated via mathematical 'absorption' in favor of *prior belief*, then RPE—which implies that behavior is optimal relative to some reward or cost function—is disqualified as an adequate representation of the function of DA. Indeed, pace PTB, DA release must exclusively encode the *precision* of representations of bottom-up sensory input. Specifically, changes in DA levels in subcortical circuits will produce changes in the synaptic gain of their principal cells, leading to changes in the representational precision encoded by those cells. The hypothesis that DA encodes precision or salience as post-synaptic gain is said not only to explain all aspects of DA operations, but to afford a single mechanism that 'provides a unifying perspective on many existing theories about dopamine' (Friston, Shiner, et al., 2012).

PTB's unifying perspective on existing theories of DA has various consequences for the three hypotheses described in Section 4. (For a more general critical assessment see Gershman & Daw, 2012.) By 'absorbing' the semantics of *reward* and *value* into *prior belief*, PTB reconceptualizes *reward* as 'just familiar sensory states' (Friston, Shiner, et al., 2012: 2). Although several kinds of 'familiar sensory states' are not rewarding, this move would at least partially reduce or eliminate the RPE model of DA, while attempting to 'explain why dopaminergic responses do not behave as reward prediction errors generally' (Friston, Shiner, et al., 2012: 17).

Where *reward* gets replaced by *prior belief*, PTB associates the concept of precision with that of salience, which suggests that it 'can also connect to constructs like *incentive salience* in psychology and *aberrant salience* in psychopathology' (Friston, Shiner, et al., 2012: 2). But connect how? Firstly, observe that (Friston, Shiner, et al., 2012, 2012: 2) understand *salience* as 'an attribute of (probabilistic) representations that determines the confidence or certainty about what is represented'. That is, within PTB, *salience* and *precision* are construed synonymously: both refer to measures of certainty about hidden states. Within IS, however, *incentive salience* is characterized as a 'magnet-like' property that makes these external stimuli or internal representation more attention-grabbing and more likely to be wanted and approached. Hence, *incentive salience* is semantically inequivalent with *salience* or *precision*; and without explication of whether and in which sense *incentive salience* and *precision* are co-referential, the bridge principles needed to effect a reduction are not in the offing. Secondly, the kinds of phenomena referred to by *incentive salience* in psychology and psychiatry concern motivation and addiction, and the methods typically used to test IS include behavioral Pavlovian and instrumental conditioning tasks with human and nonhuman participants. By contrast, the kinds of phenomena targeted by *precision* concern attention and visual search, whereas the methods typically used to test the precision hypothesis involve theoretical work and simulations of saccadic movement and motor behavior (e.g., Friston, Shiner, et al., 2012). Altogether, these disanalogies make it unclear whether and how PTB 'connects' to the psychological phenomena picked out by IS, much less in the way required to establish the explanatory unificationist, monist, and reductionist theses.

The content and explanatory power of HED should also be reconsidered under the pressure of PTB. Recent research by Rutledge, Skandali, Dayan, and Dolan (2015) suggests that L-DOPA boosts the effects of rewards on happiness and that DA plays a subtle role in both decision-making under uncertainty and the subjective feelings of happiness related to receipt of reward. Not only do their experiments exemplify the co-evolutionary dynamics between these three hypotheses (RPE, IS, HED), which is predicted by explanatory pluralism, but because PTB lacks resources to explain the impact of DA manipulation on subjective feelings of happiness, it is unclear how PTB could accommodate their data. Indeed, within PTB, the perceived hedonic value of certain stimuli along with the phenomenology of motivation plays no reducibly causal roles in behavior; consequently, psychiatric categories grounded in the construct ANHEDONIA would then be ill-fitted to reliably inform diagnoses and treatments of mental disease. So, either the theories and explanations in which they factor ought to be scheduled for elimination—going the way of constructs like MIASMA—or else will fail to be reduced by PTB. But which? Since advocates of PTB have no demonstration to offer, and since constructs like ANHEDONIA, similarly to the construct REWARD, continue to factor in serious scientific explanations, we should infer that PTB is currently in no position to explain them away.

