CHAPTER 11

MODEL CASCADES: HOW AND WHY

Computational modeling of the brain holds great promise as a bridge from brain to behavior. To fulfill this promise, however, it is not enough for models to be 'biologically plausible': models must be structurally accurate. Such structural accuracy may be supported by a model's fit of existing data, the derivation of new predictions, and the use of supported assumptions. All three sources of support require modelers to be explicit about the ontology of the model, and require the existence of data constraining the modeling. If such data is only sparsely available, a new approach to modeling might bring reprieve. If several models are constructed that together form a cascade of models constraining data can be maximally used and higher-level models can be constrained by lower-level models. Modeling the same substrate at different levels of representation, as has been practiced in this dissertation, thus has benefits that exceed the merits of each model in the cascade on it own.

11.1 INTRODUCTION

Most psychologists and neuroscientists agree that the brain produces behavior, and that ultimate theories of behavior will be ones that spell out the link between the two. In practice, however, the disciplines of psychology and neuroscience are still largely separate enterprises, the one focusing on behavior, the other on brain functioning. It has proven difficult to link specific behavioral findings to brain mechanisms. In vivo cell recordings and functional imaging do hold the promise of linking behavior to the underlying brain activity. Such techniques have indeed elucidated which brain areas are involved in certain behavioral functions (Cabeza & Nyberg, 2000; O'Keefe & Recce, 1993), but the way in which these brain areas generate behavior remains shrouded in mist.

One type of research that has often been promoted as a bridge between brain functioning and behavior is computational modeling of the brain. Ideally, a model incorporating knowledge about brain anatomy and physiology can show how certain tasks are solved by a human or animal. Such models can then generate predictions on both the physiological level and the behavioral level. In practice, however, a majority of computational models still focus on one or the other, modeling either brain anatomy and physiology in great detail, or behavior. In the first class of models, behavior is often modeled, if at all, in a very abstract way, precluding the possibility of generating behaviorally testable predictions. In the second class of models, the connection to real brain processes is often so thin as to become irrelevant. For example, model neurons in many connectionist frameworks have little to do with real neurons.

This chapter is based on a draft of a paper co-authored by Janneke Jehee, Jaap Murre and me (Meeter, Jehee, & Murre, in prep.).
This state of affairs is not very surprising. A 'mindbrain' model, one that is both adequate at the biological plane and specific about behavior, must work at levels so far apart that they seem unconnected, differing widely on the temporal and spatial scale on which events take place. It must somehow do justice to neurons which take up on average about $8 \times 10^{-15}$ ml. and show spikes that last less than a millisecond, but also to behavioral tasks which involve the whole brain (around 1300 milliliters in the human) and take seconds to minutes or more (Anderson, 2002; Jehee & Murre, subm.).

In recent years, however, more and more models have been proposed that are intermediate, being neither purely functional nor tied with much precision to brain anatomy (Bogacz, Brown, & Giraud-Carrier, 2001; Gluck & Myers, 1993; Hasselmo et al., 2002; Jensen & Lisman, 1996; Murre, 1996; Norman & O'Reilly, in press; Raffone & Wolters, 2002; Rokers, Mercado, Allen, Myers, & Gluck, 2002; Sohal & Hasselmo, 1998; van der Velde & de Kamps, 2001; chapters 5, 6, 9, 10 this dissertation). They typically take inspiration from brain anatomy to model a particular set of behavioral data, or a particular psychological function. For such models, I will use the term psychobiological models. Some of these models have generated much enthusiasm, as they hold great promise as a bridge from the brain to behavior.

Nevertheless, it is often unclear how exactly psychobiological models should be appraised. This is easier for models that are purely psychological (which may be evaluated on their ability to explain and predict behavior) or purely biophysical (which may be evaluated on their faithful representation of brain anatomy, and on their ability to explain and predict brain physiology). As models of behavior, psychobiological models are most often not up to standards set by formal models. For example, none of the biologically-inspired network models of memory yields quantitative predictions on memory experiments, something that was already achieved by functional, formal models developed in the seventies and early eighties (e.g., Raaijmakers & Shiffrin, 1981). Psychobiological models also implement brain features only coarsely, and they seldom are precise enough to generate predictions on neurophysiology. According to traditional criteria, psychobiological models are thus not particularly successful.

Such deficiencies are often said to be compensated by a high 'biological plausibility'. However, this is more often claimed than substantiated (Jacobs & Grainger, 1994). For many psychobiological models substantiating biological plausibility would also be difficult, as they operate in a void in which few data is available to constrain the modeling. Even in cases where it is more than just a claim, how biological plausibility should be weighed is not clear: does it make up for fitting a very limited set of phenomena and for fitting only at a qualitative level?

