Birte U. Forstmann
Eric-Jan Wagenmakers
Editors

An Introduction to Model-Based Cognitive Neuroscience

Springer

An Introduction to Model-Based Cognitive Neuroscience

Birte U. Forstmann • Eric-Jan Wagenmakers
Editors

An Introduction to Model-Based Cognitive Neuroscience

Springer

Editors
Birte U. Forstmann
Cognitive Science Center
University of Amsterdam
Amsterdam
The Netherlands

Eric-Jan Wagenmakers
Department of Psychological Methods
University of Amsterdam
Amsterdam
The Netherlands

ISBN 978-1-4939-2235-2 ISBN 978-1-4939-2236-9 (eBook)
DOI 10.1007/978-1-4939-2236-9

Library of Congress Control Number: 2015930523

Springer New York Heidelberg Dordrecht London

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This is the first book on model-based cognitive neuroscience, a nascent field that is defined by a reciprocal relationship between cognitive neuroscience and behavioral mathematical modeling. Traditionally, cognitive neuroscience and behavioral modeling are separate disciplines with little crosstalk. In recent years, however, neuroscientists have discovered the advantages of mathematical models of cognition and performance, whereas mathematical psychologists are increasingly aware of the fact that brain data can constrain mathematical models in ways that are useful and unique.

To stimulate the further integration between behavioral modeling and cognitive neuroscience, this book brings together 17 invited contributions from leading researchers in the field of model-based cognitive neuroscience. The main challenge in coordinating these contributions was to make the book accessible to both mathematical modelers and neuroscientists, a challenge we met in two ways. Firstly, the book starts with seven tutorial chapters: three of these chapters outline and illustrate the principles of mathematical modeling of behavior, another three chapters describe basic principles of brain function and structure, and the final tutorial chapter concerns the interaction between modeling and neuroscience. Secondly, in order to highlight the reciprocal relationship between the two fields, the five chapers in Part 2 feature applications that emphasize the value of modeling for neuroscience, whereas the five chapters in Part 3 deal with applications that center on the value of neuroscience for modeling.

The authors of each chapter have tried hard to make their work accessible. As a result of their efforts, this book can be used as the core material for an advanced undergraduate or graduate course on model-based cognitive neuroscience. To facilitate the use of the book for teaching, each chapter ends with a list of recommended readings and a series of questions. The readings can be used to expand the course materials, and the questions can be used to deepen the learning process. Teachers can obtain the answers to the questions upon request. The 17 chapters vary in scope and in difficulty, and we suggest that teachers cherry-pick the chapters they expect to be particularly relevant and appropriate for the student's background level of knowledge.

Just as the chapter authors, we are excited about the advantages and challenges that come with the integration of two disciplines, disciplines that share the same goal—to unravel the mysteries of the human mind—but so far have pursued that common goal in disparate ways. We hope that the enthusiasm with which the book was written is noticable for the reader, whether undergraduate student, graduate student, or academic staff member.

Finally, should you note any typographical errors, conceptual mistakes, glaring omissions, overgeneralizations, or anything else you feel requires correction: please do not hesitate to contact us so we can address these issues in an erratum that we will post on our websites.

Amsterdam
31-05-2014

Birte U. Forstmann
Eric-Jan Wagenmakers

Contents

Contributors

William H. Alexander Department of Experimental Psychology, Ghent University, Gent, Belgium

Anneke Alkemade University of Amsterdam, Cognitive Science Center Amsterdam, Amsterdam, The Netherlands

John R. Anderson Carnegie Mellon University, Pittsburgh, PA, USA

F. Gregory Ashby Department of Psychological & Brain Sciences, University of California, Santa Barbara, CA, USA

Rafal Bogacz Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK

Jelmer P. Borst Carnegie Mellon University, Pittsburgh, PA, USA

Joshua W. Brown Department of Psychological and Brain Sciences, Indiana University, Bloomington, USA

Scott D. Brown School of Psychology, University of Newcastle, University Avenue, Callaghan, Australia

Xiaomo Chen Department of Psychological and Brain Sciences, Johns Hopkins University, Baltimore, MD, USA

Jochen Ditterich Center for Neuroscience and Department of Neurobiology, Physiology and Behavior, University of California, Davis, CA, USA

Simon Farrell School of Psychology, University of Western Australia, Crawley, WA, Australia

Birte U. Forstmann University of Amsterdam, Cognitive Science Center Amsterdam, Amsterdam, The Netherlands

Department of Psychology, University of Amsterdam, Amsterdam, VZ, The Netherlands

Michael J. Frank Cognitive, Linguistic & Psychological Sciences, Brown Institute for Brain Science, Providence, USA

Andrew Heathcote School of Psychology, University of Newcastle, University Avenue, Callaghan, Australia

Max C. Keuken University of Amsterdam, Cognitive Science Center Amsterdam, Amsterdam, The Netherlands

Peter Kok Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Nijmegen, HB, The Netherlands

Floris P. de Lange Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Nijmegen, HB, The Netherlands

Stephan Lewandowsky University of Bristol, Department of Experimental Psychology, Bristol, UK

Gordon D. Logan Department of Psychology, Vanderbilt University, Nashville, TN, USA

Rogier B. Mars Department of Experimental Psychology, University of Oxford, Oxford, UK

Jill X. O'Reilly Centre for Functional MRI of the Brain (FMRIB), Nuffield Department of Clinical Neurosciences, John Radcliffe Hospital, University of Oxford, Oxford, UK

Thomas J. Palmeri Department of Psychology, Vanderbilt University, Nashville, TN, USA

Roger Ratcliff Department of Psychology, The Ohio State University, Columbus, OH, USA

Jeffrey D. Schall Department of Psychology, Vanderbilt University, Nashville, TN, USA

John T. Serences Department of Psychology, University of California, San Diego, CA, USA

Philip L. Smith Melbourne School of Psychological Sciences, The University of Melbourne, Melbourne, VIC, Australia

Thomas C. Sprague Neurosciences Graduate Program, University of California, San Diego, CA, USA

Veit Stuphorn Department of Psychological and Brain Sciences, Johns Hopkins University, Baltimore, MD, USA

Department of Neuroscience, Johns Hopkins University School of Medicine and Zanvyl Krieger Mind/Brain Institute, Baltimore, MD, USA

Brandon M. Turner Psychology Department, The Ohio State University, Columbus, OH, USA

Eric-Jan Wagenmakers Department of Psychological Methods, University of Amsterdam, Weesperplein 4, Amsterdam, The Netherlands

Part I
Tutorials

Chapter 1
An Introduction to Cognitive Modeling

Simon Farrell and Stephan Lewandowsky

Abstract We provide a tutorial on the basic attributes of computational cognitive models—models that are formulated as a set of mathematical equations or as a computer simulation. We first show how models can generate complex behavior and novel insights from very simple underlying assumptions about human cognition. We survey the different classes of models, from description to explanation, and present examples of each class. We then illustrate the reasons why computational models are preferable to purely verbal means of theorizing. For example, we show that computational models help theoreticians overcome the limitations of human cognition, thereby enabling us to create coherent and plausible accounts of how we think or remember and guard against subtle theoretical errors. Models can also measure latent constructs and link them to individual differences, which would escape detection if only the raw data were considered. We conclude by reviewing some open challenges.

1.1 Introduction

Your best friend introduces you to the host of the party you just joined. Seconds later, as you turn away, the horrible realization sinks in that you cannot remember the host's name. Was it James? Or Gerard? We have all experienced such reminders of the brittleness and imperfection of our memory and cognition. The principal goal of cognitive science is to understand the processes that underlie our cognition: why do we sometimes forget information within seconds? How come we still remember the name of our first-grade teacher? In this chapter, we show that to answer such

S. Farrell
School of Psychology, University of Western Australia,
35 Stirling Highway, Crawley, WA, 6009, Australia
e-mail: simon.farrell@uwa.edu.au

S. Lewandowsky
University of Bristol, Department of Experimental Psychology,
12a Priory Road, Bristol BS8 1TU, UK
e-mail: Stephan.Lewandowsky@bristol.ac.uk

B. U. Forstmann, E.-J. Wagenmakers (eds.), *An Introduction to Model-Based Cognitive Neuroscience,* DOI 10.1007/978-1-4939-2236-9_1

questions, we ought to rely on computational models rather than conducting our theorizing exclusively at a verbal level. Computational models instantiate assumptions about how cognition might operate in a precise and unambiguous manner, thereby permitting a rigorous test of those assumptions.

Several decades ago, in the early stages of the cognitive revolution that ultimately led to the overthrow of behaviorism, theorists were content with formulating theories at a verbal level. For example, they might have postulated that information enters a limited-capacity "short-term memory", which is subject to rapid forgetting unless the person makes an effort to transfer the information into a more stable "long-term memory" using a process such as rehearsal. If rehearsal is prevented because one's attention is occupied elsewhere, information such as a host's name can be forgotten within literally 1–2 s [1].

Although such verbal models have spawned considerable progress, their inherent limitations have become increasingly apparent. This chapter surveys some of those limitations and also clarifies how computational models can help overcome them.

We begin by showing how computational models can illuminate complex social interactions that are simply not amenable to verbal theorizing. How could one anticipate or predict events such as the "Arab spring"? Sparked by the self-immolation of a Tunisian roadside vendor in late 2010, this cascaded into the overthrow of several despotic governments in the Middle East. Even events that are less unexpected can be beyond the realm of verbal analysis. For example, who could verbally predict or describe the behaviour of a crowd escaping a building on fire?

Just because such complex dynamic events defy verbal analysis does not imply that they cannot be described at all: Complex social dynamics can arise, and can be understood, on the basis of some very simple assumptions about individual behavior in conjunction with computational instantiations of how those individuals interact. In a nutshell, such "agent-based" models are initialized by creating a population of agents that then interact with each other according to some very simple rules—e.g., "follow the majority of your neighbors." Over time, patterns of behavior and social structure can emerge that were not programmed into the agents and that verbal analysis would not have anticipated [2].

This emergence of social structure can be illustrated with the model presented by Kenrick et al. [3], which combined an evolutionary approach to understanding individual behavior with a dynamic representation of social interactions. Kenrick et al. conceived of psychological mechanisms as decision rules designed to address fundamental problems confronted by humans and their ancestors. A number of such basic decision rules were postulated; here we focus on a simulation of individuals' choice between cooperation with others and self-protective aggression. Both modes of behavior can deliver obvious advantages to an individual, but they can also incur drawbacks. A peaceful and cooperative person can be subject to exploitation if the environment is uniformly hostile, and a hostile person in a cooperative environment can waste time through unnecessary conflicts.

Much research on conformity and social norming [4–6] suggests that people frequently model their behavior based on what others around them are doing. If everyone in my neighborhood is conserving energy, then I am more likely also to try

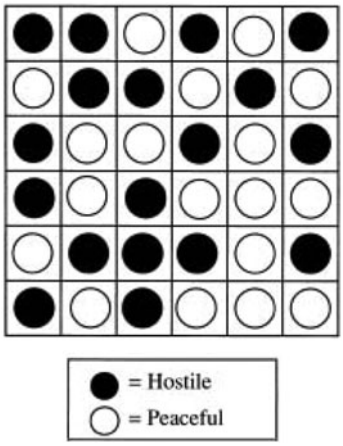


Fig. 1.1 A community of agents that is randomly initialized to being "hostile" or "peaceful" at the outset of a simulation. At each time step of the simulation, each agent will decide whether to retain its current attitude or alter it by inspecting all its immediate neighbors. If more than half the neighbors behave in the opposite manner from the agent, it will change its behavior to conform to that majority. (Figure adapted from [3], published by the American Psychological Association, reprinted with permission)

and reduce my consumption [6]. Accordingly, Kenrick et al. [3] built a simulation involving a community of autonomous "agents," each of which would change its current behavior if more than half of its neighbors acted differently. Thus, a hostile agent surrounded by a majority of cooperative neighbors would become cooperative at the next simulated time step, whereas it would remain hostile if surrounded by a majority of aggressive neighbors.

This decision rule may sound trivial at first glance, but its complexity should not be underestimated. First, the behavior of each agent has a direct effect on its immediate neighbors, and that relationship is reciprocal. That is, an agent is affected by its neighbors as much as that agent affects the neighbors. Moreover, there is a time-lagged effect of each agent's behavior on the neighbors' neighbors; whenever an agent changes its behavior, the consequences of this decision ripple through the community over time.

Suppose a community of agents is randomly initialized to contain an equal number of peaceful and hostile agents, each of which changes its behavior according to the neighbor-majority rule. Figure 1.1 shows one such random starting pattern with an equal number of hostile and cooperative agents. How might this random pattern evolve over time when agents adjust their behavior at each time step? As an exercise, pause for a moment now. Without reading ahead, think of how the initial configuration in Fig. 1.1 might evolve over time.

What did you think would happen after this random community has had a chance to interact for a number of iterations?

You may be surprised that in the model just presented, initial randomness does not beget randomness—quite to the contrary, over time this random community will

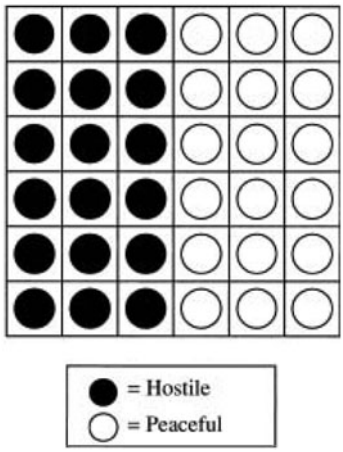


Fig. 1.2 The final state of the community of agents initialized as shown in Fig. 1.1 after a large number of time steps. No further changes will occur, regardless of how long the simulation is continued. (Figure adapted from [3], published by the American Psychological Association, reprinted with permission)

segregate into two "neighborhoods" of equal size; one consisting of cooperating agents, the other one consisting of hostile agents, with a clear boundary between the two clusters. This final configuration is shown in Fig. 1.2.

