TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 7 (2015) 259–273

ISSN:1756-8757 print / 1756-8765 online
DOI: 10.1111/tops.12130

Marr's Levels Revisited: Understanding How Brains Break

Valerie G. Hardcastle,[a] Kiah Hardcastle[b]

[a]*Departments of Philosophy, Psychology, and Psychiatry & Behavioral Neuroscience, Weaver Institute for Law and Psychiatry, University of Cincinnati*
[b]*Neuroscience Program, Stanford University*

Received 30 June 2013; received in revised form 4 April 2014; accepted 14 April 2014

Abstract

While the research programs in early cognitive science and artificial intelligence aimed to articulate what cognition was in ideal terms, much research in contemporary computational neuroscience looks at how and why brains fail to function as they should ideally. This focus on impairment affects how we understand David Marr's hypothesized three levels of understanding. In this essay, we suggest some refinements to Marr's distinctions using a population activity model of cortico-striatal circuitry exploring impulsivity and behavioral inhibition as a case study. In particular, we urge that Marr's computational level should be redefined to include a description of how systems break down. We also underscore that feed-forward processing, cognition disconnected from behavioral context, and representations do not always drive cognition in the way that Marr originally assumed.

Keywords: David Marr; Computational modeling; Basal ganglia; Explanation; Stop signal reaction time task; Computational neuroscience

1. Introduction

In 1982, David Marr advocated for three independent levels of understanding for any information-processing device, like the human brain: the level of computational theory, the level of representation and algorithm, and the level of hardware implementation (p. 25). Cognitive scientists and philosophers have discussed the viability of these distinctions since then (see, e.g., Butler, 1998; Cummins, 1989; Dennett, 1994; Egan, 1991, 1992, 1995; Gilman, 1994, 1996; Harnish, 2002; Horgan & Tienson, 1994, 1996; Kitcher,

Correspondence should be sent to Valerie G. Hardcastle, Departments of Philosophy, Psychology, and Psychiatry & Behavioral Neuroscience, Weaver Institute for Law and Psychiatry, University of Cincinnati, Cincinnati, OH 45220. E-mail: valerie.hardcastle@uc.edu

1988; McClamrock, 1991; Newell, 1982, 1986; Poggio, 2012; Polger, 2002; Pylyshyn, 1984; Shagrir, 2010; Sterelny, 1990; Stevens, 2012; Verdego & Quesada, 2011). Now, more than 30 years out, we can ask: What can we learn from contemporary research in computational neuroscience that might shed light on the feasibility of Marr's original proposal?

Much current work in the cognitive sciences, including computational neuroscience, focuses on brains performing less than optimally. That is, while the research programs in artificial intelligence and the like in the 1970s aimed to articulate what cognitive processing was in ideal terms, a substantial amount of research now looks at how and why brains or other cognitive machines fail to function as they should.[1] This focus on impairment affects how we understand Marr's three levels. In this essay, we refine Marr's distinctions using a population activity model of cortico-striatal circuitry exploring impulsivity and behavioral inhibition as a case study. In particular, we argue that the computational level should be redefined, for simply knowing the goal of a computation may not tell us much about why something has gone wrong and why the information-processing device is exhibiting abnormal behavior. How systems break down under impaired conditions should be part of the description at the computational level. We also agree with McClamrock (1991), Poggio (2012), and Stevens (2012) that feed-forward processing, cognition detached from behavioral context, and representations do not always drive cognitive processes in the way that Marr originally assumed, and we provide some additional reasons for why they are correct in their assessments.

2. The role of neuroscience in information processing theories, circa 1980

During his all-too-brief lifetime, David Courtnay Marr effectively established what is now known as computational neuroscience (Hardcastle, 2007). Frustrated with working on abstract theories of whole brain function, which he came to believe were too vague to ever be able to describe how the brain actually works, Marr concluded that we would need to delve into the details of specific brain mechanisms and the problems they are solving in order to fully explain human cognition: "Neural net theory, unless it is closely tied to the known anatomy and physiology of some part of the brain and makes some unexpected predictions, is of no value" (Marr, 1975, p. 876). At the same time, he believed that something like the computational descriptions of cognitive processes would provide the theoretical foundation for neurophysiology; neurophysiology alone was not enough to be explanatory. Marr wrote: "Neurophysiology and psychophysics have as their business to *describe* the behavior of cells or of subjects, not to *explain* such behavior" (1982, p. 15).

In pointing out the theoretical difference between algorithm and implementation, Marr also underscored an important distinction between description and explanation that philosophers of science today continue to rely on and debate (see, e.g., current controversies in the use of dynamical systems modeling in psychology; Chemero, 2009). Marr's point was that, despite the failings of then-current neural net theories, whatever it is that explains

thought and information processing is going to be something other than the influx and efflux of potassium across cell boundaries or some such, for that is simply the wrong type of description to account for cognition. Instead, we need more than just those sorts of accounts; we also need "a clear understanding of what is to be computed, how it is to be done, the physical assumptions on which the method is based, and some kind of analysis of algorithms that are capable of carrying it out" (Marr, 1982, p. 24). Explanation requires much more than just describing the interactions of component pieces. We need to know the problem the system under investigation faces and the form of the possible solutions it could enact to solve that problem. Good theories of the brain and cognition do not have to sacrifice mathematical rigor for neural data. Instead, computational neuroscience can and should combine both. Consequently, "gone is any explanation *in terms of* neurons—except as a way of implementing a method" (Marr, 1982, p. 18).

