TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 7 (2015) 243–258

ISSN:1756-8757 print / 1756-8765 online
DOI: 10.1111/tops.12132

Beyond Single-Level Accounts: The Role of Cognitive Architectures in Cognitive Scientific Explanation

Richard P. Cooper,[a] David Peebles[b]

[a]*Centre for Computation, Cognition and Modelling, Department of Psychological Sciences, Birkbeck, University of London*
[b]*Department of Behavioural and Social Sciences, University of Huddersfield*

Received 12 September 2013; received in revised form 15 July 2014; accepted 15 July 2014

Abstract

We consider approaches to explanation within the cognitive sciences that begin with Marr's computational level (e.g., purely Bayesian accounts of cognitive phenomena) or Marr's implementational level (e.g., reductionist accounts of cognitive phenomena based only on neural-level evidence) and argue that each is subject to fundamental limitations which impair their ability to provide adequate explanations of cognitive phenomena. For this reason, it is argued, explanation cannot proceed at either level without tight coupling to the algorithmic and representation level. Even at this level, however, we argue that additional constraints relating to the decomposition of the cognitive system into a set of interacting subfunctions (i.e., a cognitive architecture) are required. Integrated cognitive architectures that permit abstract specification of the functions of components and that make contact with the neural level provide a powerful bridge for linking the algorithmic and representational level to both the computational level and the implementational level.

Keywords: Cognitive architecture; Explanation; Algorithmic and representational level; Bayes' rule; Cognitive neuroscience

Explanation within cognitive science has frequently been argued to require multiple domains or levels, and several distinct multi-level accounts of cognitive scientific explanation have been proposed (e.g., Chomsky, 1965; Cummins, 1983; Marr, 1982; Newell, 1982). The account of Marr, however, has received perhaps the most attention. A central tenet of Marr's analysis is that a complete explanation of a device's behavior requires an account of that behavior at what he terms the computational level (CL), the algorithmic and representational level (ARL), and the implementation level (IL). While Marr argued

Correspondence should be sent to Richard P. Cooper, Centre for Cognition, Computation and Modelling, Department of Psychological Sciences, Birkbeck, University of London, London WC1E 7HX, UK. E-mail: r.cooper@bbk.ac.uk

that an account at each of the three levels was required for a complete explanation, he also emphasized the primacy of the most abstract of his levels, the CL. Specifically, he argued that while "algorithms and mechanisms are empirically more accessible, [...] the level of computational theory [...] is critically important from an information-processing point of view [... because ...] the nature of the computations that underlie perception [and, by extension, cognition] depends more upon the computational problems that have to be solved than upon the particular hardware in which their solutions are implemented" (Marr, 1982, p. 27). This position has face validity, particularly with respect to naturally intelligent systems where the particular hardware in which the solutions are implemented is shaped by evolutionary factors, and so will arguably be well-suited, or perhaps even optimized, for the computational problems that an evolved agent such as ourselves must solve (though see Jacob, 1977).

Marr was critical of approaches to the understanding of cognitive systems that were not rooted in the CL. For example, in discussing the work of Newell and Simon (1972) on the application of production systems to problem-solving, he wrote that "mechanism-based approaches are genuinely dangerous. The problem is that the goal of such studies is mimicry rather than understanding, and these studies can easily degenerate into the writing of programs that do no more than mimic in an unenlightening way" (Marr, 1982, p. 347).[1]

But is the primacy afforded by Marr to the CL justified? One could make plausible arguments for the primacy of Marr's two other levels. Surely, for example, the IL provides a privileged starting point because we can directly study the operation and connectivity of neurons in different brain structures and regions—and in different species—and knowing how neurons function and interact can give us insights into the operation of small-scale neural networks, which in turn can allow us to understand the capabilities of larger scale networks and structures. We will argue below that this purely reductionist strategy is likely to be of limited utility—in part because of the difficulties of "reverse inference" (inferring function from neural-level activation). Similarly, one might argue that a privileged role should be accorded to the ARL, given the role that information-processing models continue to play in the development of cognitive theory. This is the argument that we develop, though we do so in the context of cognitive architectures—theories of the organization and interaction of the hypothesized subsystems subserving cognition.

1. The role of the computational level

Part of Marr's justification for CL explanation was a reaction against early AI computer programs that did not take seriously "the distinction between what a program did and how it did it" (Marr, 1982, p. 28), thereby obscuring a mechanism's purpose, but Marr was equally clear that a CL explanation should address two questions: "what a device does and why" (p. 22). The "what" question might be answered by a specific mathematical function (i.e., by set-theoretic descriptions of the inputs and outputs and a mapping from inputs to outputs), but to answer "why" questions requires additional machinery—typically appealing to over-arching background theories such as (in the case of the cognitive sciences) evolution. "Why" questions are undoubtedly critical for explanation, but, at least with respect to explanations of cognitive functioning, their potential answers are generally more speculative and far more difficult to address in concrete, unequivocal terms than "what" questions (see Anderson, 2015). A pragmatic approach is therefore to de-conflate "what" and "why" and approach them (where possible) as independent questions.