In this chapter, I will discuss how psychobiological models can be evaluated, taking into account their particular position in between the brain and the mind. In a way, the chapter can be read as an analysis of the claim of biological plausibility. I will argue that there are in general three ways in which models can garner support. One is by matching existing data, a second way is by leading to predictions, the third is by being a priori plausible. All three ways require that modelers be specific on the ontology of their model, and they all require that there exists data constraining the model. In many cases, little constraining data may be available. As a solution to this problem, a new approach to modeling will be proposed: that of constructing a model cascade. Such a cascade of complementary models (a hierarchy of models) may optimize the way data is incorporated into the model, and could bridge the gap between the seemingly unconnected levels of biology and psychology. Models in the cascade can also constrain
one-another, and pose restrictions on modeling where behavior and physiology pose too few.

The next sections will be concerned with what a computational model of the brain is, and how such a model can be appraised. Thereafter, I will present strategies for good modeling, with special emphasis on the solution expounded here: model cascades. Several examples of such cascades will briefly be reviewed. One such example is, of course, the relations between the models presented in this dissertation.

11.2 STRUCTURE OF A MODEL

Computational modeling of the brain is the attempt to build a structure that is in some way an abstraction of the brain or a part of it. Characteristics of the functioning of this abstract structure can then be derived or established through simulation. Usually, the model produces model behavior that can be compared to empirical data. The abstract structure may be thought of as a collection of assumptions that together specify the model. In an analytical, mathematical model, these assumptions are the formulas used, the structure of the data that goes into the formulas, and parameter values. In neural networks, they may be the rules governing the behavior of the nodes, the subdivisions of the network, the connection schemes incorporated in the network, the number of nodes and other parameter values (in fact, all neural networks can be written as systems of coupled equations, making them a subcategory of analytic models).

Some assumptions going into the model may be supported by evidence from the brain and behavioral sciences. I will call these supported assumptions. As an example, many recent models of the hippocampus and its role in memory have faithfully reproduced hippocampal subdivisions, and the pathways between those subdivisions (Hasselmo et al., 2002; Jensen & Lisman, 1996; Norman & O'Reilly, in press; chapter 10 this dissertation). Assumptions describing these aspects of the networks can thus be said to be supported by our knowledge of brain anatomy. Assumptions may also be unsupported ideas about brain or behavior. These assumptions, which could possibly be true of the brain, I will refer to as untested assumptions. Some of these may be ones that are central to the new explanations offered by the model. For these assumptions, which reflect the bold new theorizing of the modeler, I will reserve the name hypotheses.

To build a working model, it is usually not enough to only abstract away from the brain, however (if all mismatch with biology would refute a model, 99% of all models would have to be rejected). Assumptions have to be added to the formalism that do not in themselves model anything in the brain and are not meant to be hypotheses about the brain, but that allow the model to produce behavior that can be compared with data. An example is error back-propagation. Only backprop diehards would say it is biologically plausible, but it features in many models to support learning in the network (Gluck & Myers, 1993; van der Velde & de Kamps, 2001). Other examples are hard winner-take-all dynamics, orthogonal input patterns, linear weight increases, clamping of input patterns, empty brains at the outset of the simulation, etc. Such assumptions, which are in all likelihood counterfactual, I will refer to as heuristic assumptions (similar use of the

1 Error backpropagation is a heuristic assumption in models of the brain. It may be a hypothesis in a functional model, if the modeler assumes that the backprop learning rule captures the functional characteristics of learning in the modeled system.
term is found in economics, where false assumptions of rationality and full information are defended as innocent heuristic devices).

As already said, parameter values can also be viewed as assumptions. In the case of a free parameter, its value may be said to be an untested assumption (it could, in principle, be true of reality that the parameter or some counterpart has the chosen value). If, on the other hand, the value is taken from data not modeled, then the value may be said to be a supported assumption.

11.3 MODELING GOALS

Computational modeling can have several goals. A computational model can be presented as only one of many possible instantiations of a verbal theory. The goal of the modeler is then to show that the model produces the theorized behavior, as an existence proof that the mechanisms in the verbal theory can work as proposed (e.g., Nadel et al., 2000). Sometimes, models are said to uncover computational limitations and tradeoffs, making it possible to derive general principles that are applicable to the brain's computations (McClelland et al., 1995), or they make statements about the computations that a certain brain area can perform (Treves & Rolls, 1994). A goal of modeling may also be to structure data and tease apart components that underlie superficial effects in the data (e.g., modeling retention to uncover differences in forgetting rates; Meeter, Murre et al., subm.; Rubin & Wenzel, 1996). When used in this way, the output of the model is input into other theorizing. Most often, however, the aim of the modeler is to explain existing findings and predict new ones (Jacobs & Grainger, 1994).

There are two ways in which ‘explaining of findings’ can be understood. A classic interpretation of theorizing in general is that a theory summarizes and systematizes observation, allowing covering laws to be extracted from the data. The theory – or computational model– would then be nothing more than a calculus with operationalizations (Hempel, 1965), from which predictions can be derived. This is in all likelihood not the view of most modelers. Most modelers, as most scientists in general, probably have the pretension that their model corresponds to something in reality. They will claim for their model what Webb (2001) has called structural accuracy. Structural accuracy refers to how well the model represents the real mechanisms underlying the target behavior. If a model is structurally accurate, it does not reproduce the target data only because it incorporates a covering law, it does so because the mechanism producing model behavior is in some ways equivalent to that producing the data in the modeled substrate. The second interpretation of a model explaining a finding, the one that most modelers would accept, is thus for the model reproducing the data, and doing so in a structurally adequate way.