Marr divides a complete explanation into three parts. At the highest and most abstract level, we have what Marr calls the computational theory. This describes the goal of the computing system, why this goal is the appropriate one, and the basic logic of the strategy by which the computations are carried out. Historically, there has been some interpretive dispute over whether Marr intended the computational level to refer to the content of mental states, or whether it only specified a mathematical function being computed by a set of articulable rules. (See, e.g., Dennett, 1994; Harnish, 2002; Newell, 1982, 1986; Pylyshyn, 1984; Sterelny, 1990 vs. Butler, 1998; Egan, 1991, 1992, 1995; Shagrir, 2010.) Without defending our interpretation here (though see Kroustallis, 2006; Shagrir, 2010; Verdego & Quesada, 2011), we will be loosely siding with those who believe the computational level is representational, as we cannot conceive of a useful way to specify what a cognitive or behavioral goal is and why the goal is appropriate without at least obliquely referring to content.

Next, "there must exist an additional level of understanding at which the character of the information-processing tasks carried out ... are analyzed and understood in a way that is independent of the particular mechanisms and structures that implement them in our heads" (1982, p. 25). This is the level of representation and algorithm. At this level, we explain how the computational theory can be modeled. That is, we define how we represent the inputs and outputs to the system and how we transform the inputs into the outputs algorithmically. And finally, there is the level of implementation. At this level, we articulate how what Marr calls the "representations" are physically realized in the brain.

As Polger (2002) notes, there need not be a tight fit between the algorithm and the implementation. So long as the implementation produces the outputs the computational description requires, given the inputs it specifies, then the algorithmic level need only approximate how the brain actually transforms its representations. (See also Cummins, 1989, for a similar view, as well as Broadbent's dispute with McClelland and Rumelhart, 1985, over whether connectionist models are algorithmic or implementational.) Marr himself distanced the computational level from the algorithmic and implementational: "The computational theory of a process is rather independent of the algorithm or implementation levels, since it is determined solely by the information-processing task to be solved"

(1982, p. 337). In sum, there are three, very distinct, aspects to any explanation in computational neuroscience.

According to Marr, one advantage for distinguishing among these three levels is that we can explain the optimal functioning of the system: "It becomes possible, by separating explanations into different levels, to make explicit statements about what is being computed is optimal in some sense and why it is guaranteed to function correctly" (1982, p. 19). Brains are complicated and often faulty affairs. But if we can abstract away from the imperfect physicality and idealize brain function into something like mathematical algorithms, then we will be able to describe cognition in its perfect form.

It is at this point that we believe that Marr's views from 30-plus years ago diverge most strongly from what actually happens in laboratories and in theorizing today. In particular, instead of having models and theories emphasize optimal performance under ideal conditions, current theories and models emphasize impaired performance under degraded conditions. Our contention goes beyond the suggestion that Marr unfairly relied on the principles behind reverse engineering to develop his computational theories (see, e.g., the discussion in Dennett, 1994; Gilman, 1994, 1996; Kitcher, 1988). That discussion hinges on the fact that humans are evolved creatures, not designed ones. However, we are less interested in how Marr came to his theories than what his theories are of. That is, making optimizing design assumptions as a way into understanding a system might be a perfectly legitimate way to approach the problem. Or it might lead one astray. Either way, we focus instead on what Marr's theory is designed to explain—optimal performance under ideal conditions—and not the assumptions that led Marr to posit the specific goals and algorithms that he did. We claim that the target of Marr's theories is insufficient from today's perspective.

In addition, instead of isolated feed-forward models from Marr's time, current models emphasize context and feedback loops. (See also McClamrock, 1991; Poggio, 2012; Stevens, 2012, for additional discussion of these points.) And, instead of divorcing cognition and action, today our descriptions and explanations have cognition and action very much intertwined. (Indeed, the whole subfield of embodied cognition is devoted to exploring this idea.) Finally, instead of Marr's assumption that transformations over representations always drive the outputs, we know now that our computational world is much more varied and complex.

While we will offer each of the four changes listed above (adding system malfunction to computational descriptions; including context and feedback loops in algorithms; expanding cognition to include action; and giving a broader interpretation of algorithms than transforming representations) as refinements to Marr's original proposal, we would also like to stress that a multi-level approach to explanation and understanding in computational neuroscience is still very much alive and well today (see, e.g., Gurney, 2009). Marr's emphasis on asking why a brain process is occurring, instead of merely looking for a differential equation or some such to describe it, has been internalized by neurophysiologists and theoretical biologists alike.

3. Basal ganglia, action selection, and impulsivity

Instead of providing a comprehensive meta-analysis of the relevant literature, we are going to use a case-study approach to illustrate how we believe a Marrian approach has grown and matured over the past 30 years. Specifically, we are going to look at models of the basal ganglia in action selection and impulsivity.

Various brain systems, which function largely independently of one another, compete with each other for which area gets to drive motor behavior. Different information streams can and often do demand different behaviors of the same motor systems. All animals must solve this computational problem: how to sort through incoming data efficiently to figure out which action to perform with their limited motor resources. Our brains resolve these internal conflicts by selecting one action over the others. They do so by damping down the lower priority responses (see Gazzaniga et al., 2013; Gurney, Prescott, & Redgrave, 2001; Ridderinkhof, van den Wildenberg, Segalowitz, & Carter, 2004; Zhang, Hughes, & Rowe, 2012, for discussion). Exactly how our brains accomplish this feat through a variety of inhibitory and excitatory connections in our basal ganglia is described in more detail below.