Answering "what" questions requires a suitable language in which to phrase the answers. Over the last 15 years, CL theorizing within the cognitive sciences has come to be dominated by approaches in which behavior is understood in probabilistic terms, with appeal to Bayes' theorem in order to determine the posterior probability of an event, given evidence and prior probabilities (e.g., Griffiths, Kemp, & Tenenbaum, 2008). Bayesian approaches have been particularly successful in providing putative accounts of behavior that may be broadly characterized as inductive, such as categorization (Tenenbaum & Griffiths, 2001) and reasoning (Oaksford & Chater, 2001), but it is important to recognize that there are alternative frameworks within which CL explanations might be developed. Marr (1982) used differential equations, but other feasible CL languages include formal logics (such as first-order predicate logic, various modal logics, defeasible logics—see Baggio, van Lambalgen & Hagoort, 2015—attribute-value logic, and combinatory logic), information theory (Shannon & Weaver, 1949), category theory (Phillips & Wilson, 2010), graph theory (Baronchelli, Ferrer-i-Cancho, Pastor-Satorras, Chater, & Christiansen, 2013), and even the problem-space theory of Newell and Simon (1972). From the perspective of Marr's CL, Bayesian accounts do not have any a priori superiority over accounts expressed in any of these other frameworks.[2]

The critical issue for cognitive science methodology is not so much the specific formalism in which CL theories should be stated, but what role (if any) CL statements should play in explaining cognition or developing cognitive theory. For Marr, the role of the CL was to formulate the computational problem under consideration precisely such that one might then consider how that problem might be addressed at the ARL. Critically, Marr was concerned not with the abstract computational problem faced by an agent (which, with respect to vision, might be expressed in Bayesian terms as being to determine the most likely interpretation of the current visual input), but with the computational problems faced by subcomponents (such as the retina or the primary visual cortex) of that agent. Thus, Marr characterized the function of the retina as computing ∇^2 and its time derivative of the retinal image (cf. Marr, 1982, p. 337). From this perspective, the primary role of the CL is to constrain the ARL by specifying the overall function that must be computed at that level.

2. Limitations of purely computational-level explanations

A critical feature of CL explanation is that it abstracts away from details of processing and implementation. To illustrate, consider the motivation given by Xu and Tenenbaum

(2007) in developing their Bayesian account of word learning. The authors' aim is "to understand in functional terms how implicit knowledge and inferential machinery guide people in generalising from examples—rather than to describe precisely the psychological processes involved" (p. 251). In other words, their aim is to present an account of their domain of interest (learning the meanings of words) that captures behavioral regularities while abstracting away from specific algorithmic and representational commitments. Similar appeals to the utility of CL explanation are common in the Bayesian literature (see, e.g., Norris, 2006). The emphasis is generally on how the CL account may serve as a highly abstract statement of theory that deliberately avoids lower level commitments. This appeal to abstraction is also apparent in Marr's original account of CL explanation. Specifically, his attack on the insufficiency of early AI programs as explanations included the complaint that "particular data structures, such as lists of attribute value pairs called property lists in the LISP programming language, were held to amount to theories of the representation of knowledge" (Marr, 1982, p. 28). By abstracting away from specific representational devices (such as LISP data structures), CL accounts side-step representational debates (including, e.g., that underlying the connectionist/symbolic divide).

In abstracting away from mechanism and process, CL explanations are generally held to be idealizations. Thus, processing constraints (e.g., working memory limitations) may mean that the actual function computed by a specific algorithm or implementation is only an approximation of the CL function. Again, Marr (1982) makes this point, citing Chomsky's distinction between linguistic competence and linguistic performance within transformational grammar (Chomsky, 1965) as corresponding to the difference between a CL explanation and an ARL one, with performance factors modulating the competence theory. The same point has been made more recently with respect to Bayesian CL theories (e.g., Bowers & Davis, 2012; Oaksford & Chater, 2007; Rogers & Seidenberg, 2011). What does not appear to have been realized is the implications of this distinction between competence and performance for the evaluation of CL theories. Franks (1995) argued that, in the absence of a performance theory, competence theories have limited predictive utility. More precisely, any competence theory prediction could be undermined by performance factors. Consequently, competence theories (and hence CL explanations) are not falsifiable in the standard Popperian sense.

