Psychonomic Bulletin & Review
1999, 6 (2), 173-182

Computational analyses in cognitive neuroscience: In defense of biological implausibility

ITIEL E. DROR
Southampton University, Southampton, England
and
DONALD P. GALLOGLY
University of Texas, Austin, Texas

Because cognitive neuroscience researchers attempt to understand the human mind by bridging behavior and brain, they expect computational analyses to be biologically plausible. In this paper, biologically *implausible* computational analyses are shown to have critical and essential roles in the various stages and domains of cognitive neuroscience research. Specifically, biologically implausible computational analyses can contribute to (1) understanding and characterizing the problem that is being studied, (2) examining the availability of information and its representation, and (3) evaluating and understanding the neuronal solution. In the context of the distinct types of contributions made by certain computational analyses, the biological plausibility of those analyses is altogether irrelevant. These biologically implausible models are nevertheless relevant and important for biologically driven research.

The goal of cognitive neuroscience researchers is to uncover the cognitive processes that underlie human intelligence and the human mind, and not to understand any computational entity, per se. Because humans are biological, neuroscience research and its integration into cognitive theory have been the focus of the transition from cognitive science to cognitive neuroscience (see, e.g., Kosslyn & Koenig, 1992). As a relatively new approach, cognitive neuroscience is still in the process of exploring and establishing the various ways knowledge from different domains can be combined into a more complete and accurate description of cognition. The foundation and strength of cognitive neuroscience are in bridging knowledge from different and distinct disciplines. As the field progresses in its empirical discoveries and its theoretical foundations, cognitive neuroscientists are beginning to realize how the integration of different domains can contribute to the overall understanding of the human mind, as well as how such an endeavor may be limited and even misleading.

Computational analyses[1] play a major role in cognitive neuroscience (Dror, 1994). The cognitive neuroscience community has reconceptualized the role of computational analyses; with the new emphasis and connection to neuroscience, biological plausibility has been put forth as a guiding principle in the construction of computational models of cognition. The acceptance of computational investigations has depended on their biological plausibility.[2] Since cognitive processes are implemented in the brain, any model of these processes that is biologically implausible must be incorrect. In this light, computational analyses need to conform to the known principles by which the nervous system operates and, specifically, to the neural underpinnings of the process being studied (see, e.g., Bower, 1990; Crick, 1989; Gluck & Thompson, 1987; Grossberg, 1987). Thus, if a model is biologically implausible, then it should be rejected prima facie.

Discussions on computational analyses within cognitive neuroscience have adopted this point of view as a basic framework and have focused on whether specific analyses are biologically plausible or not. The view that computational models must be biologically plausible seems to have been adopted a priori by the majority of researchers who try to understand the computations in the brain that underlie cognition. Cognitive neuroscience researchers who have been criticized on the grounds of the biological implausibility of their computational analyses have creatively defended their work by trying to show that their models are nevertheless biologically plausible. By doing so, they have embraced and strengthened their critics' point of view that biological implausibility rejects the possible contribution of their analyses.

This paper was supported by grants awarded by AFRL (Air Force Research Laboratory) to I.E.D. We owe thanks to many people who have read various versions of this paper or who have discussed related issues with us. These discussions about biologically implausible computational analyses in cognitive neuroscience, and perhaps the strong opposition we found to their acceptance, prompted us to write this paper. In particular, we would like to thank Steve Kosslyn, Bill Estes, Randy O'Reilly, Christof Koch, Mike Mozer, Stevan Harnad, and Allan Pantle. We also want to thank Randi Martin, David Adams, William Bechtel, Art Glenberg, and an anonymous reviewer for very helpful comments. We need to point out, however, that these people do not necessarily support our arguments (some of them even bluntly oppose them). The ideas expressed in the paper reflect our own viewpoint, and only we carry the responsibility for the arguments and views in this paper. Correspondence should be addressed to I. E. Dror, Department of Psychology, Southampton University, Highfield, Southampton SO17 1BJ, England (e-mail: dror@coglab.psy.soton.ac.uk; www: http://www.cogsci.soton.ac.uk/~dror).

To illustrate this framework, let us look at a certain class of models as a case study. The biological implausibility of the backpropagation learning algorithm has been the grounds for criticizing a variety of models (see, e.g., Aizawa, 1992; Bugmann, 1997; Crick, 1989; Crick & Asanuma, 1986; Grossberg, 1987; O'Reilly, 1996; Thompson & Gluck, 1990). Trying to justify their models, researchers have responded to such criticism by arguing that (1) although backpropagation is not literally and actually applied in the brain, it reflects the same algorithm used by the brain (e.g., Mazzoni, Andersen, & Jordan, 1991; Zipser & Andersen, 1988); (2) although the learning algorithm is implausible as a learning algorithm for the nervous system, the final configuration of the network is biologically plausible, and the fact that it was derived in an implausible way does not detract from the final "mature" model (e.g., Kettner, Marcario, & Port, 1993; Lehky & Sejnowski, 1988); (3) a single unit in the network is not analogous to a single neuron in the brain; if the backflow of information is transmitted as normal activity conveyed through back projection (rather than through the axon itself), then it could be carried out by the brain (e.g., Rumelhart, 1990); or alternatively, a single unit in the network stands for a population of neurons (e.g., Kosslyn & Koenig, 1992, p. 43); (4) although it may not be biologically plausible, it is functionally equivalent to, or can be modified into, one that is (e.g., Jordan, Flash, & Arnon, 1994; Murre, 1992; Schmajuk & DiCarlo, 1992); (5) it is a simplified, stripped-down version of the brain (e.g., Sejnowski, 1986); and (6) the criticism is unfounded, and the learning algorithm is biologically plausible (e.g., Levin, 1991).

All of these arguments have presupposed that biologically implausible models cannot contribute to the understanding of human and animal cognition, and hence have no place in cognitive neuroscience. Such discussions are not limited to a specific class of models or to a certain computational approach. Indeed, this criticism has been applied to all types of computational models.