One way in which action selection goes awry is by failing to inhibit responses as it should. Otherwise known as impulsivity, this failure is associated with several very common psychological disorders, including attention-deficit hyperactivity disorder, addiction, and perhaps even chronic pain (Egli, Koob, & Edwards, 2012; Hardcastle, 2014; Nichols & Waschbusch, 2004). We take the description of impulsivity, as well as the associated psychological disorders, to reside at the Marrian computational level, for it articulates a common way the selecting of actions—the goal of the system in question—goes awry. In other words, to be impulsive is to fail to perform the process of action selection as described at the computational level.

We can assess impulsivity by using a stop signal reaction time (SSRT) task (Alderson, Rapport, & Kofler, 2007; Band & van Boxtel, 1999). Being able to assess impulsivity is important because it turns out that the severity of this type of impairment can predict the response of individuals with related disorders to pharmacological interventions (van der Oord, Geurts, Prins, Emelkamp, & Oosterlaan, 2012). In SSRT tasks, subjects are cued to make relatively simple, well-trained responses. However, on a specific proportion of trials, they are presented with a stop signal, which tells the subject that the well-trained response should be withheld. We then measure how often subjects fail to inhibit their responses as a function of how much time they take between signal and action. The beauty of this simple experimental paradigm is that it can be used equally well in both humans and other animals, including rats, with high translational accuracy (Eagle et al., 2008; Winstanley, Eagle, & Robbins, 2006).

Briefly, for rats, an SSRT trial initiates when the rats poke their nose into a water port, which triggers a lever on their left to extend. Pressing the left lever extends a right lever. In most of the trials, the "Go" trials, pushing the right lever is the correct response; in this case, pressing the right lever results in reward access to the water port. However, in about a fifth of the trials (the "Stop" trials), the right lever extension is paired with a stop

signal; in this protocol, a light turns on. In the "Stop" trials, refraining from pressing the right lever within a specified duration is the correct response and is also rewarded with access to the water port (see Fig. 1).

The basal ganglia are a group of subcortical nuclei that most neuroscientists believe play a central role in prioritizing the urgency or "salience" of requests for action or movement by various parts of the brain. Here is where the Marrian computation is implemented. As a central switch, the basal ganglia determine which actions will be performed at the expense of which others. There is a broad range of algorithmic and implementational models for the basal ganglia switching function, ranging from conductance-based models of individual neurons to models of the spiking function of clusters of neurons to system-level models of neural circuitry to model robots that exhibit the appropriate behavior. In this essay, we focus on a population activity model, a model in the middle of the range of possibilities.

A variety of experimental data not only demonstrate that specific basal ganglia structures contribute to SSRT performance but also illustrate how neural structures underlie or implement the failure of behavioral inhibition (Fig. 2). For example, excitotoxic lesions (lesions induced by excessive neurotransmitter stimulation) of the medial striatum increase SSRT (Eagle & Robbins, 2003a), as do injection of dopamine receptor D2 antag-

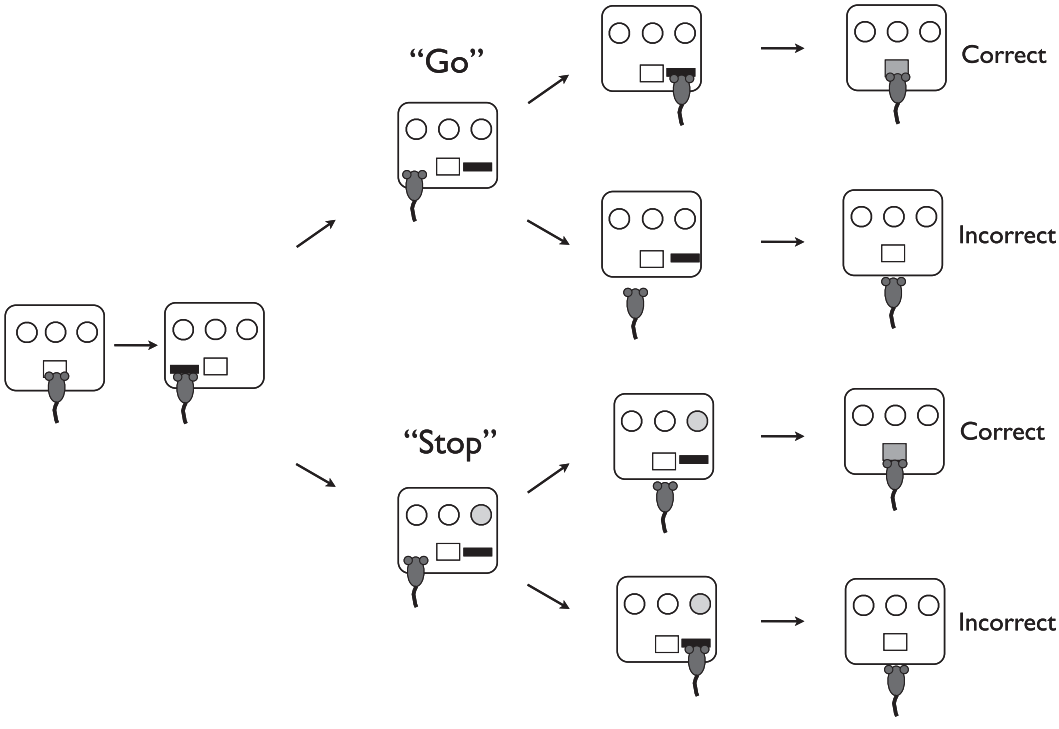

Fig. 1. Diagram of the stop signal reaction time task for rats. Adapted from Hardcastle et al. (2013).