Structural accuracy is a predicate like truth in empirical sciences: one can never be 100% certain that a model is structurally accurate, but the structural accuracy of a model can be supported. For a computational model, such support can take three forms. First, a model may be said to be supported by its ability to reproduce existing findings. Second, predictions may be derived from the model. If these are subsequently proven to be correct, the model is strongly supported (if, at least, the prediction was risky; Popper, 1934; Roberts & Pashler, 2000). Thirdly, structural accuracy may be made a priori plausible by showing that the model is built on many supported assumptions, and few untested or heuristic ones.
These three ways of supporting a model are not specific to computational work. Computational modeling is just a form of theorizing, and the sources of support for models are those of theories in general: verbal theories are also evaluated on how they explain data, on the fate of their predictions, and on how plausible they are. Presenting a theory in the form of a computational model has several advantages over verbal theorizing, however. It makes hidden assumptions explicit, and makes it easier to detect fudges, ad hoc assumptions and inconsistencies between the assumptions underlying the model. Moreover, when a model is applied in new ways it is still relatively clear what a computational model predicts (although see later). In verbal theorizing, what counts as a prediction of the model is sometimes more hotly debated than the predictions themselves. Even if the complexity of models sometimes inspires uncritical acceptance or just as uncritical rejection, computational models thus are in principle easier to evaluate objectively than verbal theories.

**A priori plausibility**

Outsiders would rather believe a model with assumptions that they know to be true, than one with assumptions that they have to accept without backup. Having assumptions supported by empirical evidence thus makes a model a priori plausible. Adding biological features in itself, however, does not automatically make a model more likely to be structurally adequate. If certain features are unimportant for the modeled behavior, to abstract away from them does not threaten structural accuracy. It is not the presence of biological features per se that renders a model plausible, but the presence of the relevant ones as well as the absence of assumptions not in agreement with biology. Especially the presence of unsupported assumptions detracts from a model’s a priori plausibility. This can be seen from an analysis of the process of model behavior derivation. In a technical sense, producing model behavior to fit data is similar to deriving a theorem from a set of postulates. In the case of a computational model, the assumptions that a modeler would like to see as the input of the derivation are the supported assumptions, and what has been called the hypotheses (the untested assumptions that are central to novel explanations offered by the model). By what is known as the Duhem-Quine thesis, however, a single hypothesis can never be tested in isolation; it is always tested together with the unavoidable background of theory and auxiliary hypotheses: both are necessary for the derivation of a concrete prediction. In the case of the model, these include the other untested assumptions and the heuristic assumptions.

For the derived behavior to be a valid indicator of the structural accuracy of the model, the model behavior has to be caused by features that the model shares with the modeled brain tissue (i.e., supported assumptions and possibly hypotheses). If, in fact, heuristic assumptions play an important role in model behavior, then the model is probably not structurally accurate (its behavior is not produced with the same mechanisms as those in the brain). Any correspondence of model behavior and real data can then only be a coincidence. Deriving predictions from model behavior thus involves what one could call a meta-assumption of no commission: that the heuristic assumptions made are immaterial for the model behavior used as the prediction. If this meta-assumption is true, many sets of heuristic assumptions would, if combined with the supported assumptions and hypotheses, lead to the same model behavior.

There are, of course, not only sins of commission, but also of omission. Abstracting away from biological features may also ruin structural accurateness, if the elements left out in the abstraction turn out to be causally relevant for the modeled data. If the model
reproduces target data, however, it can be concluded that the model contains at least enough features to produce that data. This is not to say that all relevant assumptions are true: the model may produce the behavior because heuristic assumptions play an important role. If a model fits data, sins of omission can thus only be present in the presence of sins of commission: if the right feature is not in the model, it takes a wrong feature to make up for it and still produce the right behavior. There is thus an asymmetry, where sins of commission have worse consequences for structural accuracy than sins of omission. A priori plausibility therefore hinges strongly on the absence of unsupported features, not the presence of supported ones.

In an extreme case, there are no sins of omissions or commissions at all. This would make the model’s behavior plausible as a prediction independent of any fit of real data. This is no fata morgana. In fact, all models that purport to show computational constraints valid for the brain must follow the above reasoning. If, for example, the memory capacity of the hippocampus can validly be inferred from the model of Treves and Rolls (1994), it must be the case that the model is wholly structurally accurate. The validity of constraints derived by such models crucially depends on the meta-assumption of no commission described above, that no heuristic assumption is critical for the behavior. If that meta-assumption is not true, the derived constraint may only apply to the model, not to the brain.\footnote{McClelland et al. (1995) support their claim that certain memory functions cannot be subserved by a single memory store by a simulation showing catastrophic interference in a PDP network. The validity of this argument rests on the questionable assumption that catastrophic interference is a general phenomenon in all neural networks (French, 1999; Page, 2000).}

The constraint is also invalidated when features abstracted away from are important for the behavior (unlike with mathematical axioms, added assumptions are not necessarily restrictions on the range of behaviors of a computational model).