An intriguing aspect of this simulation is that if one inspected the final outcome only (Fig. 1.2), without any knowledge of how those segregated neighborhoods arose, it might be tempting to speculate that the simulation involved two different classes of agents, one with a propensity to cooperate, the other one with a disposition toward hostility. The simulation shows otherwise: All agents in the community are identical, and they are capable of expressing one or the other behavior depending on the context in which they find themselves. The regularity in the behavior of the system emerges entirely from the history of local interactions between the agents.

Notwithstanding, individual differences can matter a great deal as shown in the next simulation summarized in Fig. 1.3. The top left shows the initialization of the community, which now contains slightly fewer hostile ($N = 16$) than peaceful ($N = 20$) agents. First consider what happens when all individuals act in the same way, by applying the majority-neighbors rule, and there are no individual differences. In that case the minority's propensity for hostility is washed out over time and the community settles into a uniformly peaceful state (Panel a in Fig. 1.3).

Now consider what happens when a single "short-fused" individual is inserted into the community at the outset (in the second column, second row from the bottom, as marked by the arrow in Panel b). This individual differs from all the others because the agent will act aggressively even if only a single one of his or her neighbors is hostile. This single individual, despite being a single voice among many, prevents the community from reaching a stable state of peaceful bliss. Instead, as shown in Panel b, the community evolves to retain a pocket of hostility that is focused around that one short-fused individual.

However, this does not mean that a single individual with hostile tendencies will always inject pockets of aggression into a community: If that same person were inserted into a different, more predominantly peaceful, location in the community (denoted by the arrow in Panel c), then the community would absorb this single

individual and peace would ultimately prevail. Indeed, even the "short-fused" individual would end up being peaceful, simply because the "peer pressure" would be sufficiently great for the individual to change his or her behavior.

It follows that the evolution of social structure results from the interplay of an individual's propensity to engage in a certain behavior (in this case, to engage in hostility) and the micro-structure of the surrounding community. It is difficult to see how this interplay could have been discovered by verbal means alone: The differences in initial state between Panels b and c are quite subtle and the consequences of the different initializations became observable only by simulation.

To underscore that point, ask yourself what would happen if the community were seeded with *both* "short-fused" individuals (i.e., from Panels b and c) at the outset. Again, pause a moment and without reading ahead, try to answer that question.

The result is shown in Panel d of Fig. 1.3. With two short-fused individuals among them, the entire community converges into a state of persistent mutual hostility—notwithstanding the fact that at the outset only a minority of 16 out of 36 agents were acting in an aggressive manner.

The research by Kenrick et al. that we just discussed [3] illustrates several important points about computational models. First, it is difficult to conceive how the outcome of the simulations could have been anticipated by verbal means alone—who would have thought that a single short-fused individual suffices to turn a neighborhood into a pocket of hostility, even if the majority of community members is peaceful.

Second, the only regularity that was explicitly programmed into the simulation was the rule by which each agent changed its behavior based on the neighbors' conduct. The simulations yielded insights into the emergence of social structures that went far beyond the small "investment" of the rule that was programmed into the simulation.

Third, a corollary of the unanticipated emergence of social structures is that it would be difficult for us as humans to do the backwards inference. When confronted with a segregated community (Fig. 1.2), we might be tempted to infer that there are two different classes of agents that are inherently different from each other. In reality, the segregation arose from a random initialization of the agents' current behavior, and the entire population behaved according to the same rules from then on. Absent a computational model that explains the underlying processes, our backward inference could have been completely wrong.

Finally, the simulations showed that even some very simple assumptions may suffice to model complex behavior, in this instance the emergence of social structures. It is this last point that we are especially concerned with in this chapter: How apparent complexity in behavior can be explained by sometimes surprisingly—and gratifyingly—simple underlying processes. Such simple processes would be difficult to infer directly from behaviour—as noted above, the most obvious inference in the first simulation would be that differences in people's behaviour reflects individual differences. One further implication is that computational modelling can look beyond differences in observed behaviour to reveal underlying regularities.

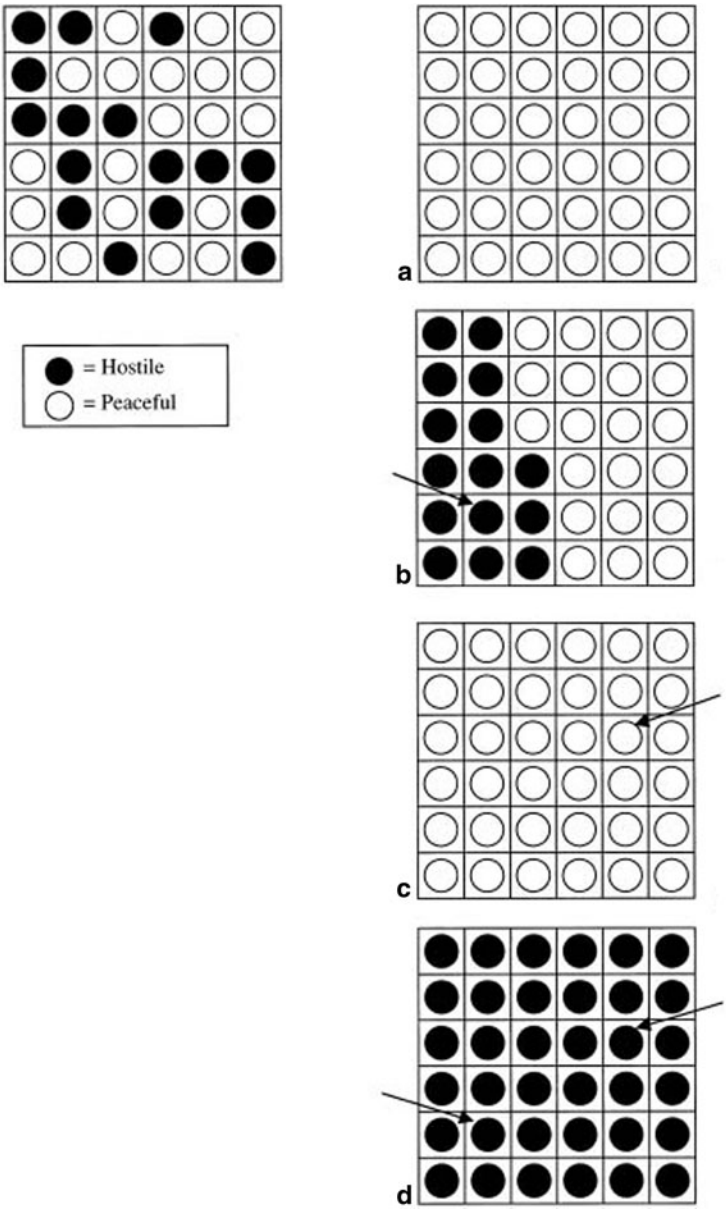


Fig. 1.3 The evolution of four communities that began with an identical initial arrangement of overt behaviors (shown at the *top left*). Panel **a** shows the final state if all individuals use the normal majority rule. Note that unlike in Fig. 1.2 all agents end up in a peaceful state because there are slightly fewer hostile than peaceful agents at the outset. Panel **b** shows the final state if one individual (marked with the *arrow*) is "short-fused" and will act aggressively if *any* of its neighbors are hostile. Panel **c** shows another final state when a different individual is short-fused instead (marked with the *arrow*). The final panel **d** shows what happens when there are two short-fused individuals (marked with *arrows*) at the outset. (Figure reprinted from [3] , published by the American Psychological Association, reprinted with permission)

Now that we know *why* computational models are attractive, we can take a more detailed look at how exactly they function.

1.2 What is a Cognitive Model?

At its most basic, a model is an abstract construct that captures structure in the data [7]. For example, a good model for the set of numbers {1, 3, 5} is their mean, namely 3. A good model for the relationship between a society's happiness and its economic wealth is a negatively accelerated function, such that happiness rises steeply as one moves from poverty to a modest level of economic security, but further increases in happiness with increasing material wealth get smaller and smaller [8]. The two models just mentioned are called *descriptive*, because they do little more than describe data or a relationship between variables.

Finding the correct description for a data set can often be helpful and address theoretical questions. For example, there has been considerable debate about how exactly a new skill is acquired with practice. Does the observed speed-up follow a "Power Law" or is better described by an exponential improvement [9]?

There is no doubt that the benefits from practice accrue in a non-linear fashion: The first time you try a new task (e.g., creating an Ikebana arrangement), completion can take a long time. For the next few trials, you will notice vast improvements, but the size of those improvements will gradually decrease. For several decades, the prevailing scholarly opinion was that the effect of practice is best captured by a "Power law"; that is, by the function (shown here in its simplest possible form),

$$RT = N^{-\beta}, \tag{1.1}$$

where RT represents the time to perform the task, N represents the number of learning trials to date, and β is the learning rate. More recently, it has been proposed [9] that the data are better described by an exponential function:

$$RT = e^{-\alpha N}, \tag{1.2}$$

where N is as before and α the learning rate. Far from being a mere technical matter, the choice of descriptive model for these data has important implications about the psychological nature of learning. The mathematical form of the exponential function implies that the learning rate, relative to what remains to be learned, is constant throughout practice [9]. That is, no matter how much practice you have had, learning continues to enhance performance by a constant proportion. By contrast, the mathematics of the power function imply that the relative learning rate slows down as practice increases. That is, although you continue to show improvements throughout, the rate of learning *decreases* with increasing practice. It turns out that the latest view on this issue favours the exponential function over the hitherto presumed "Power law" [9, 10], suggesting that our skills continue to improve at a constant relative rate no matter how much practice we have already had.

However, frequently cognitive scientists want to do more than describe the data. For example, we frequently want to *predict* new observations; we may want to know how much happiness in our society will increase if we expand our gross national product by $ 1,000,000,000. (If you live in a wealthy country, the answer is "not much.")

As important as prediction is to the pursuit of science, it is rarely the full answer. Suppose you owned a robot that sat on your kitchen bench and that on demand successfully predicted the outcome of *any* conceivable psychological experiment. However convenient this gizmo might be, it would not put an end to the basic endeavour of cognitive science because robotic predictions do not *explain* the phenomena under consideration [11].

Thus, most cognitive modeling goes beyond description and it also goes beyond prediction. Instead, most models are used as an explanatory device that formalizes our understanding of how human cognition operates. An intriguing attribute of models is that they are necessarily simpler and more abstract than the system—human cognition—they are trying to explain [12]. Models seek to retain the essential features of the system while discarding unnecessary details. By definition, the complexity of models will thus never match the complexity of human cognition—and nor should it, because there is no point in replacing one complicated thing that we do not understand with another [11]. This slightly counter-intuitive point is known as Bonini's paradox [13, 14].

What, then, would it mean to *explain* something? We illustrate the answer within the context of the practice functions just reviewed: Having described data from skill acquisition by a suitable (exponential) function, we now wish to find out how this functional form comes about. What underlying elemental cognitive process gives rise to the observed systematic decline in latency? One explanatory model for the speed-up of performance with practice is Logan's "instance model" [15]. We choose this model because it is as simple to describe as it is powerful. According to this model, performance of a task is initially accomplished by using a slow and deliberate algorithm (i.e., a recipe) to complete the task. For example, when confronted by an "alpha-arithmetic" problem, such as "$A + 4 =$?", novices typically attain the solution by stepwise increment (i.e., moving 4 letters through the alphabet from A; "B, C, D, E"). When the task has been successfully completed, the model then implements "learning by doing" by storing an instance of that solution in memory. Thus, the learner will memorize "$A + 4 = E$". When we come to perform the task again, the slow algorithm is again invoked, but it races against the retrieval of the answer from memory if the same problem and its solution has been encountered previously. Whichever process completes first is being used to respond (and to consequently lay down another instance in memory).

There are two sources of speed-up in the model. The initial speed-up arises because retrieval of instances is substantially faster than the slow algorithm. Recalling that "$A + 4 = E$" is far quicker than moving through the alphabet from A; "$B, C, D, \ldots E$." Instances are therefore more likely to win the race than the algorithm, and hence as more instances are encoded, response times will rapidly decrease because fast instance-based responding is increasingly likely to dominate over the

slow algorithmic processing. While this initial speed-up may be easy to describe verbally, the explanation for the continuing but slower speed-up later in practice is more subtle.

Indeed, the continuing speed-up may appear problematic for the instance model at first glance: After all, once all stimuli in the training ensemble have been presented, then (assuming a perfect memory) all subsequent responding to the same problems should be driven by retrieval of the appropriate instance. So why does responding continue to improve? The answer lies in the fact that each trial results in the storage of a new instance—even for stimuli that have already been encountered previously. Thus, over time multiple instances of each problem are encoded into memory. There are multiple copies of "$A+4=E$", all of which race for retrieval when the problem "$A+4=?$" is shown again. This is where things get to be interesting.

In an excellent demonstration of the benefits of computational modelling, Logan [15] analyzed the statistical properties of the minima of a certain distribution called the Weibull distribution. The Weibull distribution has a long tradition of being used to model response times in psychology [16], and the instance model builds on this tradition by assuming that the retrieval time of each instance is described by a Weibull distribution. Because *all* memorized copies of a problems compete in the race to respond, the model's response time is determined by the copy of the solution that is retrieved most quickly. It follows that response time is described by the properties of the minimum of a set of samples from a set of Weibull distributions. It turns out that as the number of copies of a given instance in memory (e.g., "$A+4=E$") increases, the "winning" response time—i.e., the minimum—will decrease as a power function of the number of copies. The more instances there are, the faster is the minimum retrieval time across the samples from the corresponding set of Weibull distributions. It follows that the model will continue to exhibit a (relatively slow) performance improvement, even beyond the point at which the slow algorithm no longer meaningfully competes with the retrieval of instances. At the time of its introduction, the instance model's ability to explain the then-prevalent power "law" of practice based on some very simple psychological principles was impressive indeed.