5.3. Higher-level help

Not infrequently, neurobiologists working on DA eschew explanatory unificationism, monism, and reductionism, and instead 'look up' levels to the psychological sciences for further evidence and constraints, such as clinical data or functional neuroimaging results to help situate large-scale task-relevant information-processing operations (Wright, 2007: 265).

For instance, Robbins and Everitt concluded that, 'even leaving aside the complications of the subjective aspects of motivation and reward, it is probable that further advances in characterizing the neural mechanisms underlying these processes will depend on a better understanding of the psychological basis of goal-directed or instrumental behavior' (1996: 228). Likewise, Berridge and Robinson suggested, 'further advances will require equal sophistication in parsing reward into its specific psychological components. [...] Neuroscientists will find it useful to distinguish the psychological components of reward because understanding the role of brain molecules, neurons, and circuits requires understanding what brains really do—which is to mediate specific behavioral and psychological functions' (2003: 507). Interestingly, Berridge also averred that further scientific breakthroughs will require development, not of lower level concepts like FREE ENERGY, but of higher-level concepts like MOTIVATION:

> [m]otivational concepts are becoming widely recognized as needed to help neuroscience models explain more than mere fragments of behavior. Yet, if our motivational concepts are seriously wrong, our quest for closer approximation to brain-behavior truths will be obstructed as much as if we had no concepts at all. We need motivational concepts, and we need the right ones, to properly understand how real brains generate real behavior (2004: 180).

These calls for increasingly sophisticated higher-level resources—a common refrain in neurobiology—are inconsistent with reductionism, and thus with PTB's drive to be a grand unifying theory. Inter alia, DA_1- and DA_2-like molecules perform numerous signaling and neuromodulatory operations, which are not fully described by any of RPE, IS, or HED; these hypotheses provide explanations for different aspects of DA operations in highly complex multi-level mechanistic explanations of brain reward function (Colombo, 2014; Wright, 2007).

5.4. Pluralism and co-evolution

All three DA models surveyed are incomplete and gappy. Yet, as explanatory pluralists predict, these lacunæ have competitively stimulated numerous extensions and refinements; in doing so, they illustrate a so-called *co-evolutionary research ideology*, where hypotheses evolve over time by borrowing from the other two or by drawing from conceptual advancements and findings in neighboring fields of inquiry (McCauley, 1996).

Consider proposals about how the neurocomputational resources of reinforcement learning (Sutton & Barto, 1998) help to formally capture the concept INCENTIVE SALIENCE and relate it more exactly to the concept REWARD PREDICTION ERROR. According to McClure, Daw, and Montague (2003), INCENTIVE SALIENCE should be formalized as expected future reward; for then some of IS's explananda, such as the dissociation between states of wanting and liking, are explained by appealing to the role of DA in biasing action selection (coherently with RPE) in a reinforcement-learning algorithm. Their proposal is that DA release assigns incentive salience to stimuli or actions by increasing the likelihood of choosing actions that lead to rewards. Accordingly, DA receptor antagonism reduces the probability of selecting any action, because estimated values for each available option would also decrease.

Whether McClure & colleagues' proposal correctly construes incentive salience (see Zhang, Berridge, Tindell, Smith, & Aldridge, 2009 for an alternative), they have initiated some co-evolution between the RPE and IS hypotheses. Specifically, the use of computational methods from reinforcement learning—informed and constrained by experimental paradigms and evidence from affective psychology and neuroscience—has helped emphasize the deep entanglement of dynamic DA operations once thought to be neatly isolable.

Dayan and Berridge (2014) drew on computational and psychological results about the interactions of Pavlovian and instrumental-learning mechanisms, traditionally associated with model-free and model-based reinforcement-learning computing, to conclude that Pavlovian learning involves its own form of model-based computations. While this conclusion blurs the distinction between Pavlovian model-free mechanism and instrumental model-based mechanism, it also calls for a re-examination of 'the role of dopamine brain systems in reward learning and motivation' (Dayan & Berridge, 2014). In keeping with a pluralist, opportunistic approach, this re-examination may focus researchers' attention on the roles of DA operations in 'tipping the balance between model-based and model-free Pavlovian predictions,' which may be experimentally studied 'using manipulations such as the reversible pre- and infralimbic lesions or dorsomedial and dorsolateral neostriatal manipulations [...] that have been so revealing for instrumental conditioning' (Dayan & Berridge, 2014).