Models used to infer computational constraints are most often highly abstract, analytical models that only yield very abstract behavior. If a modeler wants to derive or simulate more complex, more ‘interesting’ behavior, adding untested and heuristic assumptions to the model is unavoidable. These additions lessen the a priori plausibility of the model, but this loss may be compensated by more support from the other two sources of model support.

**Matching data**

Reproduction by the model of empirical data is a traditional standard for evaluating a model. It is referred to as the extent to which the behavior of the model ‘matches’ the to-be modeled data sets (Webb, 2001), or how well it ‘fits’ the data, or is ‘descriptively adequate’ (Jacobs & Grainger, 1994). Data match may be evaluated quantitatively (e.g., by the model’s ability to reproduce data curves), or only at a qualitative level (e.g., by comparing the direction of effects of independent variables in the data with model behavior). Psychobiological models often fit data only at a qualitative, not quantitative level.

How well a model fits data is the domain of a vast technical literature, which I will not attempt to summarize here (e.g., Pitt, Myung, & Zhang, 2002; Zucchini, 2000). One important aspect to mention, however, is that match can only be evaluated while taking into account the flexibility of the model (Roberts & Pashler, 2000) and the freedom of the modeler. Flexibility refers to the amount of possible data patterns that a model, given its
structure, would be able to fit. Freedom refers to the lack of constraints for a modeler in building the structure of his or her model.

Flexible models can fit any data pattern. When real data is fitted by a flexible model, this does not say much as any other data pattern would just as easily have been fitted. Data fit is only impressive in an inflexible model (another way of saying the same is that only falsifiable models are worth testing; (Popper, 1934). Sophisticated techniques are available to assess data fit against the flexibility of the model (Pitt et al., 2002). With analytic models, flexibility can to some extent be estimated as the number of free parameters – though this is not enough (Pitt et al., 2002). An exhaustive parameter search, which shows the boundaries of what the model can fit, can also be used to investigate flexibility (Roberts & Pashler, 2000). If these boundaries are close, the model can be said to have little flexibility. This is seldom undertaken, however, and would be an endless affair in many simulation models. In such models, flexibility is often hard to gauge; with neural network models, for example, it is usually not clear what counts as a free parameter and what not, and which design features offer extra freedom and which not.

Moreover, extra flexibility is often hidden in the translation of model behavior to target behavior. An example is the treatment of time in models. Processes take time, and much data has either the form of a time duration (e.g., reaction times, interspike intervals) or are set in time (a learning curve, or the order of activation of different processing regions). In a computational model, a process may take up a number of cycles or events, but these are often not tied to a duration in real time. The correspondence between time in the model and real time is thus usually a free parameter. In purely functional models, the time taken up by events in the model is therefore often explicitly treated as a parameter. In ACT-R, for example, one retrieval step takes up a given amount of time (usually set to 50 ms; Anderson, 2002). In biophysical models, the time taken up by events can be derived from neurophysiological data (Hodgkin & Huxley, 1952). It is the intermediate, psychobiological models, where the treatment of time is most problematic. In models of long-term memory consolidation (Alvarez & Squire, 1994; McClelland et al., 1995; Murre, 1996), for example, one consolidation cycle may represent one night of sleep, but it may also be just a minute of dreaming.

Even in an inflexible model, match is not always convincing. If a very complex model fits a few qualitative data patterns, many observers will be underwhelmed. Such models invite the suspicion that they are custom-built to reproduce the target data, and will do nothing more than just that. Successful reproduction of data is in such cases not very convincing support for the model. This is the problem behind the gripe of many experimentalists that modelers tinker with their model until the desired behavior is obtained, with the model losing all relevance to psychology and neuroscience in the process.

To evaluate match, it is thus important to not only take into account the flexibility of the model, but also the freedom of the modeler. One constraint on that freedom is the demands for simplicity and parsimony: if a modeler can only work with few assumptions (if the model is sparse), he or she does not have much freedom to fit data. Supported assumptions also do not count in an assessment of freedom: if biology prescribes the anatomy of the model, it cannot be designed to increase the match. In models with few untested and heuristic assumptions, even qualitatively reproduction of data patterns can therefore be an impressive match.

Models can also be made less flexible with biological data, by basing as many parameter values as possible on data that is not fitted. Many parameter values in Hodgkin-Huxley
formalism models can, for example, be extracted from experimental data (Hodgkin & Huxley, 1952). Nevertheless, because of their ill-defined freedom and flexibility, judging the match between psychobiological models and data is often difficult. This match can only be assessed by eyeballing the balance between model simplicity (a proxy for flexibility and freedom) and its ability to reproduce data patterns. Needless to say that this is not a very reliable procedure.

**Deriving predictions**

Most psychobiological models reproduce behavior, but either at only a qualitative level, or with so much flexibility or freedom that the fit is not compelling support for the model. The model can nevertheless be very worthwhile, for example as a new way of seeing things. Like verbal theories (which also do not produce inflexible matches of behavior), such models are usually assessed on the basis of the predictions they make. The strongest support any model or theory can receive is that daring predictions it made are subsequently proven to be correct (Roberts & Pashler, 2000).