Popperian falsifiability is not necessarily a requirement for scientific theorizing (Lakatos, 1970), but the implications of the argument concerning performance factors are illustrated by the simulations of Cooper, Yule, and Fox (2003), who compared the behavior of three classes of model—Bayesian, associationist, and hypothesis testing—with that of subjects on a sequential category-learning task (medical diagnosis). During the task, subjects completed a series of trials where they were able to query the presence/absence of a virtual patient's symptoms (e.g., headache, temperature, etc.) before making a diagnosis. Initially subjects were only able to guess, but feedback allowed them to learn the symptom/disease associations (which were probabilistic). Subjects were also encouraged to minimize the number of symptoms queried on each trial when making their diagnoses. Good performance on the task required that, for each virtual patient, subjects query those symptoms that were most informative, adjusting the order of their queries on the fly based on the information they received in response to their queries. The authors found that all three classes of model, when supplemented with appropriate performance factors, were able to replicate the key behavioral effects. Thus, for the Bayesian model it was necessary to assume that the initial probabilities (priors) of each symptom given each disease were less than 0.5 and that symptoms where queried based on an information-theoretic measure which embodied a confirmation bias. In contrast, the associationist model required strong negative initial associations between symptoms and diseases together with a form of recency bias implemented by adding low levels of noise to associations on each trial. For the hypothesis testing model, subject performance was simulated by the use of a confirmatory strategy supplemented with moderate time-based decay of the contents of working memory. In summary, while the performance factors differed across the classes of model, the net result was one of indistinguishability. The work is particularly apposite, given that Xu and Tenenbaum (2007), in their study of word learning discussed above, argue strongly for a Bayesian model and against associationist or hypothesis testing accounts of word learning. Cooper et al.'s result suggests that, once performance factors are incorporated into each class of model, Xu and Tenenbaum's argument does not hold: At the behavioral level, a Bayesian model may be coerced into mimicking the performance of an associationist model by, for example, judicious choice of priors and the weighting of old and new evidence, just as an associationist model may be coerced into mimicking a Bayesian model by, for example, judicious choice of initial weights.

A related limitation of CL accounts of behavior is that they cannot make direct contact with one of the key traditional measures of the cognitive psychologist, namely response times. Given that CL accounts are concerned with what function is carried out by a system rather than how that function is computed, this may be justified, but it is a further barrier to relating CL accounts to behavior.

A focus on function is also more easily justified when an argument can be made that the processes under investigation can be decomposed into a linear chain of processing, as in Marr's conception of vision whereby the input is represented and re-represented in a sequence of stages as it is transformed from a retinal map via first the primal sketch and then a 2½ dimensional viewer-centered sketch to a three-dimensional, object-centered representation. A linear chain of this kind supports feedback between successive stages (and hence allows for necessary interactions between top-down and bottom-up processing), but it is still feasible to view each stage in the chain as performing a self-contained function. Moreover, one end of the chain at least is anchored in the external observable world. While this may be an appropriate decomposition for vision, it is less clear that higher cognitive functioning can be decomposed so neatly into stage-wise processing. Thus, ARL accounts of cognitive processes generally conceive of observable behavior as the product of multiple interacting subsystems (such as working memory, attention, goal generation, and maintenance, etc.), with the interactions between those subsystems not being amenable to linearization. This poses a fundamental dilemma for CL accounts of the cognitive system that hypothesize such subsystems. Specifically, should a CL account be directed toward the cognitive system as a whole or to the various hypothesized unobservable subsystems? Just as importantly, one cannot justifiably ask the "why" question of the functioning of subsystems without first addressing why the system as a whole might comprise those subsystems. For example, one cannot ask why working memory carries out such-and-such a function or should be optimized in such-and-such a way without first addressing why the cognitive system might comprise a working memory subsystem interacting with a range of other subsystems.[3]

The above criticisms of purely CL explanations go beyond recent critiques of Bayesian accounts of cognition (e.g., Bowers & Davis, 2012; Jones & Love, 2011; Marcus & Davis, 2013). They reflect general concerns about explanations pitched solely at the CL, regardless of whether those explanations are phrased in probabilistic, logical, or set-theoretic terms.

3. The role of neuroscience evidence

We have argued that CL accounts, in the absence of ARL accounts, are limited in their explanatory force. But could one not study the operation of the brain directly with neuroscience evidence alone? Notwithstanding Marr's analysis, some might argue that this approach has been successful in understanding the functioning of peripheral systems (with low-level visual processing being perhaps the most celebrated success; Hubel & Wiesel, 1962). However, application of the reductionist approach to higher cognitive processes, such as those involved in reasoning, decision-making, planning, problem-solving, and language, is problematic for several reasons.

Neuroscience offers a great variety of methods that might be deployed to investigate higher cognitive processes. These include methods that are entirely neurophysiological in nature (mapping neural pathways and connectivity), methods for tracking neural activity while performing a task (including numerous forms of functional neuroimaging with differing temporal and spatial resolution, as well as single cell recording in behaving primates, and, very occasionally, humans), and methods based on the analysis of the effects of brain lesions (including temporary lesions created via TMS and permanent lesions due to organic damage). Yet even with all of these methods at neuroscientists' disposal, an understanding of higher cognitive processes at the neural level remains elusive. Is it just a matter of time?