Computational models have been characterized as falling along a continuum of biological plausibility, from realistic to simplistic. Realistic models remain strictly adherent to the known properties of the nervous system, as exemplified by Mead's (1989) artificial retina chips, whereas simplistic models make generalizations and simplifications. Models are often evaluated by how closely they adhere to the nervous system; models that are more realistic are considered more viable and to have more explanatory power in cognitive neuroscience.

In this paper, we question the position that the legitimacy and value of computational models necessarily depends on their biological plausibility, and we argue in defense of biologically implausible computational models and analyses. We show that even cognitive neuroscientists and other researchers that may be biologically driven can benefit from biologically implausible models. We argue not only in defense of simplistic models, but also (and mainly) for computational investigations that are completely divorced, in practice and in theory, from any aspect of the nervous system. By doing so, we show that even in the most extreme dissociation from biology, computational models can still be relevant and contribute to understanding the biological system and the cognitive processes it computes. Although biologically plausible computational analyses are critical for cognitive neuroscience, we show that biologically implausible models have also made meaningful contributions to our understanding of human and animal brain processes. Furthermore, we illustrate that biologically implausible computational analyses are necessary because one cannot completely understand the process that is being studied by considering only the mechanism that is implemented in the brain. We argue that in the context of the specific types of contributions that such models make, the biological perspective is altogether irrelevant.

To consider the role of biologically implausible computational analyses in cognitive neuroscience, we first need to examine its foundations. Cognitive neuroscience is distinct from biopsychology and neuroscience proper. In the latter, computational analyses—or to be more precise, neuromodeling—are focused solely on mimicking the biological functions of the nervous system. The aim of such disciplines and their models is limited to the biological scope, and hence it may be justified to exclude models that are biologically implausible from those disciplines. In contrast, cognitive neuroscience has a broader scope, one in which cognition and mind are intertwined with research about the biological brain. However, by itself this does not necessitate the acceptance of biologically implausible models. The conceptualization of the cognitive neuroscience framework establishes the role of computational analyses, which in turn determines the acceptance or rejection of biologically implausible models. As presented in this paper, the foundations of cognitive neuroscience are open to different interpretations in this matter.

We will start with observations on Marr's (1982) highly influential approach to understanding complex information processing systems. Then we will describe Kosslyn's (Kosslyn & Koenig, 1992) conceptualization of the different domains that contribute to our understanding of the human mind and brain. Both approaches contribute greatly to the foundations of cognitive neuroscience. Although their basic frameworks are very similar and include comparable components, they present a somewhat different mode of interaction and integration between those components. This difference puts computational analyses in a different light and has far-reaching implications for the validity and role of biologically implausible computational analyses in cognitive neuroscience.

THE COGNITIVE NEUROSCIENCE FRAMEWORK AND THE ROLE OF COMPUTATIONAL ANALYSES WITHIN IT

Marr (1982) described three levels of analysis that are needed in order to understand complex information processing machines: (1) The *computational theory* addresses

the overall goals of the computation; it examines questions such as, "What is the goal of the computation, and why is it appropriate?" (p. 25). (2) The *representation and algorithm* translates the computational theory into specific computational terms. This level examines the computational mechanisms used to map input into output: "What is the representation for the input and output, and what is the algorithm for the transformation?" (p. 25).[3] (3) The *hardware implementation* level introduces the biological manifestation of the computational analysis for the first time; this level examines how the computational analysis can be "realized physically" (p. 25).

Marr's (1982) *three levels of analysis* approach has influenced many computational researchers and has had a very strong impact in forming cognitive neuroscience. The importance of Marr's view, in the context of this paper, is that the analyses that are needed for understanding information processing are conceived as levels within a hierarchy: computational theory at the top, representation and algorithm at the center, and hardware implementation at the bottom.

Although the three levels depend on each other, they are not completely determined by the other levels. Rather, higher levels define the framework for, and are constrained by, the lower levels. As Marr (1982) stated, "These three levels are coupled, but only loosely. The choice of an algorithm is influenced, for example, by what it has to do and the hardware in which it has to run" (p. 25). That is, although the levels are only loosely coupled because there is a one-to-many relation between higher and lower levels (i.e., a single algorithm may be implemented in a variety of different types of hardware), they are coupled together because what is defined at the higher levels must be supportable by the lower levels (i.e., the algorithm needs to be one that can be carried out with the available hardware). Hence, the computational investigations at the level of representation and algorithm must be *biologically plausible.*

Following Marr's (1982) approach, many cognitive neuroscientists have adopted a framework whereby computational analyses that are biologically implausible are viewed as making no contribution to the understanding of human and animal cognition. If they are biologically implausible, then they cannot contribute to the understanding of the biological entity; that is, they are automatically refuted as viable options and are dismissed. To use Marr's terms, computational analyses that cannot be realized physically at the hardware implementation level have no place in the hierarchy needed to understand information processing.

The view that computational models must be biologically plausible seems to have been adopted a priori by the majority of researchers who try to understand the computations in the brain that underlie cognition. Cognitive neuroscience researchers who have been attacked on the grounds of the biological implausibility of their computational analyses have creatively defended their work by trying to show its possible biological plausibility (see specific examples in the introduction).

Using a perspective comparable to that of Marr (1982), Kosslyn and Koenig (1992) developed the wet mind approach as a foundation for cognitive neuroscience. However, their description is not composed of a hierarchy in which higher levels build on lower levels. Instead, Kosslyn and Koenig described the relationship as a "Cognitive Neuroscience Triangle" (p. 49). Behavior, in their conception, is placed at the top because the ultimate goal of our science is to understand behavior. At the other vertices are computational analyses and computer models, and neurophysiology and anatomy. This way of thinking about the field illustrates that, while the three disciplines work relatively independently, they all may draw upon and contribute to one another.

Following Kosslyn and Koenig's (1992) framework, a new notion of computational analyses may emerge. Computational analyses do not necessarily have to be biologically plausible. However, Kosslyn and Koenig did not pursue this notion and proceeded, as have many others, to emphasize biologically plausible models. As Kosslyn and Koenig stated, "The goal of computational analyses and computer modeling is not merely to fathom *any* possible way in which the behavior could be produced; instead, we want to know how a device *with the structure and properties of the brain* could generate the behavior" (p. 50, original emphasis).