Since then, the power law of practice has been largely abandoned and there is now considerable agreement that practice functions are better characterized as being exponential in shape [9]. What are the implications of this realization for the instance model? On the one hand, this empirical re-assessment clearly challenges the instance model in its original form: As originally formulated, the fundamental properties of the model necessarily imply the power law, and the rejection of that law therefore falsifies the instance model. On the other hand, the basic assumption that performance involves a race between memorized instances is compatible with an exponential form of the practice function, provided that assumptions concerning the distribution of race times and response selection are relaxed [17]. Taken together, we can draw the following conclusions that illustrate the power of computational modeling: First, the original instance model is not compatible with the empirical database as it has evolved since the 1980's. Second, one of the attractive psychological principles underlying the instance model, namely the encoding of copies of all stimuli and selection of a response via a race between competing alternatives, has survived empirical scrutiny

and continues to be a contender for an explanation of human skill acquisition and indeed other phenomena [17].

1.3 Why Do We Model?

1.3.1 Validating our Reasoning and Understanding

Science depends on reproducibility. That is why Method sections must offer sufficient detail to permit replication of a study by another lab, and that is why replication of findings is such an important endeavour. Concerns about replicability of psychological findings have therefore become an important research topic in their own right [18].

There is another aspect to reproducibility that is tacitly taken for granted by most researchers but often fails to be explored in sufficient depth: Scientists assume that we are all *reasoning* on the same terms. However, like it or not, communication among scientists resembles a game of "telephone" (also known as "Chinese whispers") whereby theories and models are formulated and recorded on paper, before being read by the next scientist who needs to understand them, and may summarise those ideas in their own paper. Each step in this chain involves cognitive reasoning, and is thus subject to the known limitations of human cognition—from forgetting your host's name in a few seconds to the confirmation bias, to name but two [19].

The implications of this inescapable reliance on human reasoning can be illustrated with the popular "spreading activation theory" [20, 21] which postulates that concepts in memory (i.e., our knowledge of *dog* or *cat*) are represented by an interconnected network of nodes. Nodes are activated upon stimulus presentation, and activation spreads through the connections to neighboring nodes. To understand and communicate the notion of spreading activation, several analogies might be used: Some researchers liken the spread to electricity passing through wires [22] whereas others liken it to water passing through pipes (as one of us has done in lectures to undergraduates). Akin to the way in which we understand physical systems [23], the analogy adopted will determine people's precise understanding of the operation of the model. The water analogy necessarily implies a relatively slow spread of activation, while an electricity analogy will imply almost instantaneous spreading of activation. As it turns out, the data agree with the electricity analogy in showing activation of distal concepts to be almost instantaneous [24]. This problem—that the choice of analogy will affect a scientist's understanding of her own model—will undoubtedly be compounded when theorizing involves groups of scholars who communicate with each other. What the group considers to be a shared understanding of a model may in fact be limited to a shared understanding of only some core features. The implications for a potential lack of reproducibility under these circumstances are obvious.

Those reasoning problems can be alleviated by using computational models in preference to verbal theorizing. A principal advantage of computational models is

that we are forced to specify all parts of our theory. In the case of spreading activation, we must answer such questions as: Can activation flow backwards to immediately preceding nodes? Is the amount of activation unlimited? Is there any leakage of activation from nodes? Such questions have been answered in Anderson's [25] implementation of spreading activation in a memory model based in the computational framework of his ACT (Adaptive Control of Thought) theory (see also [26], this volume). This theory represents knowledge as units (or nodes) that are associated to each other to varying degrees. Closely related concepts ("bread–butter") have strong connections and concepts that are more distant ("bread–flour") have weaker connections. When concepts are activated, the corresponding units comprise the contents of working memory. Units in working memory become sources of activation, and pass their activation on to other units to an extent that is proportional to their own activation and the connection strengths.

The model has an effective limit on the amount of activation by assuming some loss of activation from the source units. The model also assumes that activation can flow back along activation pathways. The model uses these and other assumptions about encoding and retrieval to explain spreading activation and numerous other phenomena, such as serial order memory over the short term [27] and practice and spacing effects [28, 29]. Detailed specifications of this type, which verbal theories omit altogether, render a computational model more readily communicable—e.g., by sharing the computer code with other scholars—and hence more testable and falsifiable.

1.3.2 Examine Necessity and Sufficiency

Let's suppose, then, that our computational model has prevented us from going astray in our reasoning and theorizing. Let's furthermore suppose that we have fit the model to a broad range of data. What conclusions can we legitimately draw from this successful modeling?

At the very least, a running model can (to quote Fum et al.) " . . . be considered as a sufficiency proof of the internal coherence and completeness of the ideas it is based upon" (p. 136) [12]. However, it does not follow that the model is also *necessary*—that is, that the model provides the sole unique explanation for the data. Just because you flew from Lagos to Tripolis on your last African trip does not mean that you could not have taken an overland caravan instead. This point may appear obvious but in our experience it is frequently overlooked and its implications are rarely fully recognized.

The fact that a good model fit provides evidence of sufficiency but not necessity implies that the fit represents only fairly weak evidence in the model's favor: The model is only one of many *possible* explanations for the data but not *the only* possible explanation. This is an in-principle problem that has nothing to do with the quality of the data or the model. There are always other ways in which the available data could have been accommodated. For example, there exists an infinite number

of possible models of planetary motion because relative motion can be described with respect to any possible reference point—it just so happens that the Sun is a particularly convenient point. This indeterminacy of a model is often referred to as the "identifiability problem," to acknowledge the fact that the "correct" model almost necessarily escapes identification. The identifiability problem has led some researchers to suggest that process modeling should be abandoned altogether in favor of alternative approaches [30].

Our preferred interpretation of the problem acknowledges its existence but highlights what it does *not* imply: First, the fact that many alternative models exist *in principle* does not imply that any of those models are trivial or easy to come by—on the contrary, constructing cognitive models is an effortful and painstaking process that is far from trivial. Second, the existence of an unknown number of potential models does not preclude comparison and selection from among a limited set of actually instantiated models—after all, there is an infinite number of possible models of planetary motion; however, this has not precluded selection and universal acceptance of the heliocentric model.

Moreover, there are some circumstances in which even a demonstration of sufficiency, via a good model fit, may be impressive and noteworthy. Those instances arise when the model is a priori—i.e., on the basis of intuition or prior research—*unlikely* to handle the data.

We now present two examples that illustrate those points: How model selection can provide us with at least some sense of the necessity of a model, and how sometimes even in the absence of model comparison the success of a single model can be highly informative.

1.3.2.1 Model Comparison

One field in psychology that has received considerable attention from mathematical psychology is categorization. Categorization refers to our ability to map the potentially unlimited number of objects that we might encounter in the world into a relatively small number of discrete categories. This enables us to recognize some small furry animals as cats and others as dogs, thereby permitting us to predict whether the animal might meow or bark. It also permits us to communicate to others that a tomato is red and a canary yellow, without having to worry about the fact that our visual system can discriminate thousands of colors. Two models that have been influential in accounting for this ability to categorize are General Recognition Theory (GRT; [31]) and the Generalized Context Model (GCM; [32]).

According to GRT, we categorize objects on the basis of boundaries that partition the multidimensional space in which objects exist into different categories. Categorization errors arise from the noisiness in locating objects in that space. For example, cats are defined by a multitude of features, ranging from the length of their fur to the number of legs and length of whiskers and so on. Each particular cat in the world is defined by the conjunction of those features, and hence occupies a distinct point in a multidimensional space whose axes are defined by those features. Dogs

can likewise be represented as points in this multidimensional space, although they would tend to occupy a different part of it—because they are larger, for example, and their vocal output occupies a different frequency range. GRT proposes that there is a boundary that bisects the size dimension somewhere, such that all stimuli to one side are considered dogs and those to the other side are considered to be cats (analogous boundaries exist along all other dimensions as well).

GCM likewise assumes a multidimensional representational space, but instead of proposing the presence of boundaries, it assumes that each object that we have previously encountered in our lives is represented in memory together with its category label, in the same way that Logan's instance model—discussed above—assumes that we accumulate traces of experiences with practice. In GCM, new objects are categorised by matching them to all instances or "exemplars" in memory; this matching is done on the basis of a similarity calculation that involves an exponential transformation of the distance in multidimensional space between the new object and each stored object. If the summed similarity between a new object and, say, all dogs in memory is greater than the summed similarity involving cats, then the new object is classified as a canine.

Both the GCM and GRT account for a wealth of data from categorization learning tasks [33–35]. Indeed, it has been argued that GCM and GRT are so good at fitting empirical data that it is difficult to tell the models apart [36]. Accordingly, to differentiate between the models, Rouder and Ratcliff [36] designed a series of studies with an ingenious selection of stimuli that permitted a direct quantitative model comparison. The experiments used probabilistic feedback, such that the category membership of each stimulus was imperfectly defined: On any given trial, a stimulus might be classified as belonging to category A or to category B, albeit with differing probabilities (e.g., 70 % A vs. 30 % B). The task was meaningful because each stimulus had associated with it a response (e.g., "A") that was preferentially—albeit imperfectly—considered correct.

The probabilistic feedback was designed to elicit differentiable predictions from the models: GRT predicted that certain objects further from the boundary of two categories were more likely to be assigned to a particular category (category A) than objects closer to the boundary, whereas GCM predicted the reverse. The reasons for this reversal are subtle, but in a nutshell it arises because the GRT exclusively relies on the distance of objects from a boundary and must therefore predict objects to one side (especially if far from the boundary) to be consistently assigned to that one category. The GCM, by contrast, is sensitive to local structure—that is, specific frequency with which specific instances occur—and can therefore respond more heterogeneously with respect to distance from the boundary.

Rouder and Ratcliff put the models in direct competition by fitting both to the same data (see also [37]). This approach implied that although both models might fit well, any quantitative advantage in fit could be seen as favoring that model over the other. One interesting outcome of Rouder and Ratcliff's model comparison was that neither model was a clear "winner": GCM better accounted for data from experiments with relatively non-confusable stimuli, whereas GRT's fit was superior for confusable stimuli. This modeling result suggests that categorization is supported by

multiple cognitive processes and that people can rely on one or the other depending on circumstances: When the stimuli are distinct from each other and few in number, people prefer to memorize them and base their decisions on instance-based comparisons, exactly as predicted by the GCM. When there are many stimuli, and they all resemble each other, people instead compare test stimuli to a category boundary, exactly as expected by the GRT. More recently, it has become apparent that people can choose quite flexibly between different ways of solving a categorization task even when the stimuli remain identical [38, 39].

Quantitative model comparison has become a standard tool in cognitive science, and the outcome of comparisons has influenced theory development in many instances [39–43]. One issue of particular prominence in model comparison concerns the role of model *complexity*. This issue is beyond the scope of this chapter, but in a nutshell, the more complicated a model is the more data it is likely to be able to explain. This complexity advantage must be corrected in order to place models on an even footing during comparison (see, e.g., Chaps. 5 and 7 in [14]). As it happens, the models compared by Rouder and Ratcliff in the preceding example were equal in complexity, which allowed us to set aside this issue for our tutorial example.

1.3.2.2 Sufficiency of a Single Model

Many of the modelling papers in cognitive psychology take a new or existing model, and show that it is able to produce behavior that matches existing behavioral data: This approach delivers the demonstration that the implemented model under consideration is *sufficient* to account for the data. We noted earlier that this is often a relatively weak claim because it does not consider other competing explanations for the same data. Nonetheless, if a model can account for a broad range of results, ideally spread across multiple paradigms, this at least shows that the model is not a one-trick pony. Moreover, the impact and utility of sufficiency demonstrations increases in proportion to how surprising the success of a model is: When a priori human reasoning suggests that the model should *not* be able to account for data, then a success becomes particularly informative.

One such example involves the SIMPLE (Scale-Invariant Memory, Perception, and LEarning) model of Brown et al. [44]. This model treats the problem of explaining memory as fundamentally similar to the process by which we discriminate between different stimuli in multidimensional space (as in the GCM that we just discussed). One dimension that is particularly relevant to our ability to discriminate different events is time. According to SIMPLE, objects in memory are represented along an obligatory (logarithmically compressed) temporal dimension, although other dimensions may become relevant in particular situations [45].

One important feature of this model is that it eschews the distinction between memory systems operating over different time scales—specifically, SIMPLE does not differentiate between short-term memory and long-term memory—and it assumes that the same principles of temporal discrimination hold regardless of whether memory is observed over seconds, minutes, days, or even years. An important success of

SIMPLE has therefore been to show that it can account for data that are traditionally taken to suggest temporally-limited memory stores or processes.