So, after a period of competition and individual success, distinct models of DA are drawing on one another's conceptual resources and tools. The precision and flexibility of the reinforcement-learning framework, along with well-understood experimental paradigms from affective neuroscience and psychology, is leading toward theoretical and experimental integration.

These circumstances are what one would expect if explanatory pluralism were true. Again, explanatory pluralists contend that—in interlevel contexts—sets of pairwise scientific theories co-evolve and mutually influence each other without higher-level theories and hypotheses being supplanted by lower-level theories. The co-evolution of scientific research typically proceeds in ways that mutually enhance both theories, and sometimes vindicates T_R, given the fragmentary connections between theoretical projects at different levels (Dale et al., 2009; McCauley, 1986; McCauley, 1996; McCauley & Bechtel, 2001). Neuroscientists are therefore led to ask different questions about DA, and to formulate different predictions that are subsequently tested and assessed in a variety of ways (Wright, 2002).

6. Conclusion: against grand unifying theories

Neuroscientific inquiry into DA's functional profile fits well with explanatory pluralism. To arrive at this conclusion, we argued that the GUT intuitions of advocates of PTB are not satisfied; for PTB is a grand unifying theory only if PTB satisfies explanatory unificationism, monism, and reductionism with respect to central cases. In the central case of the role of DA operations in brain reward function, HED, IS, and RPE are mature, competing hypotheses; each is successful in various ways, although they are themselves not unified and none is reducible to the other. HED entails that DA operations are directly involved in motivational impairments and indirectly involved in the dysregulation of hedonic experience. IS entails that DA operations are directly involved in attributing attractiveness to representations, and in wanting and incentivizing—but not liking—rewards. And RPE entails that DA encodes the magnitude of the difference between experienced vs. actual reward. Since HED, IS, and RPE are neither unified nor reducible either to each other or to the free energy formulation of PTB without loss of explanatory content, it follows that PTB is not a grand unifying theory.

The conclusion that PTB is not the grand unifying theory its advocates make it out to be, by itself, falls short of supporting explanatory pluralism. But if explanatory pluralism about DA were true, there would exist a multiplicity of mature, competitive, and successful explanations about the DA operations that contribute to brain reward function. Since several mature, competitively successful explanations about DA operations do exist, the best explanation for this multiplicity is that explanatory pluralism is true.

This abductive argument comports well with the larger history of research in neuroscience, where the construction of grand unifying theories has proven unrewarding. In their literature review on the DA hypothesis of schizophrenia (DHS), Kendler and Schaffner arrive at a similar lesson: 'science works best when diverse theories with distinct predictions compete with one another. [I]t has been common in the history of science in general and the medical and social sciences in particular for theories to be defended with a fervor that cannot be justified by the available evidence. [...] Although very tempting, it will likely be more realistic and productive for us to focus on smaller questions, and to settle for 'bit-by-bit' progress as we clarify, in a piecemeal manner, the immensely complex web of causes that contribute to [the phenomenon to be explained]' (2011: 59). While we neglected DHS, we emphasize the same lesson. Progress in neuroscience is ill-served by fervently advancing a single grand unifying theory of mind/brain that attempts to solve all problems. Rather, it is more productive to focus experimental and theoretical research on some problems, and to generate a plurality of solutions that compete as local explanations and narrowly-conceived hypotheses.

Acknowledgments

We are grateful to Peter Dayan and to Charles Rathkopf for constructive criticisms and helpful suggestions on previous versions of this manuscript. Wright thanks Bert Leuridan and audience members at the Centre for Philosophical Psychology at Universiteit Antwerpen for their thoughtful discussion. Colombo's work on this project was supported by the Deutsche Forschungsgemeinschaft (DFG) as part of the priority program ''New Frameworks of Rationality'' ([SPP 1516]).

References

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington: APA.

Anderson, M. L., & Chemero, T. (2013). The problem with brain GUTs: Conflation of different senses of 'prediction' threatens metaphysical disaster. *Behavioral and Brain Sciences, 36*, 204–205.