Consider the case of propositional reasoning. The neurophysiological/implementation level appears to be the wrong level to begin an analysis of a complex ability such as this. It is unclear where one would start. Simply scanning subjects while they attempt reasoning tasks (with no cognitive-level theory of how such tasks are solved) is unlikely to yield informative results. In fact, there are many neuropsychological and neuroimaging studies of reasoning (see Shallice & Cooper, 2011, pp. 464–478 for a review). The naïve observer might expect that these would have informed cognitive-scientific understanding of the ability. Yet the studies present a confused picture, with a complex network of frontal and parietal cortical regions implicated across a range of reasoning tasks, and, while clearly defined reasoning deficits arise in some neuropsychological patients, one cannot easily infer a theory of reasoning from the neural data. Indeed, the neural data have low discriminative power, with ongoing debates suggesting that the data are not even able to differentiate convincingly between existing cognitive-level theories (cf. Coltheart, 2006, 2011). Perhaps just as critically, Goel (2007) argues that interpretation of the neural data from reasoning experiments requires a re-evaluation of psychological theories of reasoning, with the neuroscience evidence indicating that human reasoning involves separate systems for dealing with (a) familiar and unfamiliar material, (b) conflicting information and belief bias, and (c) certain and uncertain information. The evidence from functional neuroimaging to support these dissociations is far from conclusive, and, as Shallice and Cooper (2011) point out, dual-system theories of reasoning were proposed long before neuroimaging evidence was available. Those theories draw on the same distinctions discussed by Goel (see, e.g., Evans, Barston, & Pollard, 1983). So in this domain at least neuroscience evidence does not allow one to “explain” reasoning behavior, nor does it trigger the complete re-evaluation of cognitive theory that Goel (2007) suggests.

What then is the role of the neuroscience evidence? Consider first neuroimaging in its various guises. Coltheart (2006, 2011) argues forcefully that neuroimaging evidence has not informed psychological theory. His specific concern is in the use of neuroimaging evidence to discriminate between competing psychological theories, and he provides six case studies where neuroimaging has, contrary to the various researchers’ explicitly stated intentions, failed to have any discriminating force. Coltheart’s arguments against the utility of neuroimaging in cognitive theorizing are not arguments *in principle*, and it is conceivable that future imaging studies will support the kind of theory discrimination that Coltheart seeks. In this respect the model-based methods discussed by Love (2015) appear particularly promising. Note, though, that those methods are *not* reductionist—critically they require specification of a cognitive-level theory that might be coupled to neural-level data.

One reason why neuroimaging evidence is limited in the extent to which it can inform psychological theory is the “reverse inference” problem: Can one infer the involvement of cognitive-level processes from activation at the neural level (or to be more precise, from differences in the BOLD response at the neural level)? Even if one had a clear understanding of the relation between the firing of neurons and the BOLD signal as measured by fMRI (and we do not; Maier et al., 2008), it remains possible that a single cognitive function might be implemented by a network of neural regions, or that a single region might implement multiple cognitive functions. For this reason cognitive neuropsychology—the investigation of cognitive deficits following neural damage or degeneration—provides an invaluable complementary source of neural-level evidence. Cognitive neuropsychology allows one to draw conclusions about the function of the intact cognitive system from patterns of behavior following breakdown of that system. Shallice and Cooper (2011) argue that neuroimaging and neuropsychology make different inferential assumptions linking function to structure. While those assumptions may not be valid in any specific case, inferences that are supported by *both* neuroimaging and neuropsychological evidence are likely to be robust because in the event that the assumptions of both methods are violated, there is no reason why they should then yield a consistent localization of the function under consideration.

An alternative approach to solving the reverse inference problem has been developed by Poldrack and colleagues (e.g., Lenartowicz, Kalar, Congdon, & Poldrack, 2010; Yarkoni, Poldrack, Nichols, Van Essen, & Wager, 2011). This approach automatically extracts activation foci (coordinate representations of brain regions showing activation or deactivation in one task relative to a control task) from reported studies and associates these with keywords in the corresponding write-up of the study. Machine learning techniques are then used to discover commonalities between studies associated with the same keyword and discriminate these from studies associated with other keywords. While the approach introduces many sources of noise into the process (e.g., by not differentiating between activation and deactivation of a region and by not carefully examining the reported contrasts and classifying them appropriately), this noise is countered by the large database of imaging studies that has developed over the last 25 years, and the approach has been successful at drawing distinctions between the neural activity associated with a range of keywords. Thus, Yarkoni et al. (2011) report that the neural regions extracted for, for example, the keywords “attention” and “executive” discriminate between studies of attention and studies of executive function with a mean classification accuracy of 64% (which is significantly above the chance level of 50% at the $p = .001$ level on the dataset used).

This automated meta-analytic approach attempts to understand brain function without recourse to either the ARL or the CL. While it draws upon cognitive constructs such as “attention” and “executive,” it does so in an informal way without attempting to characterize those constructs in any way beyond their use in psychological discourse. While the approach might serve as a useful first step in relating cognitive function to neural structure, it fails to advance understanding of cognitive function in the Marr sense for several reasons. Firstly, it is necessarily coarse in its discriminations. While Yarkoni et al. (2011) consider 25 keywords, those keywords refer to very broad concepts (“pain,” “emotion,” “imagery,” etc.), and any attempt to focus in on more specific or more clearly defined subfunctions will be met with major difficulties in detecting a signal amid the many sources of noise in the data. Furthermore, even with broad concepts, the discrimination accuracy, while statistically significant, is low. More critically, however, even if discrimination accuracy was high, the method would only allow one to answer “where” questions (i.e., where is function X computed?). It cannot answer true IL questions (i.e., how do neurons in the identified regions *compute* the function?), or directly answer the “what,” “why,” or “how” questions that characterize Marr’s higher levels.