For example, classical amnesia usually involves an impairment for temporally distant items during a free recall task, with the most recently presented items being spared. This has often been interpreted within a "dual-store" theory as reflecting a selective impairment of long-term memory (LTM), accompanied by an intact short-term memory (STM) system [46]. In simulations with SIMPLE, Brown and colleagues [47] showed that this pattern of impairments could be accounted for by assuming that amnesic patients have uniformly worse memory, and that the apparent sparing of STM arises from patients' tendency to rehearse information as it is presented, rather than trying to rehearse entire sequences of information: implementing this rehearsal schedule in SIMPLE, along with the assumption of worse temporal discrimination to produce overall worse memory, produced predictions that were closely aligned to the data. It is important to emphasise the implications of this modeling result: A pervasive pattern of memory impairment that had hitherto been preferentially interpreted as reflecting the operation of two memory systems was found to be explainable by a single-process model once a potential explanatory variable (viz. rehearsal schedule) was controlled for. By establishing an alternative explanation, this outcome considerably weakens the link between the observed pattern of memory impairment and the dual-store model. At the same time, by showing that a single process can account for the data, a sufficiency proof for SIMPLE has been obtained. In this instance, the sufficiency proof is particularly impressive because at first glance the data seemed strongly to point towards a different theoretical alternative.

The ability of SIMPLE to provide surprising sufficiency proofs is not limited to one phenomenon. SIMPLE has similarly provided a unifying explanation for different rates of forgetting at different time scales, which previously had sometimes been ascribed to different memory systems [48]. Likewise, SIMPLE can handle a variety of results that previously had been taken to present evidence for the existence of consolidation, a putative process that solidifies memories after initial encoding for some ongoing period of time [49]. The fact that a single-process model can handle a number of results that have conventionally been interpreted as reflecting the involvement of multiple memory systems is notable: Although in all instances the model has merely been shown to be sufficient, rather than necessary, the fact that sufficiency was established "against the odds" and in light of strong competing explanations permits one to take greater confidence in the results.

1.3.3 Using Models to Measure Underlying Individual Variables

The last few decades have seen an increasing trend towards describing and analyzing the behavior and cognition of individuals rather than groups of people (unless, of course, we are interested in group behavior, as in our introductory example involving peaceful and hostile agents). Arguably, modeling of aggregate group data (e.g., an average score in a condition) is no longer acceptable in many circumstances.

The emphasis on individual-level performance has given rise to a novel application of computational models, namely the identification and description of individual differences via model parameters that would otherwise escape detection by conventional statistical means. A good example of this is the use of response times to infer underlying differences in mental processes or abilities. Even without computational modeling, individual differences in the speed of responding can be informative: for example, the capacity of people's working memory (WMC) has been found to correlate with the time taken to make certain types of *voluntary* eye movements, such as a saccade *away* from an orienting cue. For automatic eye movements, such as a saccade towards an orienting cue, no correlation with WMC is observed [50]. The fact that a working memory measure, such as the complex-span task in which people must encode a list of items for immediate serial recall while also performing a distractor task (e.g., simple arithmetic), correlates with a task that involves no memory component—viz. voluntary eye movements—implicates a shared mechanism of voluntary control, arguably best characterized as executive attention, that links the two tasks and that varies between individuals. (The further fact that WMC does not correlate with nearly identical eye movements that do not involve voluntary control further sharpens the focus on attention as opposed to some more general ability factor.)

However, considering only an individual's overall speed of responding fails to identify the reason for why individuals may differ in their speediness. More information can be gathered when response-time models are used to analyse response times in tasks in which people must choose between actions or answers (e.g., whether a word stimulus represents a plant or an animal). Several models exist that describe such tasks and they offer several cognitive mechanisms that could produce a faster or slower response time [51]. Specifically, people may be faster or slower at collecting evidence in favour of making a response; they may require less (or more) evidence before making any choice; they may be biased towards a particular response; or they may simply have faster motor responses, independent of the core cognitive processes of interest. Consequently, observing a correlation between mean response time on a task and some other measure is unlikely to tell us the whole story.

Enter an approach known as "cognitive psychometrics" [52, 53], which allows researchers to extract maximal information from response time data. Researchers have begun to fit response-time models to peoples' behaviour so as to extract these underlying variables, and examine the relationship between those "hidden" variables and other factors of interest (e.g., WMC).

We illustrate this approach by considering the work of Schmiedek and colleagues [54], who examined the relationship between WMC (extracted from performance on a battery of working memory tasks, including several complex-span tasks) and various components of response time estimated from a number of speeded choice tasks. That is, instead of considering people's "raw" response times, Schmiedek et al. fit a response-time model [55] to the data of each participant. Via model fitting, the raw data were replaced by estimates of the model's parameters, and it was those parameters that were then related to WMC. In a nutshell, in the same way that a slope and intercept can capture the important trends that characterize a bivariate

point cloud, fitting response-time models to an individual's ensemble of responses can provide estimates of the important underlying cognitive components that drive those responses.

Schmiedek et al. found a strong relationship between WMC and the so-called drift rate, which is a parameter that captures the speed with which people can collect the evidence from a stimulus on which to base their decision. At the same time, an estimate of the non-decision time (i.e., the component of response time that was required for things such as motor execution of the response, encoding of the stimulus, and so on) was found to be independent of WMC. That is, although WMC was correlated with speed of responding, this correlation arose only from the speed with which people can extract information from a stimulus—none of the other component processes correlated with WMC. This selective association points to the likelihood that drift rate reflects a central cognitive ability that is supporting higher-order cognition.

Lest one think that only response times are amenable to model-based individual-differences analysis, we briefly present an investigation by one of the present authors [56] that sought to examine the role of WMC in category learning. In Lewandowsky's study, people's WMC was again measured by a number of tasks. In addition, participants completed 6 different categorization tasks involving three-dimensional binary stimuli—namely, the famous problems introduced by Shepard, Hovland, and Jenkins [57]. Performance on all 6 tasks was found to be uniformly related to WMC, such that people with greater working memory capacity tended to learn all 6 categorization tasks more quickly. This result is surprising, because it runs counter to the predictions of a popular view of categorization which invokes different memory systems as underpinning different tasks—and at least one of those systems is predicted to function independently of working memory [58]. But what ability exactly was it that drove performance in all tasks? To find out, the categorization performance of each individual on each task was modeled by ALCOVE [59], a model based on the GCM discussed earlier but with the additional ability to learn and improve performance over trials. Similar to the results of Schmiedek et al. only one of the model parameters was found to relate to WMC; namely, the learning rate. This suggests that working memory is associated with the ability to form long-term associations between instances in memory and the corresponding responses, rather than contributing to categorization performance in some other way. For example, Lewandowsky found that the precision of exemplar memory, represented by another parameter in ALCOVE, was not related to WMC.

Finally, recent work reinforcing one of the general themes of this book—the benefits of modelling neural processes with computational models—has examined correlations between model parameters and brain activity, showing that changes in response urgency specifically correlated with changes in activity in the basal ganglia [60]. These are just a few of the growing number of instances in which modeling has served to uncover individual differences that would have otherwise remained hidden, thereby attesting to the great promise of cognitive psychometrics [52, 53].

1.4 Open Challenges

We have only scratched the surface of huge number of papers that fruitfully employ computational modelling to understand cognition and behaviour, and we anticipate these methods will become increasingly useful—if not necessary—as we explain more complex phenomena. Looking to the future, it is clear that some open challenges remain for computational modelling and mathematical psychology. One is the expertise required to competently design and test models, and the consequent challenge of "skilling up" new generations of researchers in these methods. Townsend [61] noted the apparent drop in availability of training in mathematical psychology at both the undergraduate and graduate levels, in contrast to the increasing popularity of graduate training in areas such as cognitive neuroscience. These difficulties have no doubt been exacerbated by the lack of introductory textbooks on the topic, although recent entries (including the book you are reading now; see also [14, 62]) may encourage researchers with this expertise to start offering courses in their own universities. Other potential limitations to the adoption and use of computational modelling have also been addressed in recent years. For example, a common complaint amongst critics of computational modelling was its lack of replicability. Leaving aside the even more pronounced lack of replicability of *mental* simulations, models have been made more open and accessible by researchers placing the source code for their models online, and recent successive editors of *Psychological Review*—Keith Rayner and John Anderson—have adopted the excellent policy of requiring modellers to submit their model source code as supplementary material on acceptance of their paper. Such activities make modellers' work more open to inspection by other researchers, and provide good, concrete examples for those just learning how to model.

1.5 Concluding Comments

In summary, given the increasing complexity of theories of mind and brain, computational modelling is becoming increasingly important as a tool that allows us to look beyond the immediate data to make inferences about underlying processes. We've discussed how we can use models to validate our own reasoning, and ensure that the behaviour we intuit from our inherently limited reasoning abilities matches to the *actual* behaviour of the model. This is particularly the case in models of complex social systems, where the emergent behaviour cannot be easily predicted using mental simulation. Modelling also aids us in making claims about the sufficiency and necessity of theoretical mechanisms, by using the relative fit of several competing models to identify mechanisms that were more likely to have generated the data, and by identifying models that provide unifying accounts across paradigms or research areas. Finally, we've seen how we can use models to measure the core variables underlying observed performance, and use the correlations between these model variables and cognitive or even neuroscientific measures to examine individual differences and link brain and behaviour.

Exercises

1. Hintzman [63] gives the following finding from sociobiology as a target for modelling: "While adultery rates for men and women may be equalizing, men still have more partners than women do, and they are more likely to have one-night stands". Spend 10 min or so thinking about how to formalize this problem, and what factors might explain this discrepancy, before looking at the answer.
2. Try running your own replication of the initial simulation from Kenrick et al. [3]. This requires little mathematical knowledge, and primarily relies on a little programming knowledge. Be sure to update the strategies of individuals at time t based on the activities of their neighbours at time $t + 1$; the easiest way to accomplish this might be to have two matrices, one representing the strategies at time t, and the other representing the updated strategies at time $t + 1$. If you have no programming experience, but have some time and patience, you could simulate this model using a checkers board or pen and paper—just be careful not to introduce any human errors!
3. Having replicated the simulation, try using different starting patterns; these could be random or hand-coded. One interesting pattern to try is a striped arrangement, so that successive rows alternate between containing only peaceful individuals and only hostile individuals.
4. Finally, replicate Kenrick et al.'s [3] Fig. 4, showing how the number and placement of "short-fused" individuals affects the final stable state of the simulation. Using the same starting state, explore different placements of short-fused individuals. For example, if short-fused individuals are arranged across the diagonal of the grid, does this produce overall aggressive strategies like in panel c of Kenrick et al.'s Fig. 4?

Further Reading

1. Busemeyer, J. R., & Diederich, A. (2010). *Cognitive modeling*. Thousand Oaks, CA: Sage.
2. Farrell, S., & Lewandowsky, S. (2010). Computational Models as Aids to Better Reasoning in Psychology. *Current Directions in Psychological Science, 19*, 329-335.
3. Hintzman, D. L. (1991). Why are formal models useful in psychology? In W. E. Hockley & S. Lewandowsky (Eds.), *Relating theory and data: Essays on human memory in honor of Bennet B. Murdock* (pp. 39–56). Hillsdale, NJ: Lawrence Erlbaum.
4. Lewandowsky, S., & Farrell, S. (2011). *Computational Modeling in Cognition: Principles and Practice*. Thousand Oaks, CA: Sage.
5. Lewandowsky, S. (1993). The rewards and hazards of computer simulations. *Psychological Science, 4*, 236-243.
6. McClelland, J. L. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science, 1*, 11-38.

7. Norris, D. (2005). How do computational models help us build better theories? In A. Cutler (Ed.), *Twenty-first century psycholinguistics: Four cornerstones* (pp. 331–346). Mahwah, NJ: Lawrence Erlbaum.

References

1. Muter P (1980) Very rapid forgetting. Mem Cognit 8:174–179
2. Latané B (1996) Dynamic social impact: the creation of culture by communication. J Commun 46:13–25
3. Kenrick DT, Li NP, Butner J (2003) Dynamical evolutionary psychology: individual decision rules and emergent social norms. Psychol Rev 110:3–28
4. Cialdini RB, Goldstein NJ (2004) Social influence: compliance and conformity. Annu Rev Psychol 55:591–621. doi:10.1146/annurev.psych.55.090902.142015
5. Cialdini RB, Demaine LJ, Sagarin BJ, DW Barrett, K Rhoads, PL Winter (2006) Managing social norms for persuasive impact. Soc Influ 1:3–15
6. Schultz PW, Nolan JM, Cialdini RB, Goldstein NJ, Griskevicius V (2007) The constructive, destructive, and reconstructive power of social norms. Psychol Sci 18:429–434
7. Luce RD (1995) Four tensions concerning mathematical modeling in psychology. Annu Rev Psychol 46:1–26
8. Inglehart R, Foa R, Peterson C, Welzel C (2008) Development, freedom, and rising happiness. Perspect Psychol Sci 3:264–285
9. Heathcote A, Brown S, Mewhort DJ (2000) The power law repealed: the case for an exponential law of practice. Psychon Bull Rev 7:185–207
10. Myung IJ, Kim C, Pitt MA (2000) Toward an explanation of the power law artifact: insights from response surface analysis. Mem Cognit 28:832–840. doi:10.3758/BF03198418
11. Norris D (2005) How do computational models help us build better theories? In: Cutler A (ed) Twenty-first century psycholinguistics: four cornerstones. Lawrence Erlbaum, Mahwah, pp 331–346
12. Fum D, Missier FDel, Stocco A (2007) The cognitive modeling of human behavior: why a model is (sometimes) better than 10,000 words. Cognit Syst Res 8:135–142
13. Lewandowsky S (1993) The rewards and hazards of computer simulations. Psychol Sci 4:236–243
14. Lewandowsky S, Farrell S (2011) Computational modeling in cognition: principles and practice. Sage, Thousand Oaks
15. Logan GD (1988) Toward an instance theory of automatization. Psychol Rev 95:492–527
16. Rouder JN, Lu J, Speckman P, Sun D, Jiang Y (2005) hierarchical model for estimating response time distributions. Psychon Bull Rev 12:195–223
17. Logan GD (2002) An instance theory of attention and memory. Psychol Rev 109:376–400
18. Pashler H, Wagenmakers E (2012) Editors introduction to the special section on replicability in psychological science: a crisis of confidence? Perspect Psychol Sci 7:528–530. doi:10.1177/1745691612465253
19. Evans JSBT (1989) Bias in human reasoning: causes and consequences. Lawrence Erlbaum Associates, Hove
20. Anderson JR (1996) ACT: a simple theory of complex cognition. Am Psychol 51:355–365
21. Collins AM, Loftus EF (1975) A spreading activation theory of semantic processing. Psychol Rev 82:407–428
22. Radvansky G (2006) Human memory. Pearson, Boston
23. Gentner D, Gentner DR (1983) Flowing waters or teeming crowds: mental models of electricity. In: Gentner D, Stevens AL (ed) Mental models. Lawrence Erlbaum Associates, Hillsdale, pp 99–129
24. Ratcliff R, McKoon G (1981) Does activation really spread? Psychol Rev 88:454–462