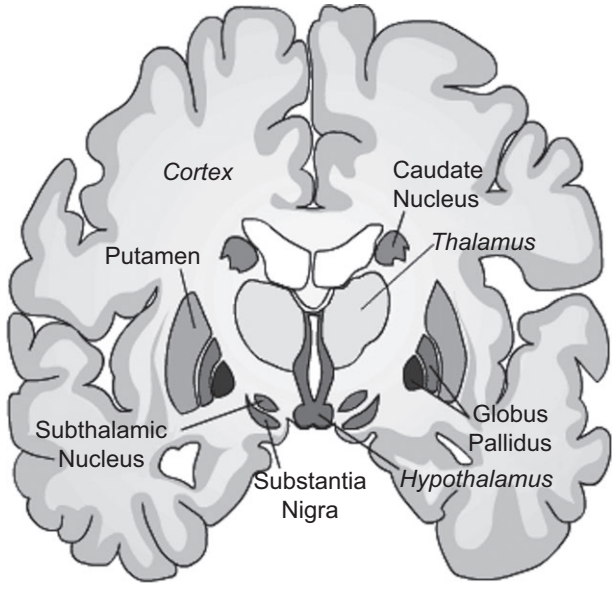

Fig. 2. The basic structure of the basal ganglia, including the caudate nucleus, globus pallidus, hypothalamus, putamen, substantia nigra, and subthalamic nuclei. Adapted from http://www.dana.org/news/brainwork/detail.aspx?id=6028, as accessed in May 2012.

onists into the dorsomedial striatum. Lesions of the subthalamic nucleus (STN) decrease mean reaction time on "Go" trials and decrease accuracy on "Stop" trials (Eagle & Robbins, 2003b). All of these interventions are implicated in impulsive behavior. In contrast, injection of dopamine receptor D1 antagonists into the same region decreases SSRT (Eagle et al., 2011), which increases behavioral inhibition, which might then decrease impulsive action selection. These sorts of findings highlight that several regions within the basal ganglia contribute in different ways to behavioral inhibition, both enhanced and impaired, at least as measured by the SSRT task.

We can use data such as these to model particular basal ganglia structures critical to this task. The study in question examines whether a Wilson–Cowan type population activity model of cortico-basal ganglia processing can predict performance in an SSRT task. A Wilson–Cowan type model is a system of nonlinear differential equations that represent the activity of multiple populations of interacting neurons (Wilson & Cowan, 1972). (Horgan & Tienson, 1994, 1996, give us one way of understanding dynamical systems approaches from a Marrian perspective.) The study asks whether a particular system of nonlinear differential equations comprises the algorithm that allows the basal ganglia to implement a brain's computational goal of selecting the most salient action. The mathematical details describing this particular model are available in Hardcastle, Smith, and Burke (2013) and are summarized in the Appendix but Fig. 3 captures the basic architecture of the simplified basal ganglia network described by the equations—it illustrates the basic implementation as suggested by the algorithmic model. It is important to note that

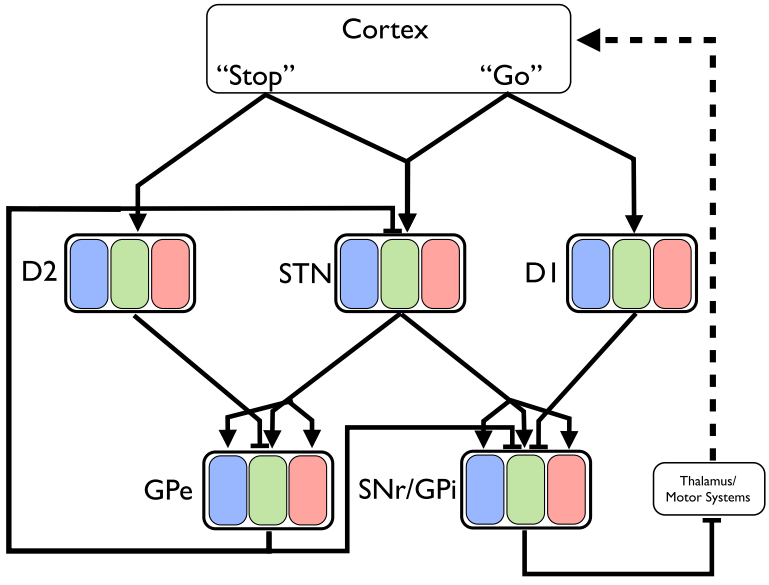

Fig. 3. Model of the basal ganglia. Cortical input is either "Go," which promotes action selection, or "Stop," which promotes action suppression. The cortico-basal ganglia network contains three separate channels that each corresponds to a unique action, illustrated by different colors here; for clarity, we have only shown the connectivity among a single channel as the black lines. The pointed arrow represents excitation, while the bar represents inhibition. Adapted from Hardcastle et al. (2013).

all parameters in this model (i.e., the spontaneous and maximum firing rates, mean cortical input, connectivity strength, time constants) were found either in the literature (Bevan & Wilson, 1999; Humphries & Gurney, 2002; Humphries, Stewart, & Gurney, 2006; Mink, 1996) or selected through various parameter studies.

Excluding the cortex, this network contains five populations, each containing three subpopulations, or channels. Each subpopulation corresponds to a particular action: pressing the left lever press, pressing the right lever, or doing something else entirely. The basal ganglia maintain an organized spatial topography that is largely preserved across its subpopulations such that connections between the subpopulations effectively become microcircuits (Redgrave et al., 2010).