Suppose, though, that neuroscience was to yield a complete understanding of the neural basis of behavior, such that, for a simple task (say choice reaction time), we understood how neural firing in low-level perceptual areas ultimately led to neural firing in peripheral motor neurons. Such an “explanation” would still be incomplete if it did not abstract away from the neural hardware of the individual. More than that, however, the explanation would need presumably to abstract away from neural hardware altogether if we are, for example, to be able to reimplement the relevant functioning in different hardware. In other words, the explanation would need to operate at a level above the IL. Thus, the critical “multiple realisability” argument made by Marr (1982) against reductionist neuroscience—namely that there is no reason to believe that a cognitive system could not be realized in any number of different computational substrates—still holds. The flip side of this argument is the claim (which we accept may be more an article of faith than an empirical fact) that cognitive-level constructs, whether they be beliefs and goals (in reasoning and problem-solving), phonemes or morphemes (in linguistic theory), or action schemas (in routine action selection) are causally efficacious and therefore cognitive theory, operating at the level of such constructs, provides a level of abstraction that is critical to understanding common function across individuals and across species.

4. Beyond single-level explanations: The role of cognitive architectures

One way to address the limitations of purely CL explanations is to develop ARL models that instantiate (at least to a first approximation) relevant CL principles and that are therefore able to approximate solutions provided by CL analyses. For example, one might develop a Bayesian CL explanation into an ARL model by introducing a mechanism that instantiates Bayesian inference. The benefit of such an approach would come from the explicit mapping between ARL mechanisms and Bayes' rule, which would allow such models to impose constraints on the assumptions of the latter and suggest explanations of suboptimal observed behavior.

This approach has recently been adopted by Bayesian theorists in the form of "rational process models" (Griffiths, Vul, & Sanborn, 2012; Sanborn, Griffiths, & Navarro, 2010; Shi, Griffiths, Feldman, & Sanborn, 2010). The proposed models use Monte Carlo algorithms (e.g., importance sampling, particle filters, Markov Chain Monte Carlo methods) which approximate Bayesian inference by sampling from a probability distribution. The predictions of rational process models can be degraded from the optimal Bayesian solution to suboptimal (i.e., human-like) performance by, for example, reducing the number of samples taken.

While lauding the efforts of Bayesian theorists to acknowledge their critics and address the question of links to the ARL, the approach has several limitations. First, the direction of influence in this approach remains entirely top-down, with no constraints being placed by the rational process model upon, for example, the priors of the Bayesian model; the sole purpose of rational process models is to sample a probability distribution from the Bayesian model, with the ARL model being post hoc.

Second, Monte Carlo methods are a class of algorithms specified at an abstract level which omit details of how sampling is achieved. This degree of abstraction places the key claim that ARL approximations of CL solutions are achieved by sampling from a probability distribution alongside the claims that problem-solving is a search over a problem space or that learning is a gradient descent on an error surface; high-level characterizations of processes that may be implemented by numerous different algorithms. What unites them is their agnosticism about cognitive architecture; they contain no claims about the structure of the mechanisms involved or the nature of the limits of those mechanisms.

While such agnosticism might be justified if behavioral measures cannot distinguish between more detailed characterizations of the processes, specifying such detailed hypotheses of representations and mechanisms is the aim of ARL models, and it is this level of detail that allows them to account for additional behavioral measures, such as strategy differences and response times. An alternative, potentially more unifying approach to bridging the ARL and CL therefore would be to take existing ARL frameworks or models and to consider how they can be related to Bayesian inference (or some other CL principle).

One move in this direction is the "computational rationality" approach (Howes, Lewis, & Vera, 2009; Lewis, Howes, & Singh, 2014), which applies Russell and Subramanian's (1995) notion of bounded optimality for artificial intelligence agents to the analysis of human behavior. The underlying assumption of this approach is that human behavior is generated by cognitive processes that are adapted not only to the structure of the environment but also to the (bounded) cognitive architecture, and it is this latter feature which determines suboptimal human performance compared to an optimal Bayesian CL solution. Computational rationality differs significantly therefore from the rational process models approach in requiring not only the specification of a task environment but also a cognitive architecture and a set of programs to execute behaviors to be specified at the ARL.

A second approach has been followed by Bayesian modelers who have sought to identify memory mechanisms that implement the Monte Carlo algorithms of rational process models. This has resulted in recent demonstrations that the importance sampling algorithm can be implemented by exemplar-based memory mechanisms (Abbott, Hamrick, & Griffiths, 2013; Shi & Griffiths, 2009; Shi et al., 2010). These mechanisms approximate sampling from the prior distribution by retrieving memories (hypotheses) according to their degree of similarity to the current context and then weighting them by the likelihood function to obtain an approximation to the posterior distribution.