25. Anderson JR (1983) A spreading activation theory of memory. J Verbal Learn Verbal Behav 22:261–295
26. Borst J, Anderson JR (2013) Using the ACT-R cognitive architecture in combination with fMRI data. In: Forstmann BU, Wagenmakers EJ (eds) An introduction to model-based cognitive neuroscience. Springer, New York, pp [this volume, needs updating]
27. Anderson JR, Matessa M (1997) A production system theory of serial memory. Psychol Rev 104:728–748
28. Pavlik PI, Anderson JR (2005) Practice and forgetting effects on vocabulary memory: an activation-based model of the spacing effect Cognit Sci 29:559–586
29. Pavlik PI, Anderson JR (2008) Using a model to compute the optimal schedule of practice. J Exp Psychol: Appl 14:101–117
30. Anderson JR, Schooler LJ (1991) Reflections of the environment in memory. Psychol Sci 2:396–408
31. Ashby FG (1992) Multidimensional models of perception and cognition. Lawrence Erlbaum, Hillsdale
32. Nosofsky RM (1986) Attention, similarity, and the identification-categorization relationship. J Exp Psychol Learn Mem Cognit 115:39–61
33. Maddox WT Ashby FG (1993) Comparing decision bound and exemplar models of categorization. Percept Psychophys 53(1):49–70
34. McKinley SC, R.M. Nosofsky (1995) Investigations of exemplar and decision bound models in large, ill-defined category structures. J Exp Psychol Hum Percept Perform 21:128–148
35. Nosofsky RM, Johansen M (2000) Exemplar-based accounts of "multiplesystem" phenomena in perceptual categorization. Psychon Bull Rev 7:375–402
36. Rouder JN, Ratcliff R (2004) Comparing categorization models. J Exp Psychol Gener 133:63–82
37. Farrell S, Ratcliff R, Cherian A, Segraves M (2006) Learn Behav 34:86
38. Craig S, Lewandowsky S (2012) Whichever way you choose to categorize, working memory helps you learn. Q J Exp Psychol 65:439–464
39. Lewandowsky S, Yang LX, Newell BR, Kalish ML (2012) Working memory does not dissociate between different perceptual categorization tasks. J Exp Psychol Learn Mem Cognit 38:881–904
40. Doll BB, Hutchison KE, Frank MJ (2011) Dopaminergic genes predict individual differences in susceptibility to confirmation bias. J Neurosci 31(16):6188–6198
41. Farrell S, Lelièvre A (2009) End anchoring in short-term order memory. J Mem Lang 60:209–227
42. Jang Y, Wixted J, Huber DE (2009) Testing signal-detection models of yes/no and two-alternative forced choice recognition memory. J Exp Psychol Gener 138:291–306
43. McDaniel M, Busemeyer J (2005) The conceptual basis of function learning and extrapolations: comparison of rule-based and associative-based models. Psychon Bull Rev 12(1):24–42
44. Brown GDA, Neath I, Chater N (2007) Amnesia, rehearsal, and temporal distinctiveness models of recall. Psychol Rev 114:539–260
45. Neath I, Brown G (2006) Simple: further applications of a local distinctiveness model of memory. Psychol Learn Motiv 46:201–243
46. Baddeley AD, Warrington EK (1970) Amnesia and the distinction between long- and short-term memory. J Verb Learn Verb Behav 9:176–189
47. Brown GDA, Della Salla S, Foster JK, Vousden JI (2007) Amnesia, rehearsal, and temporal distinctiveness models of recall. Psychon Bull Rev 14:256–260
48. Brown GDA, Lewandowsky S. (2010) Forgetting in memory models: arguments against trace decay and consolidation failure. In: Della Sala S (ed) Forgetting. Psychology Press, Hove, pp 49–75
49. Lewandowsky S, Ecker UKH, Farrell S, Brown GDA (2011) Models of cognition and constraints from neuroscience: a case study involving consolidation. Aust J Psychol 64:37–45
50. Kane MJ, Bleckley MK, Conway ARA Engle RW(2001) A controlled-attention view of working-memory capacity. J Exp Psychol Gener 130:169–183

51. Ratcliff R, Smith PL (2004) A comparison of sequential sampling models for two-choice reaction time. Psychol Rev 111:333–367
52. Batchelder W, Riefer D (1999) Theoretical and empirical review of multinomial process tree modeling. Psychon Bull Rev 6:57–86
53. Vandekerckhove J, Tuerlinckx F, Lee MD (2011) Hierarchical diffusion models for two-choice response times. Psychol Methods 16:44–62
54. Schmiedek F, Oberauer K, Wilhelm O, SüßHM, Wittmann WW (2007) Individual differences in components of reaction time distributions and their relations to working memory and intelligence. J Exp Psychol: Gener 136:414–429. doi:10.1037/0096-3445.136.3.414
55. Wagenmakers EJ, van der Maas HLJ, Grasman RPPP (2007) An EZ-diffusion model for response time and accuracy. Psychon Bull Rev 14:3–22
56. Lewandowsky S (2011) Working memory capacity and categorization: individual differences and modeling. J Exp Psychol Learn Mem Cognit 37:720–738
57. Shepard RN, Hovland CI, Jenkins HM (1961) Learning and memorization of classifications. Psychol Monogr 75:1–42 (13, Whole No. 517)
58. DeCaro MS, Thomas RD, Beilock SL (2008) Individual differences in category learning: sometimes less working memory capacity is better than more. Cognition 107:284–294
59. Kruschke JK (1992) ALCOVE: an exemplar-based connectionist model of category learning. Psychol Rev 99:22–44
60. Forstmann BU, Dutilh G, Brown S, Neumann J, von CramondDY, Ridderinkhofa KR, Wagenmakers EJ (2008) Striatum and pre-SMA facilitate decision-making under time pressure. Proc Natl Acad Sci U S A 105:17538–17542
61. Townsend JT (2008) Mathematical psychology: prospects for the 21st century: a guest editorial. J Math Psychol 52:269–280
62. Busemeyer JR, Diederich A (2010) Cognitive modeling. Sage, Thousand Oaks
63. Hintzman DL (1991) Why are formal models useful in psychology? In: Hockley WE, Lewandowsky S (eds) Relating theory and data: essays on human memory in honor of Bennet B. Murdock. Lawrence Erlbaum, Hillsdale, pp 39–56

Chapter 2
An Introduction to Good Practices in Cognitive Modeling

Andrew Heathcote, Scott D. Brown and Eric-Jan Wagenmakers

Abstract Cognitive modeling can provide important insights into the underlying causes of behavior, but the validity of those insights rests on careful model development and checking. We provide guidelines on five important aspects of the practice of cognitive modeling: parameter recovery, testing selective influence of experimental manipulations on model parameters, quantifying uncertainty in parameter estimates, testing and displaying model fit, and selecting among different model parameterizations and types of models. Each aspect is illustrated with examples.

2.1 Introduction

One of the central challenges for the study of the human mind is that cognitive processes cannot be directly observed. For example, most cognitive scientists feel confident that people can shift their attention, retrieve episodes from memory, and accumulate sensory information over time; unfortunately, these processes are latent and can only be measured indirectly, through their impact on overt behavior, such as task performance.

Another challenge, one that exacerbates the first, is that task performance is often the end result of an unknown combination of several different cognitive processes. Consider the task of deciding quickly whether an almost vertical line tilts slightly to the right or to the left. Even in this rather elementary task it is likely that at least four different factors interact to determine performance: (1) the speed with which perceptual processes encode the relevant attributes of the stimulus; (2) the

A. Heathcote (✉) · S. D. Brown
School of Psychology, University of Newcastle, University Avenue,
Callaghan, NSW 2308, Australia
e-mail: Andrew.Heathcote@newcastle.edu.au

S. D. Brown
e-mail: Scott.Brown@newcastle.edu.au

E. J. Wagenmakers
Department of Psychological Methods, University of Amsterdam,
Weesperplein 4, 1018 XA, Amsterdam, The Netherlands
e-mail: EJ.Wagenmakers@gmail.com

B. U. Forstmann, E.-J. Wagenmakers (eds.), *An Introduction to Model-Based Cognitive Neuroscience,* DOI 10.1007/978-1-4939-2236-9_2

efficiency with which the perceptual evidence is accumulated; (3) the threshold level of perceptual evidence that an individual deems sufficient for making a decision; and (4) the speed with which a motor response can be executed after a decision has been made. Hence, observed behavior (i.e., response speed and percentage correct) cannot be used blindly to draw conclusions about one specific process of interest, such as the efficiency of perceptual information accumulation. Instead, one needs to untangle the different cognitive processes and estimate both the process of interest and the nuisance processes. In other words, observed task performance needs to be decomposed in terms of the separate contributions of relevant cognitive processes. Such decomposition almost always requires the use of a cognitive process model.

Cognitive process models describe how particular combinations of cognitive processes and mechanisms give rise to observed behavior. For example, the linear ballistic accumulator model (LBA; [1]) assumes that in the line-tilt task there exist two accumulators—one for each response—that each race towards an evidence threshold. The psychological processes in the LBA model are quantified by parameters; for instance, the threshold parameter reflects response caution. Given the model assumptions, the observed data can be used to estimate model parameters, and so draw conclusions about the latent psychological processes that drive task performance. This procedure is called cognitive modeling (see Chap. 1 for details).

Cognitive modeling is perhaps the only way to isolate and identify the contribution of specific cognitive processes. Nevertheless, the validity of the conclusions hinges on the plausibility of the model. If the model does not provide an adequate account of the data, or if the model parameters do not correspond to the psychological processes of interest, then conclusions can be meaningless or even misleading. There are several guidelines and sanity checks that can guard against these problems. These guidelines are often implicit, unspoken, and passed on privately from advisor to student. The purpose of this chapter is to be explicit about the kinds of checks that are required before one can trust the conclusions from the model parameters. In each of five sections we provide a specific guideline and demonstrate its use with a concrete application.

2.2 Conduct Parameter Recovery Simulations

One of the most common goals when fitting a cognitive model to data is to estimate the parameters so that they can be compared across conditions, or across groups of people, illuminating the underlying causes of differences in behavior. For example, when Ratcliff and colleagues compared diffusion-model parameter estimates from older and younger participants, they found that the elderly were slower mainly due to greater caution rather than reduced information processing speed as had previously been assumed [2].

A basic assumption of investigations like these is adequate parameter recovery—that a given cognitive model and associated estimation procedure produces accurate and consistent parameter estimates given the available number of data points. For

standard statistical models there is a wealth of information about how accurately parameters can be recovered from data. This information lets researchers know when parameters estimated from data can, and cannot, be trusted. Models of this sort include standard statistical models (such as general linear models) and some of the simplest cognitive models (e.g., multinomial processing trees [3]).

However, many interesting cognitive models do not have well-understood estimation properties. Often the models are newly developed, or are new modifications of existing models, or sometimes they are just existing models whose parameter estimation properties have not been studied. In these cases it can be useful to conduct a parameter recovery simulation study. An extra advantage of running one's own parameter recovery simulation study is that the settings of the study (sample sizes, effect sizes, etc.) can be matched to the data set at hand, eliminating the need to extrapolate from past investigations. When implementing estimation of a model for the first time, parameter recovery with a large simulated sample size also provides an essential bug check.

The basic approach of a parameter recovery simulation study is to generate synthetic data from the model, which of course means that the true model parameters are known. The synthetic data can then be analysed using the same techniques applied to real data, and the recovered parameter estimates can be compared against the true values. This gives a sense of both the bias in the parameter estimation methods (accuracy), and the uncertainty that might be present in the estimates (reliability). If the researcher's goal is not just to estimate parameters, but in addition to discriminate between two or more competing theoretical accounts, a similar approach can be used to determine the accuracy of discrimination, called a "model recovery simulation". Synthetic data are generated from each model, fit using both models, and the results of the fits used to decide which model generated each synthetic data set. The accuracy of these decisions shows the reliability with which the models can be discriminated.

When conducting a parameter recovery simulation, it is important that the analysis methods (the model fitting or parameter estimation methods) are the same as those used in the analysis of real data. For example, both synthetic data and real data analyses should use the same settings for optimisation algorithms, sample sizes, and so on. Even the model parameters used to generate synthetic data should mirror those estimated from real data, to ensure effect sizes etc. are realistic. An exception to this rule is when parameter recovery simulations are used to investigate methodological questions, such as what sample size might be necessary in order to identify an effect of interest. If the researcher has in mind an effect of interest, parameter recovery simulations can be conducted with varying sizes of synthetic samples (both varying numbers of participants, and of data points per participant) to identify settings that will lead to reliable identification of the effect.