In the "Go" experimental condition, cortical input into the channel that corresponds to pressing the left lever turns on. This excitatory input increases the firing rate of that channel's striatum D1 neurons, which then directly inhibit the substantia nigra pars reticulata (SNr) and globus pallidus internal segment (GPi; via channel-to-channel projections), and the channel's STN neurons, which then diffusely excite the SNr/GPi (via channel-to-population projections). The projections from striatum D1 and STN cause an off-center, on-surround pattern of activity in the SNr/GPi (i.e., the SNr/GPi activity in a channel is suppressed while the neighboring channels are activated), which highlights the channel under direct inhibition from striatum D1. The SNr/GPi maintains channel-specific

inhibitory projections to the thalamus, which has excitatory projections back to the cortex, thus completing a cortico-basal ganglia-thalamic circuit.

Suppressing the SNr/GPi disinhibits the thalamus, which in turn excites various areas of the cortex into promoting an action. Consequently, sufficient inhibition of an output subpopulation ultimately leads to action selection; in this model, continuous integration of the appropriate SNr/GPi-channel firing rate under a level relative to a selection threshold means that the model opts to press the left lever. Failure to reach this threshold signifies a failure to complete a trial. Because activation of striatum D1 can ultimately suppress activity in SNr/GPi, thereby selecting an action, the D1-STN pathway is the "Go" pathway. Once the model selects the left lever press action, cortical input for that action turns off because the salience of that particular action quickly decreases.

The next action in the "Go" trial—the right lever press—then becomes the most salient action, and cortical activation turns on for that channel and consequently inhibits the appropriate subpopulation of the SNr/GPi. Again, continuous integration of the channel's firing rate under threshold determines when, or if, that action is selected. Selection of the right lever press action concludes a successful "Go" trial.

In the "Stop" experimental condition, the action selection process for pressing the left lever is the same as in the "Go" condition. Again, selecting the left lever press action immediately increases cortical input into the channel corresponding to a right lever press. However, now a stop signal turns on, which increases cortical input to striatum D2. Similar to striatum D1, striatum D2 has an inhibitory and channel-specific projection to the globus pallidus external segment (GPe). In turn, the GPe directly inhibits the SNr/GPi as well as the STN. The STN maintains diffuse excitatory projections back to the GPe. Thus, when cortical input to a particular channel in striatum D2 increases, it inhibits the corresponding channel in GPe, which then relieves inhibition of the corresponding channel in SNr/GPi. This allows the output population activity to remain above threshold, which in turn inhibits the thalamus and the corresponding cortical areas. Consequently, the D2-GPe-SNr/GPi pathway is the "Stop" pathway. The Stop and Go pathways converge at the SNr/GPi, which functions as the output decision-maker for the network (cf., Humphries et al., 2006). We can see that the hyper-direct cortical-STN-SNr/GPi pathway is what prevents premature action selection (Frank, 2005; Graybiel, 2005).

This model reproduces a basic phenomenon in human and rat SSRT task behavior, namely that "Stop" trial accuracy declines as the time between the initiation of the pre-potent response and the stop signal increases. Thus, the results of this model are consistent with those of published empirical work (e.g., Eagle & Robbins, 2003b). Moreover, the effects of dopamine-1 receptor (D1R) and dopamine-2 receptor (D2R) blockades on SSRT performance in this model are very similar to those reported elsewhere (Eagle et al., 2011). Finally, this model approximates the response latency probability distribution for a group of rats trained in the stop signal reaction time task. Hence, this model appears to capture the population-level dynamics involved in action selection.

4. Updating Marrian theoretical approaches

What is important about the above model for our purposes is that, while it roughly approximates Marr's three levels of understanding, there are four important distinctions. First, not only does it account for how the dynamics among several areas change when they are engaged in an SSRT task, but it also accounts for both normal behavior under ideal conditions as well as impaired performance data. Indeed, the fact that it meets both these desiderata helps support the notion that the model is empirically adequate. This is a model of impaired cognition in action selection as much as it is one of optimal performance.

In particular, as mentioned above, we know the effects of D1R and D2R antagonists on the SSRT task (Eagle et al., 2011). This model can simulate the effects of a D1R blockade by decreasing the strength of the inputs to the D1-expressing striatum population, which corresponds to decreasing the default cortical input values. It turns out that, similar to the empirical data, after administering a D1R antagonist, decreasing inputs to D1Rs also decreased SSRT. In addition, when the input to D2Rs is decreased, which mimics increasing levels of a D2R antagonist, SSRT increases.

This comparison of the model to data of impaired processing supports the arguments of the authors that the model they have developed is valid, for it encompasses more data than a model that simply accounts for how the region might function only under optimal circumstances. That is, because the model explains a variety of types of data, it is considered more biologically plausible than a model that does not. We submit that these sorts of considerations did not enter into Marr's projects, if for no other reason than much of these sorts of data did not exist back then. However, we also believe that being able to model impaired performance data fits with Marr's larger philosophical commitment to connect algorithms to computational goals and neurophysiological data as tightly as possible: "the algorithm depends heavily on the computational theory . . . [and] it also depends on the characteristics of the hardware in which it is to be implemented" (1982, p. 337).

Being able to model impaired performance as such requires that the goals of the computation specify what counts as a dysfunction. There are at least three possible ways of describing how a system is functioning. It can be doing exactly what it was designed to do; it can be doing what it was designed to do, but doing it suboptimally; or it can be doing something other than what it was designed to do. To figure out which descriptor is appropriate for any given system requires that we know the function of the system—the goal of its processors—and what counts as a malfunction. This requires that we broaden Marr's computational level to include a description of what types of suboptimal performance count as a breakdown in achieving the system's goals. In order to identify the SSRT model above as one that exhibits the disorder of impulsivity, we have to know that being impulsive is not the same thing as making appropriate decisions suboptimally. As an aside, we note that separating suboptimal but proper functioning from genuine malfunctions or dysfunctions has an air of arbitrariness about it. There are many pragmatic considerations that go into drawing these lines. Nevertheless, this is what the discipline

of psychiatry is about at its core, figuring out when behaviors and mental states are truly disordered as opposed to normal but compromised.