Bechtel, W., & Wright, C. (2009). What is psychological explanation? In P. Calvo & J. Symons (Eds.), *Routledge companion to the philosophy of psychology* (pp. 113–130). New York: Routledge.

Berridge, K. (2004). Motivation concepts in behavioral neuroscience. *Physiology & Behavior, 81*, 179–209.

Berridge, K. (2007). The debate over dopamine's role in reward: The case for incentive salience. *Psychopharmacology, 191*, 391–431.

Berridge, K., & Robinson, T. (1998). What is the role of dopamine in reward: Hedonic impact, reward learning, or incentive salience? *Brain Research Reviews, 28*, 309–369.

Berridge, K., & Robinson, T. (2003). Parsing reward. *Trends in Neurosciences, 26*, 507–513.

Berridge, K. C., Venier, I. L., & Robinson, T. E. (1989). Taste reactivity analysis of 6-hydroxydopamine-induced aphagia: implications for arousal and anhedonia hypotheses of dopamine function. *Behavioral neuroscience, 103*(1), 36–45.

Bickle, J. (1998). *Psychoneural reduction: The new wave*. Cambridge: MIT Press.

Bickle, J. (2003). *Philosophy and neuroscience: A ruthlessly reductive account*. Dordrecht: Kluwer Academic.

Brigandt, I. (2010). Beyond reduction and pluralism: Toward an epistemology of explanatory integration in biology. *Erkenntnis, 73*, 295–311.

Brigandt, I. & Love, A. (2012). Reductionism in biology. In E. Zalta (Ed.), *Stanford encyclopedia of philosophy*. <http://plato.stanford.edu/archives/fall2015/entries/reduction-biology/>.

Clark, A. (2013a). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*, 181–253.

Clark, A. (2013b). The many faces of precision: Replies to commentaries. *Frontiers in Psychology, 4*.

Clark, A. (2015b). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford: Oxford University Press.

Clark, A. (2015a). Embodied prediction. In T. Metzinger & J. Windt (Eds.), *Open MIND: 7(T)*. Frankfurt am Main: MIND Group.

Colombo, M. (2014). Deep and beautiful: The reward prediction error hypothesis of dopamine. *Studies in History and Philosophy of Biological and Biomedical Sciences, 45*, 57–67.

Colombo, M., & Hartmann, S. (2015). Bayesian cognitive science, unification, and explanation. *British Journal for Philosophy of Science*. http://dx.doi.org/10.1093/bjps/axv036.

Colombo, M., & Seriès, P. (2012). Bayes in the brain. On Bayesian modeling in neuroscience. *British Journal for Philosophy of Science, 63*, 697–723.

Dale, R., Dietrich, E., & Chemero, A. (2009). Explanatory pluralism in cognitive science. *Cognitive Science, 33*, 739–742.

Dayan, P., & Berridge, K. C. (2014). Model-based and model-free Pavlovian reward learning: Revaluation, revision, and revelation. *Cognitive, Affective, and Behavioral Neuroscience, 14*, 473–492.

Der-Avakian, A., & Markou, A. (2012). The neurobiology of anhedonia and other reward-related deficits. *Trends in Neurosciences, 35*, 68–77.

Dowd, E., & Barch, D. (2010). Anhedonia and emotional experience in schizophrenia: Neural and behavioral indicators. *Biological Psychiatry, 67*, 902–911.

Endicott, R. (1998). Collapse of the new wave. *Journal of Philosophy, 95*, 53–72.

Friston, K. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences, 13*, 293–301.

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

Friston, K. (2013). Active inference and free energy. *Behavioral and Brain Sciences, 36*, 212–213.

Friston, K., Daunizeau, J., Kilner, J., & Kiebel, S. (2010). Action and behavior: A free-energy formulation. *Biological Cybernetics, 102*, 227–260.

Friston, K., Samothrakis, S., & Montague, R. (2012). Active inference and agency: Optimal control without cost functions. *Biological Cybernetics, 106*, 523–541.

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., & Dolan, R. J. (2014). The anatomy of choice: Dopamine and decision-making. *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*, 20130481.