2.2.1 Examples of Parameter Recovery Simulations

Evidence accumulation models are frequently used to understand simple decisions, in paradigms from perception to reading, and short term memory to alcohol intoxication [4, 5, 6, 7, 8, 9]. The most frequently-used evidence accumulation models for analyses such as these are the diffusion model, the EZ-diffusion model, and the linear ballistic accumulator (LBA) model [10, 11, 1]. As the models have become more widely used in parameter estimation analyses, the need for parameter recovery simulations has grown. As part of addressing this problem, in previous work, Donkin and colleagues ran extensive parameter recovery simulations for the diffusion and LBA models [12]. A similar exercise was carried out just for the EZ diffusion model when it was proposed, showing how parameter estimates from that model vary when estimated from known data of varying sample sizes [11].

Donkin and colleagues also went one step further, and examined the nature of parameters estimated from wrongly-specified models [12]. They generated synthetic data from the diffusion model and the LBA model, and examined parameter estimates resulting from fitting those data with the other model (i.e., the wrong model). This showed that most of the core parameters of the two models were comparable—for example, if the non-decision parameter was changed in the data-generating model, the estimated non-decision parameter in the other model faithfully recovered that effect. There were, however, parameters for which such relationships did not hold, primarily the response-caution parameters. These results can help researchers understand when the results they conclude from analysing parameters of one model might translate to the parameters of the other model. They can also indicate when model-based inferences are and are not dependent on assumptions not shared by all models.

To appreciate the importance of parameter recovery studies, consider the work by van Ravenzwaaij and colleagues on the Balloon Analogue Risk Task (BART, [13]). On every trial of the BART, the participant is presented with a balloon that represents a specific monetary value. The participant has to decide whether to transfer the money to a virtual bank account or to pump the balloon, an action that increases the balloon's size and value. After the balloon has been pumped the participant is faced with the same choice again: transfer the money or pump the balloon. There is some probability, however, that pumping the balloon will make it burst and all the money associated with that balloon is lost. A trial finishes whenever the participant has transferred the money or the balloon has burst. The BART task was designed to measure propensity for risk-taking. However, as pointed out by Wallsten and colleagues, performance on the BART task can be influenced by multiple psychological processes [14]. To decompose observed behavior into psychological processes and obtain a separate estimate for the propensity to take risk, Wallsten and colleagues proposed a series of process models.

One of the Wallsten models for the BART task (i.e., "Model 3" from [14], their Table 2) has four parameters: α, β, γ^{+}, and μ. For the present purposes, the precise specification of the model and the meaning of the parameters is irrelevant (for a detailed description see [15, 14]). What is important here is that van Ravenzwaaij and colleagues conducted a series of studies to examine the parameter recovery

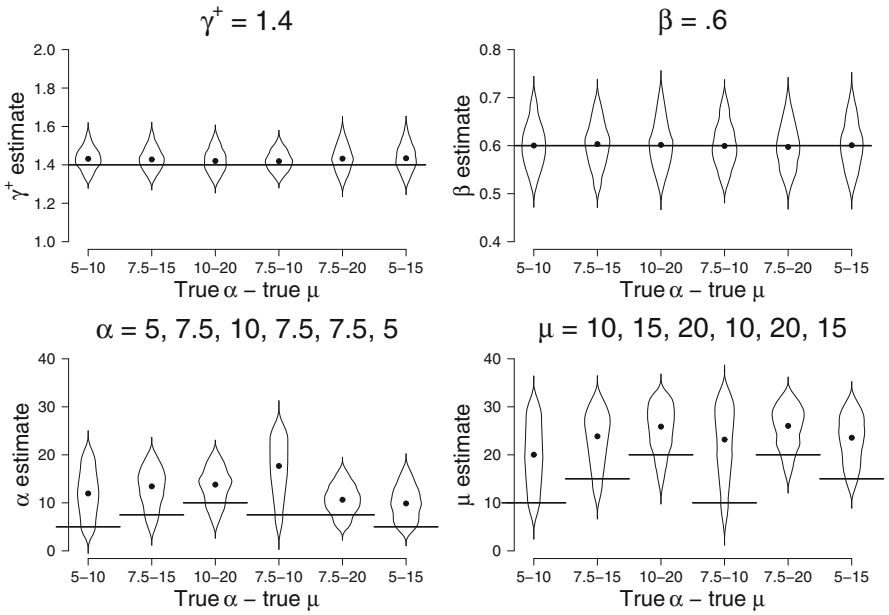


Fig. 2.1 The 4-parameter BART model recovers parameters γ^+ and β, but fails to recover parameters α and μ (results based on a 300-trial BART). The *dots* represent the median of 1000 point estimates from 1000 different BARTs performed by a single synthetic agent. The *violin shapes* around the *dots* are density estimates for the entire distribution of point estimates, with the extreme 5 % truncated [16]. The *horizontal lines* represent the true parameter values

for this model [15].[1] The results of one of those recovery studies are presented in Fig. 2.1. This figure shows the results of 1000 simulations of a single synthetic participant completing 300 BART trials[2], for each of six sets of data-generating parameter values. For each of the 1000 simulations, van Ravenzwaaij et al. obtained a point estimate for each parameter. In Fig. 2.1, the dots represent the median of the 1000 point estimates, and the "violins" that surround the dots represent density estimates that represent the entire distribution of point estimates, with the extreme 5 % truncated. The horizontal lines show the true parameter values that were used to generate the synthetic data (also indicated on top of each panel).

Figure 2.1 shows good parameter recovery for γ^+ and β, with only a slight overestimation of γ^+. The α and μ parameters are systematically overestimated. The overestimation of α increases when the true value of μ becomes smaller (in the bottom left panel, compare the fourth, second, and fifth violin from the left or compare the leftmost and rightmost violins). The overestimation of μ increases when the true value of α becomes larger (in the bottom right panel, compare the first and

[1] Extensive details are reported here: http://www.donvanravenzwaaij.com/Papers_files/BART_Appendix.pdf.

[2] With only 90 trials—the standard number—parameter recovery was very poor.

the fourth violin from the left). Both phenomena suggest that parameter recovery suffers when the true value of α is close to the true value of μ. For the six sets of data-generating parameter values shown on the x-axis from Fig. 2.1, the correlations between the point estimates of α and μ were all high: 0.97, 0.95, 0.93, 0.99, 0.83, 0.89, respectively.

The important lesson here is that, even though a model may have parameters that are conceptually distinct, the way in which they interact given the mathematical form of a model may mean that they are not distinct in practice. In such circumstances it is best to study the nature of the interaction and either modify the model or develop new paradigms that produce data capable of discriminating these parameters. The complete set of model recovery studies led van Ravenzwaaij and colleagues to propose a two-parameter BART model ([15]; but see [17]).

2.3 Carry Out Tests of Selective Influence

Cognitive models can be useful tools for understanding and predicting behavior, and for reasoning about psychological processes, but—as with all theories—utility hinges on validity. Establishing the validity of a model is a difficult problem. One method is to demonstrate that the model predicts data that are both previously unobserved, and ecologically valid. For example, a model of decision making, developed for laboratory tasks, might be validated by comparison against the decisions of consumers in real shopping situations. External data of this sort are not always available; even when they are, their ecological validity is not always clear. For example, it is increasingly common to collect neural data such as electroencephalography (EEG) or functional magnetic resonance imaging (fMRI) measurements simultaneously with behavioral data. Although it is easy to agree that the neural data should have some relationship to the cognitive model, it is not often clear what that relationship should be—which aspects of the neural data should be compared with which elements of the cognitive model.

An alternative way to establish model validity is via tests of selective influence. Rather than using external data as the benchmark of validity, this method uses experimental manipulations. Selective influence testing is based on the idea that a valid model can titrate complex effects in raw data into separate and simpler accounts in terms of latent variables. From this perspective, a model is valid to the extent that it make sense of otherwise confusing data. For example, signal detection models can explain simultaneous changes in false alarms and hit rates—and maybe confidence too—as simpler effects on underlying parameters (i.e., sensitivity and bias). Similarly, models of speeded decision-making can convert complex changes in the mean, variance, and accuracy of response time data into a single effect of just one latent variable.

Testing for selective influence begins with *a priori* hypotheses about experimental manipulations that ought to influence particular latent variables. For instance, from the structure of signal detection theory, one expects payoff manipulations to influence

bias, but not sensitivity. Empirically testing this prediction of selective influence becomes a test of the model structure itself.

2.3.1 Examples of Selective Influence Tests

Signal detection theory has a long history of checking selective influence. Nearly half a century ago, Parks [18] demonstrated that participants tended to match the probability of their responses to the relative frequency of the different stimulus classes. This behavior is called probability matching, and it is statistically optimal in some situations. Probability matching requires decision makers to adjust their decision threshold (in SDT terms: bias) in response to changes in relative stimulus frequencies. Parks—and many since—have demonstrated that decision-makers, from people to pigeons and rats, do indeed change their bias parameters appropriately (for a review, see [19]). This demonstrates selective influence, because the predicted manipulation influences the predicted model parameter, and only that parameter. Similar demonstrations have been made for changes in signal detection bias due to other manipulations (e.g., the strength of memories: [20]).

Models of simple perceptual decision making, particularly Ratcliff's diffusion model ([5, 21, 10]), have around six basic parameters. Their apparent complexity can be justified, however, through tests of selective influence. In seminal work, Ratcliff and Rouder orthogonally manipulated the difficulty of decisions and instructions about cautious vs. speedy decision-making, and demonstrated that manipulations of difficulty selectively influenced a stimulus-related model parameter (drift rate) while changes to instructions influenced a caution-related model parameter (decision boundaries). Voss, Rothermund and Voss [22] took this approach further and separately tested selective influences on the diffusion model's most fundamental parameters. For example, one experiment manipulated relative payoffs for different kinds of responses, and found selective influence on the model parameter representing bias (the "start point" parameter). These kinds of tests can alleviate concerns about model complexity by supporting the idea that particular model parameters are necessary, and by establishing direct relationships between the parameters and particular objective changes or manipulations.

Deciding whether one parameter is or is not influenced by some experimental manipulation is an exercise in model selection (i.e., selection between models that do and do not impose the selective influence assumption). Both Voss et al. and Ratcliff and Rouder approached this problem by estimating parameters freely and examining changes in the estimates between conditions; a significant effect on one parameter and non-significant effects on other parameters was taken as evidence of selective influence. Ho, Brown and Serences [23] used model selection based on BIC [24] and confirmed that changes in the response production procedure—from eye movements to button presses—influenced only a "non-decision time" parameter which captures the response-execution process. However, a number of recent studies have rejected the selective influence of cautious vs. speedy decision-making on decision boundaries [25, 26, 27]. In a later section we show how model-selection was used in this context.

2.4 Quantify Uncertainty in Parameter Estimates

In many modeling approaches, the focus is on model prediction and model fit for a single "best" set of parameter estimates. For example, suppose we wish to estimate the probability θ that Don correctly discriminates regular beer from alcohol-free beer. Don is repeatedly presented with two cups (one with regular beer, the other with non-alcoholic beer) and has to indicate which cup holds the regular beer. Now assume that Don answers correctly in 3 out of 10 cases. The maximum likelihood estimate $\hat{\theta}$ equals $3/10 = 0.3$, but it is evident that this estimate is not very precise. Focusing on only a single point estimate brings with it the danger of overconfidence: predictions will be less variable than they should be.

In general, when we wish to use a model to learn about the cognitive processes that drive task performance, it is appropriate to present the precision with which these processes have been estimated. The precision of the estimates can be obtained in several ways. Classical or frequentist modelers can use the bootstrap [28], a convenient procedure that samples with replacement from the original data and then estimates parameters based on the newly acquired bootstrap data set; the distribution of point estimates across the bootstrap data sets provides a close approximation to the classical measures of uncertainty such as the standard error and the confidence interval. Bayesian modelers can represent uncertainty in the parameter estimates by plotting the posterior distribution or a summary measure such as a credible interval.

2.4.1 Example of Quantifying Uncertainty in Parameter Estimates

In an elegant experiment, Wagenaar and Boer assessed the impact of misleading information on earlier memories [29]. They showed 562 participants a sequence of events in the form of a pictorial story involving a pedestrian-car collision at an intersection with a traffic light. In some conditions of the experiment, participants were later asked whether they remembered a pedestrian crossing the road when the car approached the "stop sign". This question is misleading (the intersection featured a traffic light, not a stop sign), and the key question centers on the impact that the misleading information about the stop sign has on the earlier memory for the traffic light.[3]

Wagenaar and Boer constructed several models to formalize their predictions. One of these models is the "destructive updating model", and its critical parameter d indicates the probability that the misleading information about the stop sign (when properly encoded) destroys the earlier memory about the traffic light. When $d = 0$, the misleading information does not affect the earlier memory and the destructive updating model reduces to the "no-conflict model". Wagenaar and Boer fit the destructive updating model to the data and found that the single best parameter estimate was $\hat{d} = 0$.

[3] The memory for the traffic light was later assessed by reminding participants that there was a traffic light at the intersection, and asking them to indicate its color.