The word ‘prediction’ has been subject to some inflation: many modelers will call any model behavior a prediction, even when the data fitted has already been around for a long time (Roberts & Pashler, 2000). Here, I mean with prediction something derived from the model behavior of which the modeler did not know whether it was the case. More suspicious minds may speak of predictions only if nobody knew, at the time that the model was presented, whether the prediction was true or not.

Computational models of the brain are seldom presented only after daring predictions have been proven to be correct (an exception is Rokers et al., 2002). Many papers presenting a new model, however, close with a long list of predictions. These are then left open for experimentalists to confirm or refute (e.g., Bogacz et al., 2001; Raffone & Wolters, 2002; Sohal & Hasselmo, 1998; chapter 10 this dissertation).

Modelers present their model as applied to one or more particular domains of phenomena. Such domains may be, for example, a set of psycholinguistic tasks, a particular type of neuronal responses, or activity levels in a particular brain region generated by certain tasks. However, a good model is not bound to this ‘native’ domain, and testing it should not be restricted to the list of predictions furnished by the modeler. Instead, models should be able to generate predictions outside its original domain (Jacobs & Grainger, 1994). This makes the model interesting for people outside of the narrow field of the modeler, it enables models to survive changes in fashion in empirical methodology, it ascents that predictions are not all riskless, and it just follows from the fact that the model models the brain (e.g., even with a broad domain of modeled tasks, it is unlikely that the modeled part does nothing else than performing those tasks).

Jacobs and Grainger (1994) distinguish two ways in which models may generalize outside their domain: horizontal generality refers to the ability of models to be applied to different tasks, behavioral measures and circumstances than originally envisioned. Vertical generality refers to the ability of a model to account for the behavior of the modeled system at different scales – for example different temporal scales. Although not mentioned by Jacobs and Grainger (1994), one could also add to this the ability of the model to account for data at different levels. In theory, such generality is the strong point of psychobiological models: they can generate relevant hypotheses both at a behavioral and at lower levels. Although, for example, a psychobiological model of memory may not be able to explain behavioral data as parsimoniously as a functional model does, by linking certain processes to particular brain regions it can generate novel predictions on
the outcome of brain lesions (Gluck & Myers, 1993; Norman & O'Reilly, in press). This will of course only work if the model is a genuine representation of the biological substrate being modeled — if it has structural accuracy. If others are to derive predictions from a model, the modeler must also be specific about what part of reality the whole model refers to. Else, any failure of a prediction of the model may always be disavowed on the basis that the model was misunderstood and in fact did not make the prediction that was tested. Such discussions do, in fact, occur in the literature.

11.4 MODEL CONSTRAINTS

Modeling strategies

Models can thus be supported by their match of real data, by generation of daring predictions, or by a priori plausibility.

- For a priori plausibility, all or almost all assumptions are supported by empirical evidence.

- For a model to be supported by its match of the data, it should be inflexible and offer little freedom to the modeler (or, at least, its freedom and flexibility should be small compared to the quantity of data fitted). This may be the case for two reasons: either because the model contains few assumptions (is sparse), or because all or many assumptions are bound by biology (are supported assumptions, in the terminology used here).

- For the model to be testable via the derivation of predictions, it should be vertically and horizontally general. For that, the ontology of the model should be clear.

From this analysis, two strategies for modeling immediately become apparent. The first would be to build a sparse, inflexible model that can be genuinely tested against data using techniques that punish the model for flexibility (Pitt et al., 2002; Zucchini, 2000). Examples of such models abound in the mathematical and connectionist memory model literature (Chessa & Murre, 2002; Raaijmakers & Shiffrin, 1981; Shiffrin & Steyvers, 1997), and, for example, in low-level models of neurons (Hodgkin & Huxley, 1952; Kistler, Gerstner, & van Hemmen, 1997; Volny-Luraghi, Maex, Vosdagger, & De Schutter, 2002).

This strategy would also be appropriate for models that are used as an existence proof or to derive computational constraints. Data analysis via a computational model, for example, is only appropriate if the model is transparent and simple, so that the relation between the empirical data and the model outcome is clear. An existence proof is valid independent of its complexity or the support for its assumptions. Nevertheless, an existence proof is more compelling when it is transparent, and when few assumptions are necessary to produce the wished-for behavior.

For many psychobiological models, however, this strategy is not plausible. They often reproduce data only at a qualitative level and are quite elaborate and flexible. Moreover, for models to be assessed on their match of existing data, model flexibility must to some extent be quantifiable. This is often not the case with psychobiological models.

The second strategy, which overlaps partly with the first, is to bind as many aspects of the model as possible to biology, by using supported assumptions and by making the ontology of the model explicit. This second strategy allows scientists other than the
modelers themselves to derive predictions from the model. In this way, the model becomes maximally testable. Moreover, binding assumptions to empirical evidence from biology makes the model a priori more plausible, and reduces flexibility of the model. This second strategy seems, given the difficulties in assessing flexibility, most appropriate for psychobiological modeling. It allows the natural strength of psychobiological models, their vertical generality, to be played out by generation of hypotheses at different levels.