Friston, K. J., Shiner, T., FitzGerald, T., Galea, J. M., Adams, R., Brown, H., Dolan, R. J., et al. (2012). Dopamine, affordance and active inference. *PLoS Computational Biology, 8*(1), e1002327.

Friston, K., & Stephan, K. (2007). Free energy and the brain. *Synthese, 159*, 417–458.

Gaillard, R., Gourion, D., & Llorca, P. (2013). Anhedonia in depression. *Encephale, 39*, 296–305.

Gershman, S. J., & Daw, N. D. (2012). Perception, action and utility: The tangled skein. In M. I. Rabinovich, K. Friston, & P. Varona (Eds.), *Principles of brain dynamics: Global state interactions* (pp. 293–312). Cambridge: MIT Press.

Gładziejewski, P. (2016). Predictive coding and representationalism. *Synthese, 193*, 559–582.

Glimcher, P. (2011). Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences, 108*, 15647–15654.

Hohwy, J. (2013). *The predictive mind*. Oxford: Oxford University Press.

Hohwy, J. (2014). Reflections on predictive processing and the mind: Interview with P. Nowakowski & P. Gładziejewski. *Avant, 5*, 145–152.

Hohwy, J. (2015). The neural organ explains the mind. In T. Metzinger & J. Windt (Eds.), *Open MIND: 19(T)*. Frankfurt am Main: MIND Group.

Horan, W., Kring, A., & Blanchard, J. (2006). Anhedonia in schizophrenia: A review of assessment strategies. *Schizophrenia Bulletin, 32*, 259–273.

Howes, O., & Kapur, S. (2009). The dopamine hypothesis of schizophrenia: Version III—The final common pathway. *Schizophrenia Bulletin, 35*, 549–562.

Huang, G. (2008). Is this a unified theory of the brain? *New Scientist* (2658).

James, W. (1902). *Varieties of religious experience: A study in human nature*. London: Longmans, Green, & Co.

Kaiser, M. (2015). *Reductive explanation in the biological sciences*. Dordrecht: Springer.

Keedwell, P., Andrew, C., Williams, S., Brammer, M., & Phillips, M. (2005). Neural correlates of anhedonia in major depressive disorder. *Biological Psychiatry, 58*, 843–853.

Kellert, S., Longino, H., & Waters, C. K. (2006). Introduction: The pluralist stance. In S. H. Kellert, H. E. Longino, & C. K. Waters (Eds.). *Minnesota Studies in Philosophy of Science, vol. 19: Scientific Pluralism* (pp. vii–xxix). Minneapolis: University of Minnesota Press.

Kendler, K., & Schaffner, K. (2011). The dopamine hypothesis of schizophrenia: An historical and philosophical analysis. *Philosophy, Psychiatry, and Psychology, 18*, 41–63.

Kitcher, P. (1989). Explanatory unification and the causal structure of the world. In P. Kitcher & W. Salmon (Eds.), *Scientific explanation* (pp. 410–505). Minneapolis: University of Minnesota Press.

Knill, D., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neuroscience, 27*, 712–719.

Lee, T., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of Optical Society of America A, 20*, 1434–1448.

Looren de Jong, H. (2001). Introduction: Symposium on explanatory pluralism. *Theory and Psychology, 11*, 731–735.

McCauley, R. (1986). Intertheoretic relations and the future of psychology. *Philosophy of Science, 53*, 179–199.

McCauley, R., & Bechtel, W. (2001). Explanatory pluralism and heuristic identity theory. *Theory and Psychology, 11*, 736–760.

McCauley, R. (1996). Explanatory pluralism and the coevolution of theories in science. In R. McCauley (Ed.), *The churchlands and their critics* (pp. 17–47). Oxford: Blackwell Publishers.

McClure, S., Daw, N., & Montague, R. (2003). A computational substrate for incentive salience. *Trends in Neuroscience, 26*, 423–428.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *The Journal of Neuroscience, 16*(5), 1936–1947.

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology, 53*(3), 139–154.

Niv, Y., & Schoenbaum, G. (2008). Dialogues on prediction errors. *Trends in Cognitive Sciences, 12*, 265–272.