Second, the model described above takes into account known biological feedback loops. Its responses occur in distinct phases that are associated with the different stages of the simulated SSRT task ("left lever press," "right lever press," etc.). While cortical drive propagates in a feed-forward manner throughout the model basal ganglia populations during each phase, the model's ultimate responses depend on feedback as well as feed-forward information because the simulated SSRT task progresses to each subsequent phase based on selected actions (i.e., suppression of SNr/GPi output), which change the configuration of the cortical drive in a "closed loop." The feedback loops add a measure of biological plausibility—and of complexity—that Marr's models lacked (due largely to the immature modeling techniques of the day; cf., Stevens, 2012).

Third, and part and parcel of how feedback drives the output of the model, cognition and action are significantly intertwined in this model and in our brains. Which actions are selected is fed back into the model, which then influences how the model decides which action is to be selected next. Decision-making, a cognitive activity, depends upon the immediate behavioral context as well as the priority assigned to various actions via the cortex. This process, too, is different from what Marr's models assume, in that they separated the cognitive outputs from any sort of behavioral feedback (see also McClamrock, 1991, for a different approach to a similar conclusion).

Finally, it is important to notice how the model does and does not use or account for representations or content. Marr identified mental processes with how internal representations are "obtained and how they interact" (1982, p. 6). In contrast, in this model, the cognitive process of decision-making does not rely on representations interacting with each other. While it is clear that cortical inputs and behavioral outputs of the system are meant to correspond to or at least index perceptual and motor representations in the standard sense, it would be incorrect to claim that the algorithmic transformations that occur are transformations over representations. In fact, the representations corresponding to pressing the left lever, pressing the right lever, and so forth, remain unchanged and do not interact with each other at all as the model processes information. What changes are the activation levels for each of the representations, and that ultimately drives which representation emerges the victor. This model explains what might happen to a mental representation after it has been built. It turns out that deciding which action to take does not require that the representation for the "winning" action be the one that is constructed in the end. Instead, the "winner" is the one whose activation is the least suppressed.

This is a very different approach to mental representation and cognition than what Marr articulates in his discussion of his tripartite scheme, for it focuses on the dynamic interactions among several areas as they engage in a task, not on transforming a representation via some algorithm. Marr explains that second level of analysis involves two things: "(1) a *representation* for the input and for the output of the process and (2) an *algorithm* by which the transformation may actually be accomplished" (1982, p. 23, italics his). This model does not transform a representation; instead, it chooses which representation to enact.

And yet, even though we do find differences between what Marr did 30 years ago and what goes on in computational neuroscience today, we can still put the current model into an approximation of Marr's three levels of understanding. That is, at the computational level, in which we need to articulate the goal of the computation, why that goal is appropriate, and what the logic of the strategy is by which it is carried out, we can see that the above is a model of action selection that must choose among competing demands to select the most salient behavior. We would just add that we should also outline how the computation fails, and, in this case, it malfunctions by exhibiting impulsivity.

Similarly, at the representation and algorithm level, we can answer how the computational theory is implemented, how the input and output are represented, and how the algorithm transforms the data. In this model of the basal ganglia, we find a series of coupled differential equations, with variables for inputs and outputs, but integration at the output neurons determining which action is selected. And, finally, at the level of hardware implementation, where Marr asks how the representation and algorithm is physically realized, we can point to the neural subareas of the basal ganglia.

However, neurophysiology—far from being a mere implementer of algorithms and computations—influences, constrains, and shapes both the computational theory and the hypothesized representations and algorithms (see also Poggio, 2012; Stevens, 2012). Instead of holding that the business of neurophysiology is only to describe the behavior of cells, and not to explain it, and instead of holding that explanations in terms of neurons should be replaced by computational goals and modeling algorithms, we believe that in order to have a clear understanding of what, why, and how, something in the brain is being computed, "neurons implementing a method" are fundamental to the explanation itself. Only by studying the brain itself do we understand the deep and primary importance of feedback loops, context, activity suppression, dynamic nonlinear interactions, and the like, to the strategies our brains use in solving problems and making decisions. This information then informs the proposed algorithms and hypothesized computations. We must work our theories from the bottom-up as well as the top-down.[2]

Notes

1. Though supporting this claim would require a separate article, it is our opinion that this change in focus has been largely driven by the funding priorities of the National Institute of Health, at least in the United States.
2. We would like to acknowledge the support and contributions of Gregory D. Smith and Joshua A. Burk to the modeling project, as well as the generous financial support of the Howard Hughes Medical Institute and the College of William and Mary. We also owe a debt of thanks to two anonymous reviewers and Serge Thill for their detailed comments on earlier drafts. Our article is much stronger as a result.

References

Alderson, R. M., Rapport, M. D., & Kofler, M. J. (2007). Attention-deficit/hyperactivity disorder and behavioral inhibition: A meta-analytic review of the stop-signal paradigm. *Journal of Abnormal Child Psychology*, *35*, 745–758.

Band, G. P., & van Boxtel, G. J. (1999). Inhibitory motor control in stop paradigms: Review and reinterpretation of neural mechanisms. *Acta Psychologica (Amsterdam)*, *101*, 179–211.