The critical contribution and importance of biologically plausible computational analyses is self-evident, and we do not wish in any way to understate the value of biologically plausible models. However, we think it is critically important to legitimize and support the construction of biologically implausible models within the cognitive neuroscience community that researches the brain–behavior aspects of the mind. As we hope to demonstrate, biologically implausible models are an integral part of cognitive neuroscience. The computational level does not need to be restricted by the biological domain. For some types of contributions, the biology is important in guiding and constraining computational models, but for other types of contributions, the analyses do not need to be biologically plausible and should not be discussed in relation to biology.

CONTRIBUTIONS OF BIOLOGICALLY IMPLAUSIBLE COMPUTATIONAL ANALYSES TO COGNITIVE NEUROSCIENCE

Understanding and Characterizing the Problem

Regardless of their biological plausibility, computational analyses can provide important insights into understanding and characterizing the problem being studied. Cognitive neuroscientists explore how the brain solves very complex computational problems. A key factor in the attempt to understand the brain is a better un-

derstanding of the problems that the system encounters and resolves. Computational analyses that lead to such an understanding of the problems are very important, and their biological plausibility is irrelevant to their contribution. The mere question of whether they are biologically plausible reflects a basic misconception of the models and their contributions. Such confusion is caused by viewing the computational analyses as belonging to a category that needs to mirror (or at least relate to) biology. The very question of *if* (and how much, and in what sense) the model is biologically tenable is derived from conceptualizing the models in an incorrect context of the types of contributions they make (see Ryle, 1949, for discussion of such mistakes). In the context of the specific contribution of these analyses (see examples below), the biological dimension is irrelevant.

Many different types of computational analyses can be used to explore and characterize problems. These analyses include, for example, examining whether a problem is composed of computationally distinct subproblems. Biologically implausible computational investigations, such as the "split network" and "modular architecture" techniques, have been used to examine the computational compatibility of "what" and "where" tasks in visual processing (R. A. Jacobs, Jordan, & Barto, 1991; Rueckl, Cave, & Kosslyn, 1989) and categorical and coordinate spatial judgments (Kosslyn, Chabris, Marsolek, & Koenig, 1992).

In the split network technique, a problem is computed by a single undifferentiated network and by a split network within which the problem is decomposed into subproblems. Comparing the performance efficiency of the two networks reveals whether the problem was composed of computationally distinct subproblems: If the problem was computationally a single problem, then the undifferentiated network would outperform the split network due to positive crosstalk between the training examples. In contrast, if the problem is composed from distinct computational subproblems, then the undifferentiated network would suffer from negative crosstalk while the split network would allow specialization for each computational subproblem, and thus the split network would outperform the undifferentiated network (for more details, see Rueckl et al., 1989). Not only is this type of computational analysis biologically implausible because it uses backpropagation, but also, its basic format for encoding information and its architectural topology make it virtually dissociated from what we know about the underlying biology. For example, only a few units in the model are used to reflect the computations that underlie an entire cerebral hemisphere; the transformation of the image input to activation vectors does not encode information through the use of receptive fields or visual–spatial filters. The goal of the modular architecture technique is similar to that of the split network technique. However, in this technique the network itself divides the problem into distinct subproblems, rather then an a priori division by the experimenter. This is achieved by a gating network that moderates competition for learning between different modules (for more details, see R. A. Jacobs et al., 1991).

The results of such analyses are very important in cognitive neuroscience research; if a problem is composed of distinct computational subproblems, then its processing probably involves distinct cognitive modules and neuronal substrates. Indeed, visual processing is divided between the dorsal visual pathway (the "where" system) and the ventral visual pathway (the "what" system). The biologically implausible computational analyses helped explain and characterize the problem and contributed to the understanding of the biological system and the theory of visual processing (see, e.g., Kosslyn & Koenig, 1992).

There are a wide variety of other types of computational analyses that can help explain and characterize problems. For example, problems can be classified by their level of computational complexity. Such a classification establishes a hierarchy of complexity that reflects *qualitative* or *quantitative* distinctions. A qualitative hierarchy classifies the computational complexity of problems by the type of mechanism that is required to compute their underlying computations. A quantitative hierarchy classifies the computational complexity of problems by the amount of effort required to carry out the computations. Hence, a qualitative hierarchy is based on the differences in the structural mechanisms that are needed to compute the problems, whereas a quantitative hierarchy is based on the number of steps that are needed to perform the computations using a general computing mechanism, such as a Turing machine (see, e.g., Lewis & Papadimitriou, 1981).

An example of a qualitative hierarchy is the complexity of languages. Here, the distinction between each level in the hierarchy reflects a structural difference in the mechanism required to perform the computations underlying that language (Chomsky, 1956, 1962). In order of increasing complexity are regular languages, context-free languages, context-sensitive languages, and recursively enumerable languages. For the regular languages, the mechanism of the finite automata is sufficient for dealing with their complexity, the context-free languages require a pushdown automaton, the context-sensitive languages require a linear bounded automaton, and the recursively enumerable languages require a Turing machine. An example of a quantitative hierarchy is a classification of languages by the number of steps required by a Turing machine to carry out their underlying computations—for example, whether a language is decideable in a polynomial time bound, or whether certain conceptualizations of visual search can be computed in polynomial time (see discussion on visual search below).

The computational analyses of languages and their classification in a hierarchy of computational complexity enable one to gain an important understanding of the problem domain. The hierarchy not only gives a wider context that is insightful, but it also elucidates the computational characteristics of the problem we are trying to

explore. In this context, whether a certain type of mechanism is biologically plausible or not is altogether irrelevant. The computational analysis of the problem and its place within the complexity hierarchy are intended to help us understand and characterize the problem domain as well as the specific problem we are investigating. These mechanisms are used as a tool for computational analyses, and they are capable of achieving that goal regardless of the biological facts.