Ontology of the model

Sometimes, however, the link between an assumption of the modeler and the evidence cited for that assumption are not wholly transparent to the outsider. Whether or not assumptions are supported can only be ascertained when it is clear what the model refers to in reality. If modelers are vague about what modules in their model stand for, then it would be disingenuous to claim support from neuroanatomy for the architecture of the model. Similarly, if modelers are vague about the time scale in the model of events such as spiking, it would be disingenuous to claim that the inclusion of complex spike-time dependent learning rules makes the model more plausible. For model assumptions to garner support, it is thus crucial that modelers are clear about the ontology of their model. This is even more true for generality: predictions can only be unambiguously derived from a model if it is clear to outsiders what the model is a model of.

A useful notion here is that of the ‘level’ of a model, one of Webb’s (2001) dimensions of model comparison. She specifically notes that Marr’s (1982) notion of levels of analysis is not intended here; one could instead speak of level of representation. Several hierarchies for such levels have been proposed, for example one consisting of the neuron, network, map, system, and whole-brain level (Sejnowski & Churchland, 1993). In setting up such hierarchies, there is no claim that each level has its own ontological status, or that phenomena at those levels are independent of one-another. Rather, each level is its own description of the same organ, the brain, that is the focus of research in neuroscience and psychology (Bakker & de Dulk, 1999). For evaluating support for models, a hierarchy specifying what kind of data can falsify an assumption, match or a prediction, is perhaps most useful. Such a hierarchy could, for example, consist of the following levels:

- The **behavioral level**: the behavior and algorithm of the model correspond to the behavior and algorithm of the modeled organism. The model thus represents part or whole of the ‘functional architecture’ of the whole brain. At this level, the model must match behavioral data, and can be expected to generate behavioral predictions.

- The **neuropsychological level**: identifiable parts of the model can be functionally mapped onto identifiable brain parts (such a part can be a map, a field, an area or a system). If this is the case, a lesion of the model part can be said to functionally correspond to a lesion of the brain part, allowing predictions to be derived about brain-lesioned humans and animals (e.g., Gluck & Myers, 1993; Norman & O’Reilly, in press). Assumptions on this level may be supported by lesion data, or the findings from fMRI research.

- The **systems level**: the part-to-part mapping stretches to the state of model parts during task performance. For example, more activity or plasticity in the model part may be thought to correspond to more neuronal activity or plasticity in the modeled brain part. In this case, the model may predict for example concrete fMRI signals (Fincham, Carter, van Veen, Stenger, & Anderson, 2002), plasticity
as revealed by cFos mapping, or in vivo multiecell or single cell recordings, and assumptions may be supported by, for example, gross anatomy.

- The *neurophysiological level*: the units of the model correspond to real neurons, for example sample neurons from a certain brain area (e.g., Sohal & Hasselmo, 1998). In this case, all physiological data at the cell level, and all anatomical data, can in principle be used to refute the model’s predictions, or its assumptions.

- The *biophysical level*: components of modeled neurons correspond to components of real, identifiable neurons (e.g., Volny-Luraghi et al., 2002). In this case, physiological data at subcellular level may be predicted and used for support for assumptions.

A psychobiological model is one that is situated at both the behavioral level, and at least one level below it. The more levels of representation a model has, the more data could refute the model: an outsider could derive more testable predictions from the model. Even if a model resides at just one level, the number of hypothetical predictions is very large given that it represents something. If it is clear what the model refers to, it can be combined with other theories of the same substrate, and/or with any number of conjunctive hypotheses, so that new predictions can be derived (Devitt, 1997). Whether or not these predictions can be tested with present techniques is of course another matter.

Psychobiological modelers are often not very clear on which level they see their model—whether or not, for example, their units can be seen as a sample of real neurons, or whether or not fMRI can falsify their model. The imprecision may concern one of the structures modeled (e.g., stating that a model stands for the ‘Medial Temporal Lobe’ without specification of this term). It can also be about the organism modeled (e.g., a model matches only human behavior, but the underpinning of its assumptions come from animal physiology). Although it can be claimed that a model is true of the target structure in several related species, such generality cannot be taken for granted or deduced from similarities in behavior (see Treves & Samengo, 2002, for an example). Finally, the imprecision can be one of the events modeled, for example if the time scale of the simulation is unclear. Moreover, heuristic assumptions that determine the behavior of model nodes may make it impossible to compare behavior of model parts with activity in the brain regions modeled. For example, if model nodes in a neural network have both positive and negative activations (as in the Hopfield network), it becomes difficult to compare node activation with neural firing in the target structure. The heuristic assumptions specifying behavior of the model may thus make derivation of predictions at a low level impossible.