A similar approach to bridging CL and ARL theories also does so by memory retrieval mechanisms based on relevance to the current context (Anderson, 2007; Lebiere et al., 2013; Thomson & Lebiere, 2013). Implemented within the ACT-R cognitive architecture, this method assumes that declarative knowledge chunks in long-term memory have a level of activation that determines their probability of retrieval. In this conception, chunk activation is proportional to log-likelihood, and a chunk's activation is updated (the likelihood function) from its initial base level (prior) to its final level (posterior) according to its similarity to the current context. This mechanism is essentially identical to the exemplar-based models of importance sampling mentioned above (although a formal equivalence has not been demonstrated), opening the possibility that many mechanisms exist that exhibit the same properties.

Given this possibility, the question then arises as to how one should compare and evaluate these alternatives. One obvious criterion is the number and nature of the constraints each mechanism imposes on the CL theory. In contrast to the exemplar-based models which are relatively simple in their representational and algorithmic assumptions, the other two approaches are both embedded within the broader theoretical commitments of a cognitive architecture (ACT-R) that imposes strong constraints on declarative and procedural learning and the learning of activation and utility values. These ARL constraints severely limit the number and nature of the hypotheses at the CL and therefore the kinds of inferences that can be made. In addition, the mechanisms determining these learning and control processes are consistent with both behavioral and neural data (Lebiere et al., 2013; and see below), leading to the possibility that through these connections to both the CL above and the IL below, cognitive architectures like ACT-R can provide a bridge that unites Marr's three levels.

One can make many arguments for the utility of developing cognitive models within a cognitive architecture (see, e.g., Cassimatis, Bell, & Langley, 2008; Newell, 1990), but adopting the concept of cognitive architecture is in fact highly consistent with Marr's original approach to vision. His decomposition of visual processing into a series of stages effectively specifies a visual processing architecture. Importantly, Marr does not provide a CL account of vision as a whole. Rather he argues for the provision of CL (and ARL and IL) accounts of each component of his visual processing architecture. There is, therefore, a critical difference in the starting point of Marr and that of many Bayesian approaches to cognitive processing, such as those of Norris (2006) or Griffiths et al. (2008). Trying to provide a CL account of the organism as a whole fails to take account of the (assumed) functionally modular substrate that supports cognitive processing. Cognitive architectures decompose the cognitive system into interacting functional subcomponents, and, following Marr, it is these components that are most usefully characterized at the CL, ARL, and IL.[4] At the same time, a cognitive architecture by itself does not normally uniquely determine how a specific task might be accounted for at the ARL. The strength of the "computational rationality" approach of Howes, Lewis, and colleagues discussed above is that it addresses this relationship. At the same time, the approach of Lebiere and colleagues described above aims to provide task-independent CL accounts (with links to the IL) of the functional subcomponent(s) responsible for retrieval from long-term memory.

Work involving those cognitive architectures in which functional subcomponents are associated with neural regions or structures also promises to help bridge the neuroscience and psychological levels. For example, over the last decade Anderson and his colleagues have attempted to map the core modules of the ACT-R cognitive architecture onto brain regions via fMRI data relating to the BOLD response (e.g., Anderson, 2005; Anderson, Qin, Sohn, Stenger, & Carter, 2003; Qin et al., 2003, 2004). This has led to the identification of the five core processing modules of ACT-R with associated brain regions. The mapping of elements between two levels in this manner has allowed models and data from each to influence the other. Thus, in fMRI studies of the development of children's algebra skill, a computational model of eye movement sequences, memory retrievals, problem-solving steps, and responses was able to account for the relative changes in patterns of activation in different brain regions due to learning. The direction of influence is not just in the direction of ARL to IL. In Anderson's work, imaging data have also influenced developments in the cognitive architecture. For example, the original ACT-R architecture contained a single goal buffer containing both control knowledge and problem state knowledge. BOLD response patterns suggested, however, that changes in the former correlated with activity in the anterior cingulate, while changes in the latter corresponded to activity in a parietal region. This led to a separation in the cognitive architecture between the goal buffer for control state knowledge and an "imaginal" buffer for problem state knowledge (Anderson, 2005).[5]

While the mapping between the IL and the ARL within ACT-R is arguably too coarse (e.g., at the time of writing there are regions of cortex, as well as subcortical structures, with no associated ACT-R module, and the ACT-R mapping assumes that all IL processes are organized bilaterally), the work illustrates the profitable interaction between the ARL and the IL. The architectural approach also addresses a key limitation of Poldrack's approach to reverse inference as discussed in Section 3. Recall that this approach aims to map structure to function by determining, across a very large database of functional neuroimaging studies, the brain regions or networks that allow one to discriminate between studies of cognitive-level constructs such as attention and executive function. A significant concern is that, as operationalized by Poldrack and colleagues, those cognitive-level constructs are described informally. However, this may be addressed by the development of process models within a cognitive architecture for the tasks in question, thus ensuring that one has robust hypotheses concerning how different architectural functions or components are implicated in different tasks.