In the domain of language, such hierarchies have not only established one of the bases for researching and understanding languages, but have also raised important questions about the characteristics of natural languages and their usage. Questions such as, "Are natural languages regular, context free, context sensitive, or recursively enumerable?" were raised only after computational analyses that are disconnected from the biological facts provided an important framework for understanding and characterizing languages (e.g., Moyne, 1985). Then questions about their usage may follow—for example, why certain sentences that are computationally correct (for instance, sentences with triple-center embedding) are nevertheless not used by English speakers.

The use of computational complexity analysis is not limited to the study of language, but it is applicable to a wide variety of cognitive domains. Tsotsos (1990, 1991) has applied computational complexity analysis to the study of visual search. He has demonstrated that "considerations about the computational complexity of the perceptual task are critical and lead directly to 'hard' constraints on the architecture of visual systems, both biological and computational" (Tsotsos, 1990, p. 424).

Tsotsos (1990) examined the computational complexity involved in visual search—that is, searching a visual input for a target item. The biological system seems to compute this type of search on a regular basis and with great ease. Laboratory experiments that examine visual search measure the time participants require to detect targets in a visual display. When a target is defined by conjunction of more than one feature, response time may increase linearly with the number of items in the display. However, when a target is defined by a single feature (such as color), response time may be constant and independent of the number of items in the display; the target item is said to "pop out" immediately (e.g., Treisman & Gelade, 1980; Treisman & Gormican, 1988).

Tsotsos's (1990) complexity analysis showed that a bounded visual search task—where the target is explicitly provided in advance—is computationally linear. An unbounded visual search task—where the target is implicitly expressed (e.g., when it is defined by its relations to other stimuli in the display, such as "the odd one out")—is inherently NP complete, regardless of implementation (for details, see Tsotsos, 1990). In other words, the computational complexity analysis showed that in this case the task is intractable; that is, it cannot be computed in polynomial time because the time requirements are exponential functions of the problem length. In any case, the performance of participants on a variety of visual search tasks presented apparent contradictions between the pattern of performance and the pattern of underlying computational operations needed to perform the task. Thus, the biological system (or any other system) does not conduct visual search, as was previously assumed. This prima facie contradiction focused attention on what visual search is, and forced researchers to rethink what it encompasses. Resolving the contradiction eventually required reshaping what visual search encompasses through the inclusion of a host of approximations and optimizations as well as architectural constraints (for more details, see Tsotsos, 1990).

This example shows that computational theory (in this case computational complexity analysis) can be used in a meaningful and important way to explore a problem. In our example, Tsotsos (1990) showed a mismatch between the complexity of the problem and the computational resources, forcing researchers to rethink and reconceptualize the problem so as to achieve a match between the problem and the resources that solve it. The computational complexity analysis itself is divorced and detached from the biology, but nevertheless has played a role in understanding and deciphering what visual search is and what it encompasses.

Most of the computational analyses discussed in this paper can be categorized as being either biologically plausible or biologically implausible. However, other computational analyses—such as certain computational complexity analyses discussed earlier, and sequential sampling models (see Dror, Busemeyer, & Basola, in press)—are computational tools that examine processes without regard to the biology. This type of computational analysis can be argued to be neither biologically plausible nor biologically implausible. They can be characterized as being "nonplausible" or *aplausible* in reference to biology; that is, they transcend the biological plausibility issue rather than being implausible per se. The important point in the context of our argument is that although these models may be disconnected from the biological facts, they are nevertheless important in cognitive neuroscience research.

Our broad claim is that for certain contributions, even to biologically driven research, the relation of the models to biology (be it contradictory or altogether irrelevant) is not the issue because the type of contribution that the model makes transcends the biological perspective. In the specific section above, we have shown that biologically implausible computational analyses (in one case analyses that contradict biological facts, and in another case analyses that disregard the biological facts) can still play an important role in understanding and characterizing the problem being studied. This type of understanding is an important stage in deciphering the operation of the brain and the cognitive processes it computes.

Examining the Availability of Information and its Representation

Computational analyses can examine the availability of information and its representation. This step is critical

for understanding and exploring any cognitive phenomenon. The biological plausibility of such analyses does not hinder or strengthen the results of these analyses. Again, the mere question of whether such analyses are biologically plausible reflects a basic misconception of the models and their contributions.

Pattern recognition, in its many variations, encompasses a large portion of the research in cognitive neuroscience. As an illustration of our claim, we will demonstrate how computational analyses (regardless of biological plausibility) can contribute to the understanding of the underlying mechanism used by echolocating bats to recognize 3-D objects on the basis of sonar sounds.

Bats use sonar sounds to navigate and capture prey in the dark. The behavior of bats in nature and in laboratory experiments shows that the shape of targets is encoded in sonar sounds (see, e.g., Busnel & Fish, 1980; Griffin, Friend, & Webster, 1965; Simmons et al., 1974). However, sonar sounds have many different dimensions (amplitude changes, frequency composition, cross-correlation between the emitted sound and the returning echo, and so on). It was not clear which dimension carries the information about the shape of targets. Furthermore, based on behavioral and neurophysiological research, a variety of models have been proposed to explain the biological sonar system of bats. The various models assume that relevant information is encoded in different dimensions of the sound. For example, Miller and Pedersen (1988) proposed that the amplitude envelope is used for the perception of complex targets; Schmidt (1992) proposed that a frequency analysis is used; Simmons, Moss, and Ferragamo (1990) proposed that the bat operates as an ideal receiver by performing the neural equivalent of a cross-correlation of the emitted sounds and the returning echoes. Computational analyses can examine what information is available in the different dimensions of sound, and hence provide an important framework for sonar models. Such analyses can further guide and constrain neuroscience research.

Dror, Zagaeski, and Moss (1995) used a biologically implausible computational analysis to examine the availability of information about the shape of 3-D targets in the different dimensions of sonar sounds. In their study, they recorded sonar echoes from 3-D shapes in a variety of orientations. The sonar echoes were given to a number of identical neural networks that were required to learn to recognize the shapes regardless of orientation and to generalize and recognize the shapes in novel orientations. Each network received an identical set of examples in which each sonar sound was digitized as a 240-element input vector. However, the representational formats used for the sonar sounds differed in the various networks. For example, in one network the analog sonar sounds were digitized to process the power spectrum of the sound; in another network, the amplitude waveform was processed and represented; in yet another network, the input was based on a cross-correlation function (for more details, see Dror et al., 1995). By examining the performance of the various networks, one could assess the availability of shape information in the different dimensions of sonar sounds. Specifically, Dror et al. (1995) found that the spectrogram representation is the most efficient in conveying shape information and that the information is redundantly encoded in the lower (25.7–51.5 kHz) and higher (51.5– 95.6 kHz) ultrasonic frequencies (corresponding to the first and second harmonics emitted by the big brown bat, *Eptesicus fuscus*).