Bevan, M. D., & Wilson, C. J. (1999). Mechanisms underlying spontaneous oscillation and rhythmic firing in rat subthalamic neurons. *Journal of Neuroscience*, *19*, 7617–7628.

Broadbent, D. (1985). A question of levels: Comment on McClelland and Rumelhart. *Journal of Experimental Psychology: General*, *114*, 189–192.

Butler, K. (1998). Content, computation, and individuation. *Synthese*, *114*, 277–292.

Chemero, A. (2009). *Radical embodied cognition*. Cambridge, MA: The MIT Press.

Cummins, R. (1989). *Meaning and mental representation*. Cambridge, MA: MIT Press.

Dennett, D. (1994). Cognitive science as reverse engineering: Several meanings of "top-down" and "bottom-up". In D. Prawitz, B. Skyrms, & D. Westershal (Eds.), *Logic, methodology and philosophy of science IX* (pp. 679–689). Amsterdam: Elsevier Science.

Eagle, D., Baunez, C., Hutcheson, D., Lehmann, O., Shah, A., & Robbins, T. (2008). Stop-signal reaction-time task performance: Role of prefrontal cortex and subthalamic nucleus. *Cerebral Cortex*, *18*, 178–188.

Eagle, D. M., & Robbins, T. W. (2003a). Inhibitory control in rats performing a stop-signal reaction time task: Effects of lesions of the medial striatum and d-amphetamine. *Behavioral Neuroscience*, *117*, 1302–1317.

Eagle, D. M., & Robbins, T. W. (2003b). Lesions of the medial prefrontal cortex or nucleus accumbens core do not impair inhibitory control in rats performing a stop-signal reaction time task. *Behavioral Brain Research*, *146*, 131–144.

Eagle, D. M., Wong, J. C. K., Allan, M. E., Mar, A. C., Theobald, D. E., & Robbins, T. W. (2011). Contrasting roles for the dopamine D1 and D2 receptor subtypes in the dorsomedial striatum but not the nucleus accumbens core during behavioral inhibition in the stop-signal task in rats. *Journal of Neuroscience*, *31*, 7349–7356.

Egan, F. (1991). Must psychology be individualistic? *Philosophical Review*, *100*, 179–203.

Egan, F. (1992). Individualism, computation and perceptual content. *Mind*, *101*, 443–459.

Egan, F. (1995). Computation and content. *Philosophical Review*, *104*, 181–203.

Egli, M., Koob, G. F., & Edwards, S. (2012). Alcohol dependence as a chronic pain disorder. *Neuroscience and Biobehavioral Reviews*, *36*, 2179–2192.

Ermentrout, B. (1998). Neural networks as a spatio-temporal pattern-forming systems. *Reports on Progress in Physics.*, *61*, 353–430.

Frank, M. J. (2005). Dynamic dopamine modulation in the basal ganglia: A neurocomputational account of cognitive deficits in medicated and nonmedicated Parkinsonism. *Journal of Cognitive Neuroscience*, *17*, 51–72.

Gazzaniga, M., Ivry, R. B., & Mangun, G. R. (2013). *Cognitive neuroscience: The biology of the mind*. New York: Norton.

Gilman, D. (1994). Simplicity, cognition, and adaptation: Some remarks on Marr's theory of vision. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, *1994*(1), 454–464.

Gilman, D. (1996). Optimization and simplicity: Computational vision and biological explanation. *Synthese*, *107*, 292–323.

Graybiel, A. M. (2005). The basal ganglia: Learning new tricks and loving it. *Current Opinions in Neurobiology*, *15*, 638–644.

Gurney, K. N. (2009). Computational models in neuroscience from membranes to robots. In E. Mavritsaki & D. Heinke (Eds.), *Computational modeling in behavioral neuroscience: Closing the gap between neurophysiology and behavior* (pp. 107–136). New York: Psychology Press.

Gurney, K. N., Prescott, T. J., & Redgrave, P. (2001). A computational model of action selection in the basal ganglia: A new functional anatomy. *Biological Cybernetics*, *84*, 401–410.

Hardcastle, V. G. (2007). David Courtnay Marr. In N. Koertge (Ed.), *New dictionary of scientific biography* (Vol. 5, pp. 32–34). New York: Charles Scribner's Sons.

Hardcastle, V. G. (2014). Pleasure gone awry? A new conceptualization of chronic pain and addiction. *Review of Philosophy and Psychology*, *5*, 71–85.

Hardcastle, K., Smith, G. D., & Burke, J. A. (2013). A population activity model of cortico-striatal circuitry underlying behavioral inhibition in rats. In C. R. Gordon & T. G. Abbadelli (Eds.), *Globus pallidus: Regional anatomy, functions/dysfunctions and role in behavioral disorders* (pp. 67–92). New York: Neuroscience Research Progress, Nova Science Publishers.

Harnish, R. M. (2002). *Minds, brains, computers: An historical introduction to the foundations of cognitive science*. Malden, MA: Blackwell.

Horgan, T., & Tienson, J. (1994). A nonclassical framework for cognitive science. *Synthese*, *101*, 305–345.

Horgan, T., & Tienson, J. (1996). *Connectionism and the philosophy of psychology*. Cambridge, MA: The MIT Press.

Humphries, M. D., & Gurney, K. N. (2002). The role of intra-thalamic and thalamocortical circuits in action selection. *Network*, *13*, 131–156.

Humphries, M. D., Stewart, R. D., & Gurney, K. N. (2006). A physiologically plausible model of action selection and oscillatory activity in the basal ganglia. *Journal of Neuroscience*, *26*, 12921–12942.