Olds, J. (1956). Pleasure centers in the brain. *Scientific American, 195*, 105–116.

O'Sullivan, S. S., Wu, K., Politis, M., Lawrence, A. D., Evans, A. H., Bose, S. K., ... Piccini, P. (2011). Cue-induced striatal dopamine release in Parkinson's disease-associated impulsive-compulsive behaviours. *Brain, 134*(4), 969–978.

Peciña, S., & Berridge, K. C. (2005). Hedonic hot spot in nucleus accumbens shell: where do μ-opioids cause increased hedonic impact of sweetness? *The Journal of Neuroscience, 25*(50), 11777–11786.

Rao, R., & Ballard, D. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2*, 79–87.

Rasmussen, D., & Eliasmith, C. (2013). God, the devil, and the details: Fleshing out the predictive processing framework. *Behavioral and Brain Sciences, 36*, 223–224.

Robbins, T., & Everitt, B. (1996). Neurobehavioral mechanisms of reward and motivation. *Current Opinion in Neurobiology, 6*, 228–236.

Robinson, T. E., & Berridge, K. C. (2008). Review. The incentive sensitization theory of addiction: some current issues. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 363*(1507), 3137–3146.

Robinson, S., Sandstrom, S. M., Denenberg, V. H., & Palmiter, R. D. (2005). Distinguishing whether dopamine regulates liking, wanting, and/or learning about rewards. *Behavioral Neuroscience, 119*(1), 5–15.

Rømer-Thomsen, K., Whybrow, P., & Kringelbach, M. (2015). Reconceptualizing anhedonia: Novel perspectives on balancing the pleasure networks in the human brain. *Frontiers in Behavioral Neuroscience, 9*, 49.

Rutledge, R., Skandali, S., Dayan, P., & Dolan, R. (2015). Dopaminergic modulation of decision-making and subjective well-being. *Journal of Neuroscience, 35*, 9811–9822.

Salamone, J., Cousins, M., & Snyder, B. (1997). Behavioral functions of nucleus accumbens dopamine: Empirical and conceptual problems with the anhedonia hypothesis. *Neuroscience and Biobehavioral Reviews, 21*, 341–359.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology, 80*(1), 1–27.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science, 275*, 1593–1599.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: an introduction*. Cambridge, MA: MIT Press.

Tindell, A., Berridge, K., Zhang, J., Peciña, S., & Aldridge, J. (2005). Ventral pallidal neurons code incentive motivation: Amplification by mesolimbic sensitization and amphetamine. *European Journal of Neuroscience, 22*, 2617–2634.

Treadway, M., & Zald, D. (2011). Reconsidering anhedonia in depression: lessons from translational neuroscience. *Neuroscience and Biobehavioral Reviews, 35* [537nce a].

Van Bouwel, J., Weber, E., & De Vreese, L. (2011). Indispensability arguments in favor of reductive explanations. *Journal for General Philosophy of Science, 42*, 33–46.

Wise, R. (1982). Neuroleptics and operant behavior: The anhedonia hypothesis. *Behavioral and Brain Sciences, 5*, 39–87.

Wise, R. (2008). Dopamine and reward: The anhedonia hypothesis 30 years on. *Neurotoxicity Research, 14*, 169–183.

Wright, C. (2000). Eliminativist undercurrents in the new wave model of psychoneural reduction. *Journal of Mind and Behavior, 21*, 413–436.

Wright, C. (2002). Animal models of depression in neuropsychopharmacology qua Feyerabendian philosophy of science. In S. Shohov (Ed.). *Advances in psychology research* (Vol. 13, pp. 129–148). New York: Nova Science.

Wright, C. (2007). Is psychological explanation going extinct? In M. Schouten & H. Looren de Jong (Eds.), *The matter of the mind: Philosophical essays on psychology, neuroscience, and reduction* (pp. 249–274). Oxford: Blackwell Publishers.

Zhang, J., Berridge, K., Tindell, A., Smith, K. S., & Aldridge, J. (2009). A neural computational model of incentive salience. *PLoS Computational Biology, 5*, e1000437.