This computational analysis assessed the availability of shape information in sonar sounds and has been successfully applied to a variety of pattern recognition tasks (see Dror, Florer, Rios, & Zagaeski, 1996). The fact that such information is explicitly and easily available in one dimension of the sonar sounds and not in others does not prove that the biological sonar system utilizes the computationally efficient dimension. However, the computationally efficient dimension is a good place to start looking at the biosonar of bats, and thus it guides and suggests further research. Moreover, even if the biological system does not use the computationally efficient dimension, the implausible computational analysis still has an important role in studying the biological system. For example, it raises the question of why the biological system did not utilize the dimension of the sound that most easily carries shape information. The computational analysis furthermore demonstrates the type of extra computations that may be involved in processing the less computationally efficient sonar sound dimensions.

The Dror et al. (1995) network was biologically implausible; not only did it not make any explicit or implicit connection to anything we know about the bat physiology, but also the network used a simple feed-forward architecture and used backpropagation as its learning algorithm. Although the network was biologically implausible, it nevertheless was able to contribute to the study of biological sonar systems used by echolocating bats. The computer model, regardless of its biological plausibility, was able to examine the availability of shape information in the different dimensions of sound. How information is encoded and represented, and what information is available, are critical pieces of research data needed for cognitive neuroscience studies. A computational tool that can examine such issues must be an integral part of cognitive neuroscience research, and its ability to perform such analyses does not depend on its biological plausibility.

Evaluating and Understanding the Solution

The cognitive system operates to resolve complex information processing problems. The resolution of these problems is defined by computations that characterize cognitive modules and their interactions. Underlying the cognitive modules are neural transformations that map input to output. We have shown so far that biologically implausible computational analyses can be used to understand and characterize the problem we are studying and to examine what information is available and how it is represented to the system. In this section, we will pre-

sent the role that computational analyses can play in the investigation of the actual neural transformations that underlie cognition (i.e., the solution), even when the computational analyses are biologically implausible.

In a variety of ways, biologically implausible computational analyses are essential for evaluating and understanding the solution to a problem. For example, computational models enable an understanding of the advantages and disadvantages of certain types of information processing schemata. They provide insights into the solution by unmasking the basic tradeoffs that exist within and between solutions for the problem being studied. Furthermore, such computational models establish criteria and parameters for evaluating and examining the actual neuronal solution. In all of these cases, the biologically implausible models are essential and work in conjunction with the biological models. The biologically implausible models complement the plausible models, enabling a better understanding, evaluation, and appreciation of them.

To illustrate our point, let us first look at a classroom example of sorting algorithms (although this example does not reflect a cognitive or brain process, it is a good methodological example that is intended to help convey our theoretical and computational point; later we provide examples from the cognitive neuroscience literature). There are many possible methods for sorting, differing in a variety of computational perspectives, such as their ease of implementation, resource requirements, and number of required operations (Dromey, 1982; Press, Teukolsky, Vetterling, & Flannery, 1992). Because sorting algorithms will be used in environments with varying demands and will be implemented by systems that differ in speed and available resources, some of the algorithms are more appropriate than others in different circumstances. A very brief overview of some of these algorithms will illustrate the importance and necessity of studying a variety of possible solutions.

In bubble sort, adjacent pairs of elements are compared, and if they are out of order their positions are switched. In this way, "heavy elements sink" rapidly to the bottom of the list while "light elements bubble up" to the top. This algorithm requires N^2 comparisons and considerably more operations in order to completely sort any given list (where N is the number of items in the list). A different sorting algorithm, straight insertion, also requires N^2 comparisons but considerably fewer operations to sort any given list. Here, the first element that is found to be out of order is subsequently compared with the remaining items until an insertion point is found. Comparing these two algorithms reveals that, although both require the same number of comparisons, the straight insertion algorithm is more efficient because it requires fewer operations to form the ordered list. Hence, evaluating and understanding a sorting algorithm requires examination of not only the number of necessary comparisons, but also the number of operations needed to form the ordered list.

The Shell diminishing increment sorting method works by *divide and conquer*. In the worst case scenario, this method will require $N^{1.5}$ comparisons, and with a very large list of random items it will require on average $N^{1.25}$ comparisons. Other, even faster, sorting algorithms exist for special cases (e.g., heapsort) and for partitioning the items to be sorted when the range is known in advance (e.g., Hoare's method). An examination of such algorithms reveals a distinction between algorithms that work with the entire list throughout the processing of the list, and algorithms that divide the list into sublists. Furthermore, comparing sorting algorithms reveals that some algorithms require an identical amount of processing regardless of the order of the initial list, whereas the efficiency of other algorithms may depend on the order of the initial list. In such cases, one algorithm may be better than another on average, but the latter may drastically outperform the former when the initial list is given in reverse order.

Using indexing and ranking allows sequential access to a list without sorting the list, but requires the use of another array. Although such mechanisms require more memory, they preserve the initial raw list and are more efficient. Hence, there is a tradeoff between the use of additional memory and the efficiency of the sorting algorithm (for more details on these and other sorting algorithms, see Dromey, 1982, and Press et al., 1992). The point is clear; mechanisms for sorting can be better understood if one has studied a variety of sorting mechanisms. It enables a better understanding by means of comparison and by understanding the different parameters and tradeoffs.

We chose sorting algorithms to illustrate our point because they are relatively simple and widely used. Next, we continue with an example from the cognitive neuroscience literature, showing how biologically implausible computational analyses can contribute to biologically driven research and explain the role of the hippocampus and neocortex in learning and memory (McClelland, McNaughton, & O'Reilly, 1995). Understanding the neural mechanisms that underlie memory and learning requires that we ascertain the relative function and role of the different neural substrates involved in the acquisition and storage of knowledge.