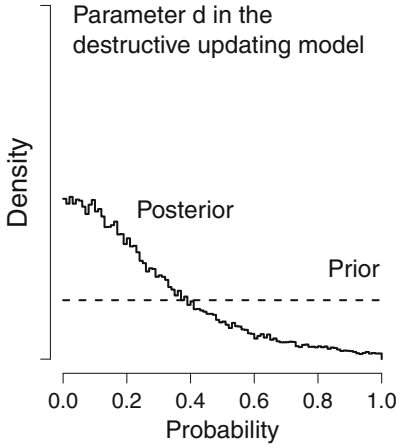


Fig. 2.2 Prior and posterior distributions for the d parameter in the destructive updating model from Wagenaar and Boer (1987), based on data from 562 participants. When $d = 0$, the destructive updating model reduces to the no-conflict model in which earlier memory is unaffected by misleading information presented at a later stage. The posterior distribution was approximated using 60,000 Markov chain Monte Carlo samples. (Figure downloaded from Flickr, courtesy of Eric-Jan Wagenmakers)

Superficial consideration may suggest that the result of Wagenaar and Boer refutes the destructive updating model, or at least makes this model highly implausible. However, a more balanced perspective arises once the uncertainty in the estimate of $\hat{d}$ is considered. Figure 2.2 shows the prior and posterior distributions for the d parameter (for details see [30]). The prior distribution is uninformative, reflecting the belief that all values of d are equally likely before seeing the data. The observed data then update this prior distribution to a posterior distribution; this posterior distribution quantifies our knowledge about d [31]. It is clear from Fig. 2.2 that the most plausible posterior value is $d = 0$, in line with the point estimate from Wagenaar and Boer, but it is also clear that this point estimate is a poor summary of the posterior distribution. The posterior distribution is quite wide and has changed relatively little compared to the prior, despite the fact that 562 people participated in the experiment. Values of $d < 0.4$ are more likely under the posterior than under the prior, but not by much; in addition, the posterior ordinate at $d = 0$ is only 2.8 times higher than the prior ordinate at value $d = 0$. This constitutes evidence against the destructive updating model that is is merely anecdotal or "not worth more than a bare mention" [32].[4]

In sum, a proper assessment of parameter uncertainty avoids conclusions that are overconfident. In the example of Wagenaar and Boer, even 562 participants were not sufficient to yield strong support for or against the models under consideration.

[4] Wagenaar and Boer put forward a similar conclusion, albeit not formalized within a Bayesian framework.

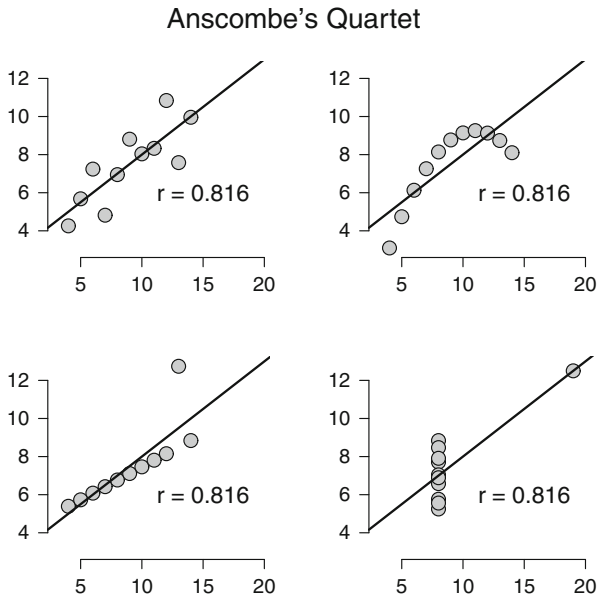


Fig. 2.3 Anscombe's quartet highlights the importance of plotting data to confirm the validity of the model fit. In each panel, the Pearson correlation between the x and y values is the same, $r = 0.816$. In fact, the four different data sets are also equal in terms of the mean and variance of the x and y values. Despite the equivalence of the four data patterns in terms of popular summary measures, the graphical displays reveal that the patterns are very different from one another, and that the Pearson correlation (a linear measure of association) is only valid for the data set from the top left panel. (Figure downloaded from Flickr, courtesy of Eric-Jan Wagenmakers)

2.5 Show Model Fit

When a model is unable to provide an adequate account of the observed data, conclusions based on the model's parameters are questionable. It is, therefore, important to always show the fit of the model to the data. A compelling demonstration of this general recommendation is known as Anscombe's quartet [33] replotted here as Fig. 2.3. The figure shows four data sets that have been equated on a number of measures: the Pearson correlation between the x and y values, the mean of the x and y values, and the variance of the x and y values. From the graphical display of the data, however, it is immediately obvious that the data sets are very different in terms of the relation between the x values and the y values. Only for the data set shown in the top left panel does it make sense to report the Pearson correlation (a linear measure of association). In general, we do not recommend relying on a test of whether a single global measure of model misfit is "significant". The latter practice is not even suitable for linear models [34], let alone non-linear cognitive process models, and is subject to the problem that with sufficient power rejection is guaranteed, and therefore meaningless [35]. Rather we recommend that a variety of

graphical checks be made and a graphical summary of the relevant aspects of model fit be reported.

Displaying and checking model fit can be difficult when data come from many participants in a complicated multiple-factor design. When the model is nonlinear, as is almost always the case with cognitive process models, fitting to data averaged over participants should be avoided, as even a mild nonlinearity can introduce systematic distortions (e.g., forgetting and practice curves [36, 37, 38]). For the purpose of displaying overall model fit it is fine to overlay a plot of the average data with the average of each participant's model fit, as both averages are subject to the same distortions. However, analogous plots should also be checked for each individual, both to detect atypical participants, and because it is common for initial fit attempts to fail with some participants. In some cases individual plots can reveal that an apparently good fit in an average plot is due to "cancelling out" of under- and over-estimation for different groups of participants. Similarly, it is important to check plots of fit broken down by all of the influential factors in the experimental design. Even when interest focuses on the effects of a subset of factors, and so it is appropriate to average over other (effectively "nuisance") factors when reporting results, such averages can hide tradeoffs that mask systematic misfit. Hence, in the first instance it is important to carry out a thorough check graphical check of fit broken down by all factors that produce non-negligible effects on data.

In the case of continuous data it is practically difficult to display large numbers of data points from many participants in complex designs. An approach often used with evidence accumulation model fit to continuous response time (RT) data is to summarize the distribution of RT using quantiles (e.g., the median and other percentiles). A common choice is the 10th, 30th, 50th, 70th and 90th percentiles (also called the 0.1, 0.3, 0.5, 0.7 and 0.9 quantiles). This five-quantile summary may omit some information, but it can compactly capture that are usually considered key features of the data. Of particular importance are the 10th percentile, which summarises the fastest RTs, the 50th percentile or median, which summarises the central tendency, and the 90th percentile, which summaries the slowest RTs. The spread between the 90th and 10th percentiles summarises variability in RT and a larger difference between the 90th and 50th percentiles compared to the 50th to 10th percentile summarises the typically positive skew in RT distribution.

Further complication arises when data are multivariate. For example, cognitive process models are usually fit to data from choice tasks. Where one of two choices is classified as correct, the rate of accurate responding provides a sufficient summary. However, participants can trade accuracy for speed [39], so in many cases it is important to also take RT into account. That is, the data are bivariate, consisting of an RT distribution for correct responses, an RT distribution for error responses, and an accuracy value specifying the rate at which correct and error responses occur. Latency-probability (LP) plots [40, 41] deal with the bivariate nature of choice data by plotting mean RT on the y-axis against response probability on the x-axis. As error responses commonly occur with low probability, error data appear on the left of the plot and correct response data on the right of the plot. In the two-choice case the x-values occur in pairs. For example, if the error rate is 0.1 then the

corresponding correct-response data must be located at 0.9. Quantile-probability (QP) plots [42] generalize this idea to also display a summary of RT distribution by plotting quantiles on the y-axis (usually the five-quantile summary) instead of the mean. Although the QP plot provides a very compact representation of choice RT data that can be appropriate in some circumstances, we do not recommend it as a general method of investigating model fit for reasons we illustrate in the following example. Rather, we recommend looking at separate plots of accuracy and correct and error RT distributions (or in the n>2 alternative case, RT distributions for each type of choice).

2.5.1 Examples of Showing Model Fit

Wagenmakers and colleagues [43] had participants perform a lexical decision task—deciding if a letter string constituted a word or nonword, using high, low and very-low frequency word stimuli and nonword stimuli. In their first experiment participants were given instructions that emphasised either the accuracy or speed of responding. They fit a relatively simple 12-parameter diffusion model to these data, assuming that instructions selectively influenced response caution and bias, whereas stimulus type selectively influenced the mean drift rate. Rae and colleagues [44] refit these data, including two extra participants not included in the originally reported data set (17 in total), in order to investigate the selective influence assumption about emphasis. They fit a more flexible 19-parameter model allowing (1) emphasis to affect the trial-to-trial standard deviation of bias as well as the mean and standard deviation of non-decision time; (2) stimulus type to affect the trial-to-trial standard deviation of the drift rate; and (3) allowing for response contamination. Their interest was in whether instruction emphasis could affect drift rate parameters, so they contrasted this 19 parameter "selective influence" model with a 27-parameter ("least constrained") model allowing speed emphasis to affect the mean and standard deviation of drift rates. We discuss this contrast in a following section but for now we focus on the fit of the selective-influence model.

Figure 2.4 is a quantile-probability plot of the selective-influence model. Data points are plotted with 95 % confidence intervals based on conventional standard errors assuming a normal distribution in order to convey an idea of the likely measurement error, and model fit is indicated by points joined by lines. Uncertainty can also be conveyed by other means, such as bootstrap methods [28] applied to the data [43] or model fits [45]. In any case, it is important to plot points for the model as well as the data; plotting only lines for either can hide mis-fit because the eye can be fooled by intersections that do not reflect an accurate fit. The figure demonstrates the utility of QP plots in illustrating an important regularity in choice RT data [46]; the overall decreasing lines from left to right in the accuracy condition show that errors are slower than corresponding correct responses, whereas the symmetric lines around $accuracy = 50\,\%$ in the speed condition indicate approximately equal correct and error speed.

Overall, Fig. 2.4 demonstrates that the model captures the majority of trends in the data, with many of the fitted points falling within 95 % data confidence intervals.

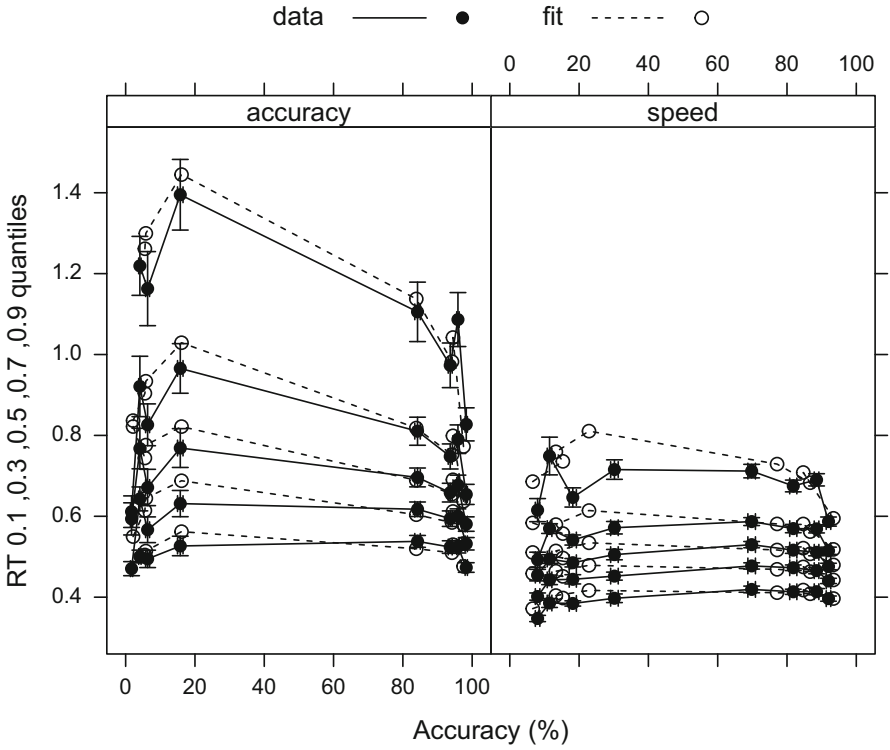


Fig. 2.4 Quantile-probability plot of the average over 17 participants from Wagenmakers and colleagues' [43] Experiment 1 and fits of the "selective influence" model. In both emphasis conditions accuracy was ordered from greatest to least: high-frequency (hf) words, nonwords (nw), low-frequency words (lf) and very-low-frequency (vlf) words. Each data point is accompanied by a 95 % confidence interval assuming a Student t distribution and based between-subject standard errors calculated as $SD(x)/\sqrt{n}$, where $SD(x)$ is the standard deviation over participants and n is the number of participants

However there is also some evidence of misfit, especially in regard to accuracy in the speed condition. Rae and colleagues [44] focused on this failure of the selective-influence model to account for the effect of emphasis instructions, motivated by similar findings for the LBA model [48], and the same pattern of under-estimation in experiments they reported using perceptual stimuli and in recognition memory (see also [49]). Figure 2.5 more clearly illustrates the speed-accuracy tradeoff induced by emphasis instructions—with accuracy displayed in the upper panels and speed of correct responses in the lower panels—and the under-estimation of the accuracy effect in the lexical decision data. For clarity, data from each stimulus type is plotted in a separate panel with emphasis condition plotted on the x-axis and joined by lines in order to emphasise the difference of interest. Each row of panels is tailored to examine a different aspect of the data. The upper panels show accuracy and the middle panels the distribution of correct RT. The lower panels plot both the central