Kitcher, P. (1988). Marr's computational theory of vision. *Philosophy of Science*, *55*, 1–24.

Kroustallis, B. (2006). Content individualtion in Marr's theory of vision. *Journal of Mind and Behavior*, *27*, 57–72.

Marr, D. C. (1975). Approaches to biological information processing. *Science*, *190*, 875–876.

Marr, D. C. (1982). *Vision: A computational investigation into human representation and information processing*. New York: W.H. Freeman.

McClamrock, R. (1991). Marr's three levels: A re-evaluation. *Minds and Machines*, *1*, 185–196.

McClelland, J., & Rumelhart, D. (1985). Levels indeed! A response to Broadbent. *Journal of Experimental Psychology: General*, *114*, 189–192.

Mink, J. W. (1996). The basal ganglia: Focused selection and inhibition of competing motor programs. *Progress in Neurobiology*, *50*, 381–425.

Newell, A. (1982). The knowledge level. *Artificial Intelligence*, *18*, 87–127.

Newell, A. (1986). The symbol level and the knowledge level. *In Meaning and cognitive structure*, edited by Z.W. Pylyshyn and W. Demopoulos, pp. 31–39. Norwood, MA: Ablex.

Nichols, S. L., & Waschbusch, D. A. (2004). A review of the validity of laboratory cognitive tasks used to assess symptoms of ADHD. *Child Psychiatry and Human Development*, *34*, 297–315.

van der Oord, S., Geurts, H. M., Prins, P. J., Emelkamp, P. M., & Oosterlaan, J. (2012). Prepotent response inhibition predicts treatment outcome in attention deficit/hyperactivity disorder. *Child Neuropsychology*, *18*, 50–61.

Pinto, D. J., Brumbery, J. C., Simons, D. J., & Ermentrout, G. B. (1996). A quantitative population model of whisker barrels: Re-examining the Wilson–Cowan equations. *Journal of Computational Neuroscience*, *3*, 247–264.

Poggio, T. (2012). The levels of understanding framework, revised. *Perception*, *41*, 1017–1023.

Polger, T. (2002). Neural machinery and realization. *Philosophy of Science*, *71*, 997–1006.

Pylyshyn, Z. W. (1984). *Computation and cognition*. Cambridge, MA: The MIT Press.

Redgrave, P., Rodriguez, M., Smith, Y., Rodriguez-Oroz, M. C., Lehericy, S., Bergman, H., Agid, Y., DeLong, M. R., & Obeso, J. A. (2010). Goal-directed and habitual control in the basal ganglia: Implications for Parkinson's disease. *Nature Reviews Neuroscience*, *11*, 760–772.

Ridderinkhof, R., van den Wildenberg, W. P. M., Segalowitz, S. J., & Carter, C. S. (2004). Neurocognitive mechanisms of cognitive control: The role of prefrontal cortex in actions selection, response inhibition, performance monitoring, and reward-based learning. *Brain and Cognition*, *56*, 129–140.

Shagrir, O. (2010). Marr on computational-level theories. *Philosophy of Science*, *77*, 477–500.
Sterelny, K. (1990). *The representational theory of mind: An introduction*. Cambridge, MA: Blackwell.
Stevens, K. A. (2012). The vision of David Marr. *Perception*, *41*, 1061–1072.
Verdego, V. M., & Quesada, D. (2011). Level of explanation vindicated. *Review of Philosophy and Psychology*, *2*, 77–88.
Wilson, H. R., & Cowan, J. D. (1972). Excitatory and inhibitory interactions in localized populations of model neurons. *Biophysics Journal*, *12*, 1–24.
Winstanley, C. A., Eagle, D. M., & Robbins, T. W. (2006). Behavioral models of impulsivity in relation to ADHD: Translation between clinical and preclinical studies. *Clinical Psychological Reviews*, *26*, 379–395.
Zhang, J., Hughes, L. E., & Rowe, J. B. (2012). Selection and inhibition mechanisms for human voluntary action decisions. *NeuroImage*, *63*, 392–402.

Appendix

The above model of the basal ganglia's subpopulations (or channels) employs the following system of Wilson–Cowan differential equations (Ermentrout, 1998; Pinto, Brumbery, Simons, & Ermentrout, 1996):

$$\tau \frac{da_i}{dt} + a_i = f_i\left(h_i + \sum_j w_{ij} a_j\right)$$

where the dependent variable a_i is the activity of subpopulation i, τ is a decay time constant, w_{ij} is the strength of connection from subpopulation j to i, and h_i represents cortical input into subpopulation i.

The firing rate of subpopulation i is given by

$$f_i(s_i) = \frac{f_i^{max}}{1 + \exp\left(\frac{-s_i - \sigma_i^0}{\sigma_i^1}\right)}$$

where $s_i = h_i + \sum_j w_{ij} a_j$ represents the input to subpopulation i. The maximum firing rate f_i^{max} and location and scale parameters σ_i^0 and σ_i^1 vary with subpopulation.

Additionally, the cortical input is modeled as a stochastically driven differential equation of the following form:

$$\frac{dh_i}{dt} = \frac{-(h_i - h_i')}{\tau_h} + \xi_i(t)$$

where τ_h is the decay time constant and h_i' is the mean cortical input. The time-dependent variable $\xi_i(t)$ represents Gaussian white noise terms, which correspond to noise that arises from a subpopulation containing a finite number of neurons, with $< \xi_i(t) > = 0$, and a two-time covariance given by $< \xi_i(t)\xi_j(t') > = \gamma_{ij}\delta(t - t')$, where $\gamma_{ij} = 0$ for $i \neq j$, and δ is the dirac-delta function.