The McClelland et al. (1995) analyses revealed a computational tradeoff between rapid learning and slow consolidation. Using biologically implausible computational analyses, they discovered that rapid learning, which takes place in the hippocampus, enables quick and focused learning of new associations. However, such a mechanism does not enable the memory system to integrate new knowledge into existing knowledge structures. Furthermore, rapid learning may result in *catastrophic interferences* (McCloskey & Cohen, 1989), whereby new knowledge interferes with what is already known. Conversely, their biologically implausible computational analyses showed that *interleaved learning*, which takes place in the neocortex, enables the memory system to consolidate new

knowledge into existing knowledge structures without interfering with what is already in memory. However, such interleaved learning requires that learning does not occur all at once, but that new knowledge be integrated gradually over time. In this case, new knowledge is acquired by slowly consolidating it into existing prior knowledge.

The McClelland et al. (1995) biologically implausible analyses provided computational evidence that the hippocampus and neocortex are not redundant dual storage mechanisms (where memories are initially stored in duplicate), but rather they are complementary systems. Initially, knowledge is stored in the hippocampus, enabling rapid learning and storage of new associations. Then, over time, knowledge is transferred and integrated into the system in the neocortex, enabling the discovery of regularities and incorporation of new knowledge into existing knowledge while avoiding the drawbacks of rapid learning, such as possible catastrophic interferences (for more details, see McClelland et al., 1995). McClelland et al. used a number of computational analyses to demonstrate this tradeoff and to show that their conclusion was not dependent on any specific technique or approach. Hence, they claimed not only that the critical computational tradeoff can be revealed through the use of a specific biological implausible model (like backpropagation), but also that a wide variety of computational approaches can be informative in such computational investigations (regardless of their biological plausibility).

The point is that such insights into the neuronal system that underlies learning and memory can be derived from biologically implausible computational analyses. The biologically implausible computational analyses conducted by McClelland et al. (1995) demonstrated the fundamental computational tradeoff between learning rapidly and remembering, and what each one entails, and how (as well as why) the hippocampus and neocortex are complementary learning systems.

The theoretically important advances that McClelland et al. (1995) provided to the understanding of memory and learning within cognitive neuroscience were made possible by the use of biologically implausible computational models. Thus, biologically driven research has benefited from biologically implausible computational analyses (and the details of how much and in what sense the model was implausible are not at all relevant to its contributions, since the computational tradeoff it demonstrated does not depend on biological fact; it transcends the biology perspective altogether).

Another research example where biologically implausible modeling has been used in understanding the solution of the biological system is in echolocating dolphins. Dolphins use underwater sonar as a full perceptual modality (Au, 1993). Although they emit multiple successive sonar clicks, it was not clear how they processed and used multiple echoes to identify a single target. Furthermore, there was a question regarding the possible computational significance of processing multiple echoes instead of using only a single echo. Moore, Roitblat, Penner, and Nachtigall (1991) were able to use computational analyses to assess the advantage of using multiple echoes in terms of efficiency (in both time and sampling needs). Furthermore, they were able to examine the computational mechanisms and architectures that may be needed to process multiple successive echoes (e.g., an integrator device that can accumulate information across multiple inputs, an apparatus that resets the integrator at the end of an input sequence, and so on). Their work entailed computational analyses that characterized the differences between two processing approaches. The computational insights from their research are important to understanding the biological mechanisms used by dolphins to perceive their environment.

The use of biologically implausible models is also important in understanding and evaluating the solutions of the biological system because it enables researchers to find and ask the appropriate questions. Biologically implausible computational analyses can determine which is the most efficient model (the simplest model that can account for the data, defined in terms of the model's fit to the data, the model's parameters, and the functional form of the model; for details, see A. M. Jacobs & Grainger, 1994; Myung & Pitt, 1997). This is an important research tool because it enables researchers to characterize the efficiency of the biological solution; if it diverges from a more efficient processing approach, then one can ask and investigate the reason why the biological system implemented the mechanism that it did. In this role, the biologically implausible analyses can guide and focus cognitive neuroscience research. Thus, biologically implausible investigations not only provide a basis for comparison and evaluation, but also raise important questions regarding neuronal function.

SUMMARY AND CONCLUSIONS

Computational analyses play a central role in cognitive neuroscience. Because cognitive neuroscience aims at understanding the neuronal system and its cognitive manifestation, biological plausibility has been taken for granted as an a priori criterion for the acceptability of computational models. In this paper, we question whether biologically driven research cannot greatly benefit from biologically implausible computational analyses. We derive an important and essential role that biologically implausible computational investigations have in cognitive neuroscience.

We argue that computational analyses are important in understanding and characterizing the problem that is being studied, in examining the availability of information and its representation, and in evaluating and understanding the neuronal solution. The critical point we make is that, in these contexts, the biological plausibility of such computational analyses is irrelevant to their contributions. We also make the case that biologically implausible models are necessary for learning and better understanding the computations of the biological mechanisms.

The contributions of the biologically implausible models discussed in this paper are not intended to cover all the possible contributions that such models can make to cognitive neuroscience. Rather, the paper is aimed at arguing that such contributions are possible and are needed. Furthermore, the examples in this paper illustrate that such contributions are not limited to one aspect of studying cognitive neuroscience, but rather, they can be applicable to almost every aspect and phase of research. In a similar way, they illustrate that such contributions are not limited to one type of computational model or to a certain cognitive domain.

Not all biologically implausible models are valuable (likewise, neither are all biologically plausible models valuable), and their possible contribution does not depend on how far they are from biology. Although the value of biologically plausible models may well depend on how close they are to the biological system, in the context of the distinct types of contributions made by the biologically implausible models, the biological dimension is altogether irrelevant.