An interesting example of prediction outside of the domain of the model, despite ontological vagueness of all three kinds, is offered by a model of human recognition memory (Norman & O'Reilly, in press). This model consists of a fairly worked-out hippocampal model, and a simple, quite abstract neocortical module that is identified with the ‘medial temporal lobe neocortex’. In this last module, patterns are stored in a set of weights between an input and an output layer. If a pattern is presented for a second time, it leads to higher activity in the ‘winners’ in the output layer of the neocortical region, which is interpreted as a familiarity signal.

In their presentation of this model, Norman and O'Reilly note an apparent contradiction between this mechanism and data from cell recordings from macaque perirhinal cortex, which seemed to show decreased firing for familiar patterns (Xiang & Brown, 1998). The ad-hoc explanation offered for this contradiction is less remarkable than the fact that Norman and O'Reilly felt compelled to comment on it. Their model is targeted at human
behavioral data, is rather abstract, and does not include activity measures on an explicit time scale. The fact that Norman and O'Reilly see monkey electrophysiology as relevant for their model implies that they see their model as situated on at least the systems level described above, and describing both the human and the macaque perirhinal cortex. Given that they derived successful predictions at the behavioral and neuropsychological levels, this gives their model considerable vertical generality.

The data gap

Being explicit about model ontology is of course on enough for assumptions to be supported by data, the data must also be there. At low levels, for example the biophysical level, this is often the case. Many parameter values in Hodgkin-Huxley formalism models can, for example, be extracted from experimental data (Hodgkin & Huxley, 1952). At higher levels no such abundance of restricting data exists, however. Both neuropsychological and imaging data is still relatively sparse, imprecise, and often not well understood. At these levels, that modeling is relatively free, unconstrained activity. The staggering variety in published computational models at these levels attests to this freedom.

It may be tempting to retreat to lower levels, where evidence is more abundant in the form of electrophysiological data; however, incorporating low-level data into a model that must also provide explanations for behavior may make the model unwieldy. Consider what would have happened if Norman and O'Reilly (in press) had set out to change their model to account for the electrophysiological data provided by Xiang and Brown (1998). It is very unlikely that assumptions needed to explain neurophysiology will be of much help in explaining mirror effects in recognition memory. They would have needed one set of assumptions to explain the phenomena at the neuronal level (as was done by Bogacz et al., 2001), and then another set to explain behavioral effects. Possibly, yet a third set of assumptions would have been needed to bridge the intermediate levels. As for many assumptions there would have been no data for support, this would have made their model top-heavy with heuristic assumptions. In such cases, when plausible biological underpinnings are not available or their inclusion requires the addition of many untested and heuristic assumptions, it may be preferable to disregard some facts in order to construct a simpler model (following the first strategy). This, in fact, was done by Norman and O'Reilly (in press).

There is an alternative, however, to falling back on an abstract model remote from biological reality, and to building a complex, unconstrained biological theory that needs many untested and heuristic assumptions to work. One may tear the complex, detailed model apart into several simpler models at different levels. In this way, a cascade of models would be created, where each model is a more or less detailed elaboration of the same idea. Models in the cascade could then constrain one-another and pose restrictions where behavior and neurobiology pose too few. This is the essence of the model cascade approach introduced below.

11.5 MODEL CASCADES

The central idea of model cascades is that one theory is not incorporated into one single model, but in several models at varying levels of representation. Lower-level models in this family are concretizations of higher-level models, higher-level models are abstractions / simplifications of the lower-level models. An example of such relations,
outside of computational modeling, is that between biochemical analysis of DNA and behavioral genetics. Our knowledge of DNA is an implementation of classical genetics (of which behavioral genetics is an offshoot), and behavioral genetics a useful abstraction of the biochemistry of DNA and meiosis. Biochemistry is, up to now, not appropriate for analyzing heritability of behavioral traits. This leaves an essential function to behavioral genetics (even though in certain details classical genetics and modern biochemical knowledge are at odds). There is thus added value to having two hierarchically organized versions of the same theory.

The relations between models at different levels in the proposed cascades are those between different levels in the hierarchy of Marr (1982) – either those between the implementation and algorithmic levels, or those between the algorithmic and competence theory levels (Webb, 2001). A lower-level model can, for example, be algorithmic implementations of higher-level models, with the higher-level model specifying the competences of the lower-level model.

These relations between models imply that no assumption in one model may be in contradiction with assumptions or behavior of the others. Ideally, all higher-level models would in principle be translatable into the framework of lower level models, without loss of function. For example, if a modeler had unlimited time and unlimited computational resources, all elements in his or her behavioral level model could be replaced by the most low-level biologically-grounded elements simulated at a millisecond scale, and the model would still be able to simulate behavioral phenomena occurring at a scale of minutes and hours. Although this may seem an unattainable ideal, given that any model needs heuristic assumptions to work, it is one with force. It excludes the use of some heuristic assumptions that preclude translation of the model into lower-level one, such as the use of negative activations, negative weights, and error-driven learning.