To summarize, then, while neuroscience evidence cannot be used in a purely bottom-up fashion to understand cognitive functioning, it may still provide a valuable complement to cognitive (i.e., algorithmic and representational) level theorizing, providing supporting evidence for specific theories of the functional organization of some faculty, particularly when neuroimaging and neuropsychology provide converging evidence for cognitive-level subfunctions, or when a theory of the organization of the cognitive architecture can bridge the ARL and the IL.

5. Conclusion

In the 30+ years since Marr proposed his tri-level hypothesis, significant theoretical advances have been made at each level of analysis. However, in our view the most critical advance has been the development of plausible cognitive architectures that decompose the cognitive system into interacting functional components. Given such a decomposition, one might begin to develop CL, ARL, and IL accounts of the components. However, this would not be sufficient for neural-level data to inform cognitive theory. For this, it is necessary to develop ARL accounts of performance on the tasks used to generate the neural-level data. We suggest this is most likely to be productive when attempted within a putative cognitive architecture in which the functional components of the architecture have CL descriptions or explanations. Such an approach will ground the function of the architectural components while allowing cognitive explanations to be linked to behavioral data, and neuroimaging and neuropsychological studies to inform our understanding of the neural instantiation of architectural components.

Notes

1. Certainly "explanation," however defined, would seem to require more than mere mimicry, but equally to label Newell and Simon's seminal work on problem-solving as "unenlightening" would seem churlish. Marr's criticism might be valid if Newell and Simon were just concerned with mimicry (i.e., data-fitting), but, in accounting for human problem-solving behavior, Newell and Simon developed the problem space computational model (PSCM)—an account of the computational problem that problem-solving is required to solve. Thus, a key claim of the work of Newell and Simon was that problem-solving should be understood as the selection of a sequence of operators that would transform a given state into a goal state. While Newell and Simon may not have been explicit about it, this level of theorizing frames the problem (of understanding human problem-solving) squarely on Marr's CL.
2. It is tempting to suggest that Bayesian accounts simultaneously address the "what" and the "why" question. If this were true, it would enhance the plausibility of any Bayesian account. However, as Bowers and Davis (2012) argue, Bayesian accounts do not necessarily do this. Indeed, Bowers and Davis characterize many such accounts as "just-so" theories, a view consistent with that developed by Marcus and Davis (2013).
3. Note that decomposing the overall function effected by a system into a function of subsystems invokes algorithmic concepts because it expresses constraints on the order of processing and on intermediate results. Strictly speaking, therefore, this decomposition already requires going beyond the CL.
4. In addition to the standard arguments for cognitive architectures, Simon (1962) argued that hierarchically organized systems (i.e., systems composed of subsystems) provide a kind of stability of subfunction that supports evolutionary processes. So the concept of a cognitive architecture is consistent with the kind of evolutionary analysis of the human cognitive system that many contemporary proponents of CL-first accounts espouse.
5. ACT-R is used here as a well-worked example, but in principle the comments apply equally to any cognitive architecture that attempts to bridge higher levels with the IL. Thus, the SAL hybrid multi-level architecture of Jilk, Lebiere, O'Reilly, and Anderson (2008) provides a further example.

Acknowledgments

We are grateful to three anonymous reviewers for insightful and constructive comments on an earlier version of this paper.

References

Abbott, J. T., Hamrick, J. B., & Griffiths, T. L. (2013). Approximating Bayesian inference with a sparse distributed memory system. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 1686–1691). Austin, TX: Cognitive Science Society.

Anderson, J. R. (2005). Human symbol manipulation within an integrated cognitive architecture. *Cognitive Science*, *29*(3), 313–341.

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* New York: Oxford University Press.

Anderson, B. L. (2015). Do teleological explanations have a place in psychological theory? *Topics in Cognitive Science*, *7*. DOI: 10.1111/tops.12135.

Anderson, J. R., Qin, Y., Sohn, M.-H., Stenger, V. A., & Carter, C. S. (2003). An information-processing model of the BOLD response in symbol manipulation tasks. *Psychonomic Bulletin & Review*, *10*, 241–261.

Baggio, G., vanLambalgen, M., & Hagoort, P. (2015). Logic as Marr's computational level: Four case studies. *Topics in Cognitive Science*, *7*(2), 287–298.

Baronchelli, A., Ferrer-i-Cancho, R., Pastor-Satorras, R., Chater, N., & Christiansen, M. H. (2013). Networks in cognitive science. *Trends in Cognitive Sciences*, *17*(7), 348–360.

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, *138*(3), 389.

Cassimatis, N. L., Bell, P., & Langley, P. (2008). Ability, breadth, and parsimony in computational models of higher-order cognition. *Cognitive Science*, *32*, 1304–1322.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: The MIT Press.

Coltheart, M. (2006). What has functional neuroimaging told us about the mind (so far)? *Cortex*, *42*, 323–331.

Coltheart, M. (2011). What has functional neuroimaging told us about the organisation of mind? *Cognitive Neuropsychology*, *28*(6), 397–402.

Cooper, R. P., Yule, P., & Fox, J. (2003). Cue selection and category learning: A systematic comparison of three theories. *Cognitive Science Quarterly*, *3*(2), 143–182.