Since cognitive neuroscience researchers are in the business of *reverse engineering* (Dennett, 1994), it is hard to always know when biologically implausible (or biologically plausible) computational analysis are more likely to be helpful. As we have shown throughout this paper, biologically implausible analyses can be helpful in making certain types of contributions. Furthermore, biologically implausible analyses are more likely to be helpful when assumptions and constraints that are present solely to account for biological facts are ignored. Conversely, biologically plausible models are likely to be helpful when specific biological underpinnings have cognitive implications. However, how can one know whether certain biological facts are relevant to the cognitive process or theory? The answer to this question may bring us back to biologically implausible models that can determine the possible significance, or lack thereof, of certain biological facts for the cognitive level. For example, Dror (1997) used such analyses to examine whether age-related cognitive changes arise from computational adaptations induced by specific biological mechanisms (for instance, despecialization of neural substrates to enable them to compute a wider variety of cognitive processes), or from higher level strategy changes that are dissociated from the biological details (for instance, more on-line processing to decrease the demands on memory). Another example would be the examination of whether age-related growth of dendrites has cognitive significance or whether it is a biological mechanism that has nothing to do with cognitive processing (Dror & Morgret, 1996). Using a computationally implausible analysis, Dror and Morgret were able to assess if, and how much, the decrease of processing power within computational units can be compensated by adding communication between processing units. The conclusion derived here is the same conclusion that we have been presenting throughout the entire paper: Understanding and evaluating the biological system and the cognitive processes it computes requires biologically implausible computational analyses.

REFERENCES

AIZAWA, K. (1992). Biology and sufficiency in connectionist theory. In J. Dinsmore (Ed.), *The symbolic and connectionist paradigms* (pp. 69-88). Hillsdale, NJ: Erlbaum.

AU, W. W. C. (1993). *The sonar of dolphins*. New York: Springer-Verlag.

BOWER, J. M. (1990). Reverse engineering the nervous system: An anatomical, physiological, and computer-based approach. In S. F. Zornetzer, J. L. Davis, & C. Lau (Eds.), *An introduction to neural and electronic networks* (pp. 3-24). San Diego: Academic Press.

BUGMANN, G. (1997). Biologically plausible neural computation. *Biosystems*, **40**, 11-19.

BUSNEL, R. G., & FISH, J. F. (1980). *Animal sonar systems*. New York: Plenum.

CHOMSKY, N. (1956). Three models for the description of language. *Transactions on Information Theory*, **2**, 113-124.

CHOMSKY, N. (1962). *Context-free grammars and pushdown storage*. Cambridge, MA: MIT Press.

CRICK, F. (1989). The recent excitement about neural networks. *Nature*, **337**, 129-132.

CRICK, F., & ASANUMA, C. (1986). Certain aspects of the anatomy and physiology of the cerebral cortex. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing* (Vol. 2, pp. 219-232). Cambridge, MA: MIT Press.

DENNETT, D. C. (1994). Cognitive science as reverse engineering: Several meanings of "top down" and "bottom up." In D. Prawitz, B. Skyrms, & D. Westerstahl (Eds.), *Proceedings of the Ninth International Congress of Logic, Methodology, and Philosophy of Science*. Amsterdam: North-Holland.

DROMEY, R. G. (1982). *How to solve it by computer*. London: Prentice-Hall.

DROR, I. E. (1994). Neural network models as tools for understanding high-level cognition: Developing paradigms for cognitive interpretation of neural network models. In M. C. Mozer, P. Smolensky, D. S. Touretzky, J. L. Elman, & A. S. Weigend (Eds.), *Proceedings of the 1993 Connectionist Models Summer School* (pp. 87-94). Hillsdale, NJ: Erlbaum.

DROR, I. E. (1997). Computational adaptations and cognitive strategy changes as compensation for age-related decline in cognitive resources. *Society for Neuroscience Abstracts*, **23**, 1457.

DROR, I. E., BUSEMEYER, J. R., & BASOLA, B. (in press). Decision making under time pressure: An independent test of sequential sampling models. *Memory & Cognition*.

DROR, I. E., FLORER, F. L., RIOS, D., & ZAGAESKI, M. (1996). Using artificial bat sonar neural network for complex pattern recognition: Recognizing faces and the speed of a moving target. *Biological Cybernetics*, **74**, 331-338.

DROR, I. E., & MORGRET, C. C. (1996). A computational investigation of dendritic growth as a compensatory mechanism for neuronal loss in the aging brain. *Society for Neuroscience Abstracts*, **22**, 1891.

DROR, I. E., ZAGAESKI, M., & MOSS, C. F. (1995). Three-dimensional target recognition via sonar: A neural network model. *Neural Networks*, **8**, 149-160.

GLUCK, M. A., & THOMPSON, R. F. (1987). Modeling the neural substrates of associative learning and memory: A computational approach. *Psychological Review*, **94**, 176-191.

GRIFFIN, D. R., FRIEND, J. H., & WEBSTER, F. A. (1965). Target discrimination by the echolocation of bats. *Journal of Experimental Zoology*, **158**, 155-168.

GROSSBERG, S. (1987). Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science*, **11**, 23-63.

JACOBS, A. M., & GRAINGER, J. (1994). Models of visual word recognition: Sampling the state of the art. *Journal of Experimental Psychology: Human Perception & Performance*, **20**, 1311-1334.

JACOBS, R. A., JORDAN, M. I., & BARTO, A. G. (1991). Task decomposi-

tion through competition in a modular connectionist architecture: The what and where vision tasks. *Cognitive Science*, **15**, 219-250.

JORDAN, M. I., FLASH, T., & ARNON, Y. (1994). A model of the learning of arm trajectories from spatial deviations. *Journal of Cognitive Neuroscience*, **6**, 359-376.

KETTNER, R., MARCARIO, J., & PORT, N. (1993). A neural network model of cortical activity during reaching. *Journal of Cognitive Neuroscience*, **5**, 14-33.

KOSSLYN, S. M., CHABRIS, C. F., MARSOLEK, C. M., & KOENIG, O. (1992). Categorical versus coordinate spatial representations: Computational analyses and computer simulations. *Journal of Experimental Psychology: Human Perception & Performance*, **18**, 562-577.

KOSSLYN, S. M., & KOENIG, O. (1992). *Wet mind*. New York: Free Press.