The advantages of model cascades are different depending on the amount of data available for constraints on modeling. Firstly, in fields of research where the general principles are not yet known (where there are too few biological and behavioral constraints to develop either abstract, testable models or elaborate theories), such modeling may allow the development of a coherent framework that is a viable theory. By splitting up the model in a model cascade, each model in that cascade can be simple and transparent, and data at all levels can be incorporated in or accounted for by the model. Behavior of the lower level model may feature as an untested assumption in the higher-level model. Without the lower level, this would just be one of possibly many assumptions that an outsider just has to accept or reject. With the lower-level model, the plausibility of the untested assumption can be assessed. Moreover, support for the lower level model would strengthen the higher-level model, and rejection of the lower-level model would mean at least a reconsidering of the higher-level model.

Several examples of such cascades have already been presented in the literature. In the retrograde amnesia literature, biological information only weakly constrains modeling, and the target behavioral data consists of a qualitative pattern of just two curves (the decreasing forgetting curve and increasing Ribot curve of graded retrograde amnesia). Here, accounting for empirical data while at the same time presenting a more basic model has been accomplished with two hierarchies of models. McClelland et al. (1995) presented a backpropagation network as an existence proof for their central mechanism, interleaved learning. With a mathematical abstraction of the same process, they fitted several retrograde amnesia curves to show how their process would apply in the real world.
Another cascade has, in this dissertation and elsewhere, been built around the TraceLink model of consolidation and amnesia (Murre, 1996; chapter 56 this dissertation). The assumptions underlying TraceLink were also included in a concise, mathematical model of learning and forgetting (Murre et al., in prep.; chapter 4 this dissertation). By fitting this model to curves from the amnesia literature, the theory will be explicited on the behavioral level, leading to, for example, estimates of the time course of consolidation. Moreover, TraceLink contains two untested assumptions about learning that were, at the high level of the TraceLink model, also not testable. TraceLink assumes that hippocampal codes are independent of the neocortical ones (i.e., if two patterns have similar neocortical representations, they will nevertheless have orthogonal hippocampal ones), and that modulation of learning produces high learning rates for novel patterns in the hippocampus, and low learning rates for old patterns. Both assumptions were fleshed out in separate, lower-level models, the first in a model of the parahippocampal gyrus (chapter 9), the second in a model of acetylcholinergic modulation of the hippocampus (chapter 10). In both cases, the lower level model makes explicit what needs to be the case for the higher level TraceLink model to be true. Moreover, they turn the untested—and untestable—assumption of TraceLink into the testable behavior of a lower level model.

With these two cascades, not much data was available to constrain modeling. In more developed areas, multilevel modeling can also play a useful role by allowing a modeler to combine the virtues of abstract, high level models with those of an inclusive, vertically general model-as-theory. One example is the relation between detailed modeling of spike generation allowed by the very complex Hodgkin-Huxley formalism (Hodgkin & Huxley, 1952), and the much simpler formalism of spike response models. By showing that these models produce approximately the same behavior, Kistler et al. (1997) in fact showed that their model was a simplification of Hodgkin-Huxley. This would allow modelers to use these neurons, enjoy its lower computational load and its reduced set of parameters, and still have the same structural accuracy as when they had used Hodgkin-Huxley formalism model neurons. In this way, the virtues of an explicit, low-level theory and a sparse, simple model can be combined.

11.6 DISCUSSION

Computational models of the brain hold great promise as a bridge between biology and psychology. Many models that attempt to bridge the gap, those that I referred to as psychobiological, have already been proposed. If these models are to fulfill this promise, however, they will have to be structurally accurate. A particular layer in a multiplayer perceptron may, for example, be said to stand for a cortical region, without there being a clear relation between the behavior of that region and the layer (cortical regions do not behave as MLP layers). What can one learn from such a model? That the MLP can generate the target behavior while the brain works in mysterious other ways is surely not a worthwhile lesson.

Structural accuracy is the goal of the model, not something that can be ascertained beforehand. Three ways to support the structural accuracy were discussed in this chapter. A model may be a priori plausible because of its use of supported assumptions. It may also be supported by its fit of existing data, provided the fits are obtained without much flexibility. Finally, it may be supported by the confirmation of predictions generated for the model. For all three sources of support, it was argued that models need to be bound to biology. As 'biological plausibility' is more often claimed than
substantiated (Jacobs & Grainger, 1994), such a claim is not enough. Instead, modelers must be explicit about the ontology of their model, and use assumptions supported by data. However, there is not always data available to support assumptions. In such cases, modeling may become the unconstrained construction of cathedrals of speculation.

In this chapter, a strategy was suggested for avoiding such unconstrained modeling in the face of lack of data, namely a model cascade. In this strategy, a hierarchy of models is developed that all share the same assumptions, and higher-level models could in principle be translated into the formalism of the lower-level models. In a sense, such a hierarchy incorporates the goals of classical reductionism, where constructs of high level are identified with entities on lower levels. These hierarchies of models allow the development of theory without unrestricted dabbling in theoretical Lalaland. Moreover, they may allow a combining of the virtues of sparse, testable models and vertically general, biology-rich models. Modeling the same substrate at different levels of representation, as proposed here and practiced in this dissertation, may thus have benefits that exceed the merits of each model in the hierarchy on itself.