Cummins, R. (1983). *The nature of psychological explanation*. Cambridge, MA: The MIT Press.

Evans, J. S., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory and Cognition*, *11*, 295–306.

Franks, B. (1995). On explanation in the cognitive sciences: Competence, idealisation, and the failure of the classical cascade. *British Journal for the Philosophy of Science*, *46*(4), 475–502.

Goel, V. (2007). Anatomy of deductive reasoning. *Trends in Cognitive Sciences*, *11*, 435–441.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *The Cambridge handbook of computational psychology* (pp. 59–100). Cambridge, UK: Cambridge University Press.

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, *21*, 263–268.

Howes, A., Lewis, R. L., & Vera, A. (2009). Rational adaptation under task and processing constraints: Implications for testing theories of cognition and action. *Psychological Review*, *116*(4), 717–751.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, *160*, 106–154.

Jacob, F. (1977). Evolution and tinkering. *Science*, *196*, 1161–1166.

Jilk, D. J., Lebiere, C., O'Reilly, R. C., & Anderson, J. R. (2008). SAL: An explicitly pluralistic cognitive architecture. *Journal of Experimental and Theoretical Artificial Intelligence*, *20*, 197–218.

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, *34*(4), 169–188.

Lakatos, I. (1970). Falsification and the methodology of scientific research programs. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (pp. 91–196). Cambridge, UK: Cambridge University Press.

Lebiere, C., Pirolli, P., Thomson, R., Paik, J., Rutledge-Taylor, M. F., Staszewski, J., & Anderson, J. R. (2013). A functional model of sensemaking in a neurocognitive architecture. *Computational Intelligence and Neuroscience*, *2013*, 1–29.

Lenartowicz, A., Kalar, D. J., Congdon, E., & Poldrack, R. A. (2010). Towards an ontology of cognitive control. *Topics in Cognitive Science*, *2*(4), 678–692.

Lewis, R. L., Howes, A., & Singh, S. (2014). Computational rationality: Linking mechanism and behavior rough bounded utility maximization. *Topics in Cognitive Science*, *6*(2), 279–311.

Love, B. C. (2015). The algorithmic level is the bridge between computation and brain. *Topics in Cognitive Science*, *7*(2), 230–242.

Maier, A., Wilke, M., Aura, C., Zhu, C., Ye, F. Q., & Leopold, D. A. (2008). Divergence of fMRI and neural signals in V1 during perceptual suppression in the awake monkey. *Nature Neuroscience*, *11*, 1193–1200.

Marcus, G. F., & Davis, E. (2013). How robust are probabilistic models of higher-level cognition? *Psychological Science*, *24*, 2351–2360.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York, NY: Henry Holt and Co.

Newell, A. (1982). The knowledge level. *Artificial Intelligence*, *18*(1), 87–127.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review*, *113*(2), 327–357.

Oaksford, M., & Chater, N. (2001). The probabilistic approach to human reasoning. *Trends in Cognitive Sciences*, *5*(8), 349–357.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford, UK: Oxford University Press.

Phillips, S., & Wilson, W. H. (2010). Categorial compositionality: A category theory explanation for the systematicity of human cognition. *PLoS Computational Biology*, *6*(7), e1000858.

Qin, Y., Carter, C. S., Silk, E., Stenger, V. A., Fissell, K., Goode, A., & Anderson, J. R. (2004). The change of the brain activation patterns as children learn algebra equation solving. *Proceedings of National Academy of Sciences USA*, *101*(15), 5686–5691.

Qin, Y., Sohn, M.-H., Anderson, J. R., Stenger, V. A., Fissell, K., Goode, A., & Carter, C. S. (2003). Predicting the practice effects on the blood oxygenation level-dependent (BOLD) function of fMRI in a symbolic manipulation task. *Proceedings of the National Academy of Sciences USA*, *100*(8), 4951–4956.

Rogers, T. T., & Seidenberg, M. S. (2011). Distinguishing literal from metaphorical applications of Bayesian approaches. *Behavioral and Brain Sciences*, *34*(4), 211–212.

Russell, S., & Subramanian, D. (1995). Provably bounded-optimal agents. *Journal of Artificial Intelligence Research*, *2*, 575–609.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*, 1144–1167.

Shallice, T., & Cooper, R. P. (2011). *The organisation of mind*. Oxford, UK: Oxford University Press.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana: University of Illinois Press.

Shi, L., & Griffiths, T. L. (2009). Neural implementation of hierarchical Bayesian inference by importance sampling. *Advances in Neural Information Processing Systems*, *22*, 1669–1677.

Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin & Review*, *17*(4), 443–464.

Simon, H. A. (1962). The architecture of complexity. *Proceedings of the American Philosophical Society*, *106*, 467–482.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalisation, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629–640.

Thomson, R., & Lebiere, C. (2013). Constraining Bayesian inference with cognitive architectures: An updated associative learning mechanism in ACT-R. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 3539–3544). Austin, TX: Cognitive Science Society.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245.

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, *8*(8), 665–670.