LEHKY, S. R., & SEJNOWSKI, T. J. (1988). Network model of shape-from-shading: Neural function arises from both receptive and projective fields. *Nature*, **333**, 452-454.

LEVIN, D. S. (1991). *Introduction to neural and cognitive modeling*. Hillsdale, NJ: Erlbaum.

LEWIS, H. R., & PAPADIMITRIOU, C. H. (1981). *Elements of the theory of computation*. Englewood Cliffs, NJ: Prentice-Hall.

MARR, D. (1982). *Vision*. San Francisco: Freeman.

MAZZONI, P., ANDERSEN, R. A., & JORDAN, M. I. (1991). A more biologically plausible learning rule than backpropagation applied to a network model of cortical area 7a. *Cerebral Cortex*, **1**, 293-307.

MCCLELLAND, J. L., MCNAUGHTON, B. L., & O'REILLY, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, **102**, 419-457.

MCCLOSKEY, M., & COHEN, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 24, pp. 109-165). New York: Academic Press.

MEAD, C. (1989). *Analog VLSI and neural systems*. Reading, MA: Addison-Wesley.

MILLER, L. A., & PEDERSEN, S. B. (1988). Echoes from insects processed using time delayed spectrometry (TDS). In P. E. Nachtigall & P. W. B. Moore (Eds.), *Animal sonar, processes and performance*. New York: Plenum.

MOORE, P. W. B., ROITBLAT, H. L., PENNER, R. H., & NACHTIGALL, P. E. (1991). Recognizing successive dolphin echoes with an integrator gateway network. *Journal of Neural Networks*, **4**, 701-709.

MOYNE, J. A. (1985). *Understanding language: Man or machine*. New York: Plenum.

MURRE, J. M. J. (1992). *Learning and categorization in modular neural networks*. Hillsdale, NJ: Erlbaum.

MYUNG, I. J., & PITT, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, **4**, 79-95.

O'REILLY, R. C. (1996). Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Computation*, **8**, 895-938.

PRESS, W. H., TEUKOLSKY, S. A., VETTERLING, W. T., & FLANNERY, B. P. (1992). *Numerical recipes in C*. Cambridge: Cambridge University Press.

RUECKL, J. G., CAVE, K. R., & KOSSLYN, S. M. (1989). Why are "what" and "where" processed by separate cortical visual systems? A computational investigation. *Journal of Cognitive Neuroscience*, **2**, 171-186.

RUMELHART, D. E. (1990). Brain style computation: Learning and generalization. In S. F. Zornetzer, J. L. Davis, & C. Lau (Eds.), *An introduction to neural and electric networks* (pp. 405-420). San Diego: Academic Press.

RYLE, G. (1949). *The concept of mind*. Hutchinson.

SCHMAJUK, N. A., & DICARLO, J. J. (1992). Stimulus configuration, classic conditioning, and hippocampal function. *Psychology Review*, **99**, 268-305.

SCHMIDT, S. (1992). Perception of structured phantom targets in the echolocating bat, *Megaderma lyra*. *Journal of the Acoustical Society of America*, **91**, 2203-2223.

SEJNOWSKI, T. J. (1986). Open question about computation in cerebral cortex. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing* (Vol. 2, pp. 372-389). Cambridge, MA: MIT Press.

SIMMONS, J. A., LAVENDER, W. A., LAVENDER, B. A., DOROSHOW, C. A., KEIFER, S. W., LIVINGSTON, R., SCALLET, A. C., & CROWLEY, D. E. (1974). Target structure and echo spectral discrimination by echolocating bats. *Science*, **186**, 1130-1132.

SIMMONS, J. A., MOSS, C. F., & FERRAGAMO, M. (1990). Convergence of temporal and spectral information into acoustic images of complex sonar targets perceived by the echolocating bat, *Eptesicus fuscus*. *Journal of Comparative Physiology A*, **166**, 449-470.

THOMPSON, R. F., & GLUCK, M. A. (1990). A biological neural network analysis of learning and memory. In S. F. Zornetzer, J. L. Davis, & C. Lau (Eds.), *An introduction to neural and electronic networks* (pp. 91-107). San Diego: Academic Press.

TREISMAN, A., & GELADE, G. (1980). A feature-integration theory of attention. *Cognitive Science*, **12**, 99-136.

TREISMAN, A., & GORMICAN, S. (1988). Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review*, **95**, 15-48.

TSOTSOS, J. K. (1990). Analyzing vision at the complexity level. *Behavioral & Brain Science*, **13**, 423-444.

TSOTSOS, J. K. (1991). Is complexity theory appropriate for analyzing biological systems? *Behavioral & Brain Science*, **14**, 770-773.

ZIPSER, D., & ANDERSEN, R. A. (1988). A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature*, **331**, 679-684.

NOTES

1. We use the term *computational analyses* in its broader meaning. It refers to conclusions and analyses based on various types of models and computer simulations. The term *biologically implausible computational analyses* is used in this paper to reflect any type of model or simulation, and analyses of them, in which some aspect is inconsistent with the known neural underpinnings (see note 2). Throughout this paper, we use the terms *computational analyses*, *computational models*, and *computational investigations* interchangeably.

2. Biological plausibility (or biological implausibility) depends on whether the computational analysis is consistent with, and conforms to, the known biology (or whether it contradicts it). This type of judgment may be highly dependent on the level of resolution at which one examines the biology. Hence, at some very fine level of resolution, any analysis would be inconsistent with some aspect of the biology, and therefore be biologically implausible. We defend even the most extreme notion of biological implausibly—that is, cases where the computational analyses have no connection to biology whatsoever, cases that go beyond the biological plausibility continuum (as we discuss later in the paper).

3. Marr (1982) did not mention computational complexity analyses as part of the discussion of computational mechanisms at the representation and algorithm level. Tsotsos (1990) noted that this exclusion is surprising (p.424); however, we speculate that Marr's exclusion of computational complexity analyses is not an oversight or surprising. Rather, its exclusion is derived from Marr's assumption of biological plausibility.

(Manuscript received January 13, 1997;
revision accepted for publication June 5, 1998.)