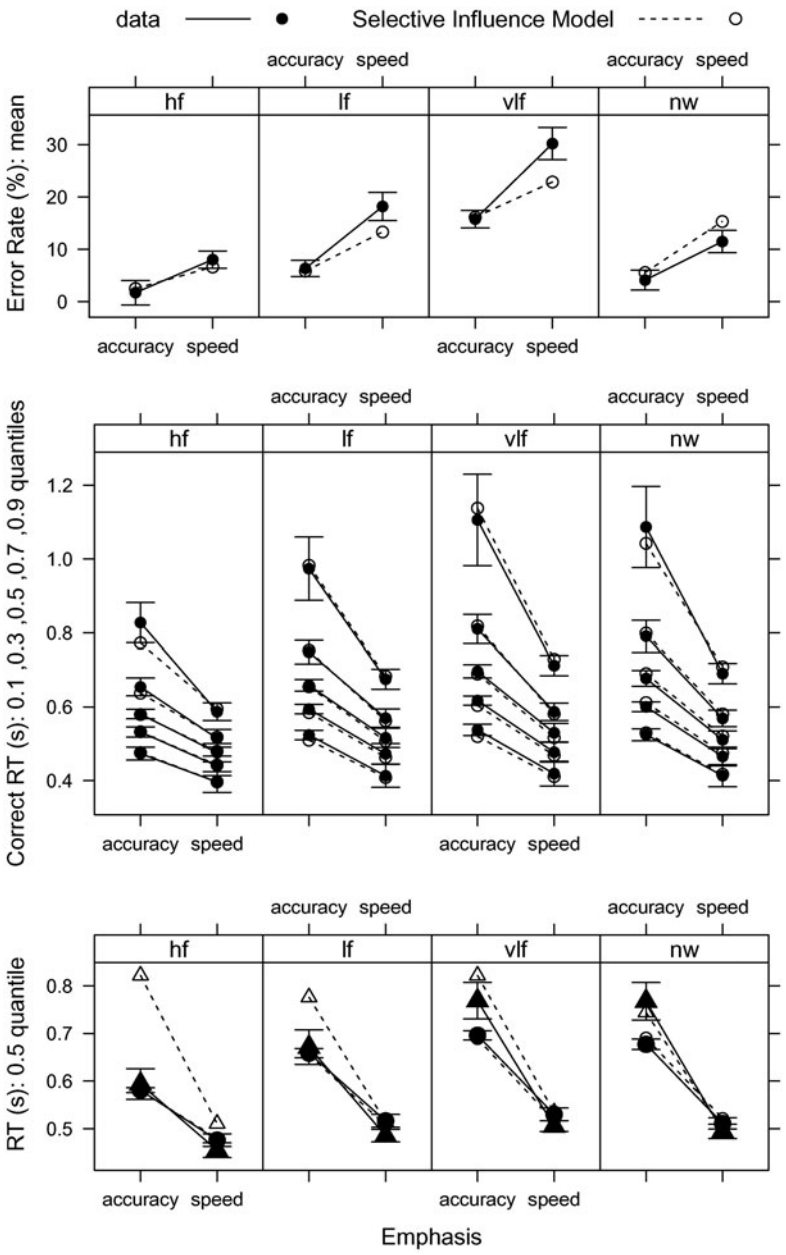


Fig. 2.5 Accuracy (*upper panels*), correct RT distribution (*middle panels*), median correct (*circle symbols*) and error (*triangle symbol*) RT (*lower panels*) plots of the average over 17 participants from Wagenmakers and colleagues' [43] Experiment 1 for high-frequency (*hf*), low-frequency (*lf*) and very-low-frequency (*vlf*) words and nonwords (*nw*). Each data point is accompanied by a 95 % confidence interval assuming a Student t distribution and based on within-subject standard errors calculated as using the bias-corrected method described in [47]

tendency (median) of correct RT (circle symbols) and error RT (triangle symbols) in order to highlight the relative speed of incorrect responses.

Figure 2.5 also uses within-subject standard errors appropriate to the focus on the (within-subject) difference between speed and accuracy emphasis. These standard errors reflect the reliability of the difference between speed and accuracy conditions, making it clear that the selective-influence model is unable to account for the effect of emphasis, particularly for very-low frequency and low-frequency words. The middle panels make it clear that, in contrast, this model provides a highly accurate account of its effect on RT distributions for correct responses, even for the 90th percentile, which has much greater measurement error than other quantiles. Finally, the lower panel clearly shows that the model predicts slow errors in the accuracy condition for all stimuli, whereas slow errors occur in the data only for the least word-like (very-low frequency and nonword) stimuli. In general, we recommend that a variety of multi-panel plots such as those in Fig. 2.5 be examined, each tailored to particular aspects of the data, with standard errors appropriate to the questions of interest. We also highly recommend plotting versions of these plots for individual participants, which is made easy using trellis graphics [50].

2.6 Engage in Model Selection

Cognitive models typically have several parameters that can sometimes interact in complex ways. How does a modeler decide which experimental design factors affect each parameter? Initial guidance is provided by conventions based on *a priori* assumptions and past research. In the realm of evidence accumulation models, for example, it has been widely assumed that stimulus-related factors selectively affect parameters related to the evidence flowing into accumulators (e.g., evidence accumulation rates) whereas instruction-related factors (e.g., an emphasis on speed vs. accuracy) affect accumulator-related parameters (e.g., bias and the amount of evidence required to make a decision) [46]. However, such conventional settings may not always hold, and in new paradigms they may not be available. Hence, it is prudent, and sometimes necessary, to engage in model selection: comparing a variety of different model parameterisations (variants) so that one or more can be selected and differences in parameter estimates among experimental conditions interpreted.

Even in relatively simple designs the number of model variants can rapidly become unmanageable. Suppose there are two experimental design factors, A and B. Each parameter might have a single estimated value (i.e., an intercept only model, often denoted ~1), a main effect of or A or B (~A or ~B), or both main effects and their interaction (denoted ~A*B = A + B + A:B, where "+" indicates an additive effect and ":" an interaction effect). If only parameterisations of this sort are considered there are 2^f models to select amongst, where f is the number of factors. If each of m types of model parameter are allowed this freedom then the total number of variants is $2^{f \times m}$. One might also consider additive models (e.g., ~A + B), in which case it is important to note that fit of an additive model depends on parameter scale (e.g., the

model ˜A + B can fit differently for a linear vs. log-scaled parameter, whereas A*B will fit identically). Further, so far we have only considered hierarchical models, where the higher terms can only occur accompanied by their constituents. If all possible non-hierarchical models are allowed (e.g., ˜A + A:B and ˜B + A:B) the increase in the number of variants with f is much faster.

Once an appropriate set of model variants is selected its members must be compared in some way. Ideally misfit is quantified by a function of the likelihood of the data under each model variant, such as in Bayesian or maximum likelihood estimation methods, or some approximation, such as in quantile-based methods like maximum probability [51, 52, 53] or minimum likelihood-ratio χ^2 (i.e., G^2) estimation [49]. As the number of estimated parameters increases a model becomes more flexible and so is able to better fit data. In particular, a nested model—a model variant that is a special case obtained by fixing one or more parameters of a more complex model–necessarily has greater misfit than models that nest it. The best model variant cannot be selected based on goodness-of-fit alone, as the least-constrained model would always be selected, and over-fitting (i.e., capturing unsystematic variation specific to a particular data set) would be rife.

One approach–similar to sequential regression methods such as stepwise selection– is to choose a model based on the significance of changes in fit as parameters are added or deleted. The maximised log-likelihood (L) is convenient for this purpose as the deviance misfit measure (D = −2×L) has a χ^2 distribution.[5] However, this approach is limited to selection amongst nested models. Preferable approaches use a model selection criterion that includes a penalty for model complexity; the model with the lowest criterion (i.e., least penalised misfit) is selected. Such criteria can be used not only to compare non-nested variants of the same model but also to compare different cognitive process models. In a Bayesian framework the Bayes factor is the ideal criterion [54], but this is rarely easy to directly compute for cognitive process models (although see [55, 56]). When estimation is achieved by posterior sampling, DIC [57], or its bias-corrected variant BPIC [58], are easy-to-compute alternatives. With maximum likelihood estimation it is convenient to use BIC = D + k× log (n), a Bayes factor approximation, or AIC = D + 2×k (n is the number of data points and k is the number of parameters) [59]. As is evident from these formulae, BIC applies a harsher penalty for complexity for typical sample sizes ($n \geq 8$).

Although we recommend using penalised-misfit criteria to guide model selection we do not recommend rigid adherence. Different criteria are optimal under different assumptions and they are often based on approximations that can be sensitive to the size of effects and samples (see, for example, the comparison of BIC and AIC in [60]). Further, it is seldom possible to check all possible models, and when the data-generating model is omitted model selection may err. Suppose, for example,

[5] The absolute value of the deviance depends on the measurement units for time and so only relative values of deviance are meaningful. Deviance is on an exponential scale, and as a rule of thumb a difference less than 3 is negligible and and difference greater than 10 indicates a strong difference.

the data were generated by an additive model, A + B but only the A, B, and A*B models are fit. Depending on the size of the A, B, and A:B effects relative to the size of the complexity penalty any of these three models may be selected. Even if the A + B model is fit, things can still go wrong if the data-generating model is additive on, for example, a logarithmic scale. Given the appropriate scale is usually not known, it is apparent that it is difficult to be absolutely sure that data-generating model is included even in quite exhaustive sets of variants (although clearly the chance of problems reduces when selection is among a large set of variants!).

In short, some judgement—albeit aided by a number of sources of evidence– must be used in model selection. For example, in smaller samples BIC can often select a model that is so simple that plots of fit reveal that the model is unable to account for practically or theoretically important trends in the data. On the other end of the spectrum over-fitting is indicated when model parameters are unstable (i.e., take on implausibly large or small values and/or values that can vary widely with little effect on fit) or take on patterns that appear nonsensical in relation to the way the parameter is interpreted psychologically. In both cases it is prudent to consider alternative model selection methods or possibly to seek further evidence. It is also worth reconsidering whether selection of a single model is required for the purpose at hand. For example, predictions averaged over models weighted by the evidence for each model are often better than predictions made by a single model [35]. Similarly, different criteria may select models that differ in some ways but are consistent with respect to the theoretical issue under investigation. We illustrate the process in the next section.

2.6.1 Examples of Model Selection

Our examples again use the lexical decision data from Wagenmakers and colleagues [43], focusing on variant selection assuming a diffusion model [44]. The example is based on maximum-likelihood fits to individual participant data, with $2^9 = 512$ diffusion variants fit with the methods described in [61]. Diffusion fits were based on quantile data, so the likelihood is only approximate [52, 53]. The least-constrained variant allowed rate parameters to vary with emphasis, that is, it did not make the conventional assumption that accumulation rate cannot be affected by instructions.

Before examining the model selection process in these examples it is important to address issues that can arise due to individual differences. When each participant's data are fit separately, different models are often selected for each participant. Considering participants as random effects provides a useful perspective on this issue. Even if there is no effect of a factor on the population mean of a parameter, when the population standard deviation is sufficiently large individual participants will display reliable effects of the factor. That is, selecting models that include the effect of a factor for some individuals does not imply that factor affects the corresponding population mean. Hierarchical models—which make assumptions about the form of population distributions– enable simultaneous fitting of all participants and direct estimation of population means. Even in this approach, however, individual participant estimates must be examined to check assumptions made by the hierarchical

model about the form of the population distribution. For example, it is possible that some individual variation results from participants being drawn from different populations (e.g., a mixture model where in one population a factor has an effect and in another it does not), in which case assuming a single unimodal population distribution is problematic. Caution must also be exercised in case the random effects model is incorrectly specified and the shrinkage (i.e., the averaging effect exerted by the assumed population distribution) masks or distorts genuine individual differences. Hierarchical modeling is best applied to relatively large samples of participants and usually requires Bayesian methods. These methods can sometimes be difficult in practice with cognitive process models where strong interactions among parameter make posterior sampling very inefficient, as was the case for the LBA model until recent advances in Markov chain Monte Carlo methods [62].

With maximum-likelihood fits to individuals it is possible to select an overall model based on and aggregate BIC or AIC.[6] The selected model, which can be thought of as treating participants as fixed effects, is usually sufficiently complex to accommodate every individual. However, it is important to be clear that selecting a model that contains a particular factor does not necessarily imply an effect of that factor on the random effect population mean. In the individual-fitting approach such questions can be addressed by testing for differences over a factor in the means of individual participant parameter estimates. These approaches to individual-participant fitting were taken by Rae and colleagues [44], with results summarised in Table 2.1. The table reports aggregate misfit measures minus the minimum deviance. Hence the best-fitting model (necessarily the least-constrained model with the largest number of parameters, k) has a zero entry in the deviance column.

The top three rows in Table 2.1 report results for the least constrained diffusion model and the models selected from the full set of 512 by aggregate AIC and BIC. The bottom three rows report results for the variant within the full set that imposes a minimal selective effect assumption—that the emphasis manipulation cannot affect rate parameters (v and sv)—and the AIC and BIC selected models among the subset of $2^7 = 128$ variants nested by the selective influence variant. Among the full set of 512 variants, AIC selected a variant where emphasis did affect the mean rate (i.e., violating selective influence), whereas BIC selected a variant that had no influence of emphasis on rate parameters. This pattern of results nicely exemplifies that fact that selection criteria can lead to theoretically important differences in conclusions, requiring researchers to seek other sources of evidence.

Rae and colleagues pointed out that the penalty for complexity imposed by BIC was likely too harsh. As shown in the upper panels of the Fig. 2.5 even the full 25-parameter selective influence LBA variant (which necessarily fits better than the 17 parameter variant selected from the overall set by BIC) fails to accommodate the effect of emphasis on accuracy. In agreement, there is a highly significant decrease

[6] Note that deviance can be summed over participants, as can AIC, but BIC cannot, due to the nonlinear $\log(n)$ term in its complexity penalty. Instead the aggregate BIC is calculated from the deviance, number of parameters and sample size summed over participants

Table 2.1 Diffusion model variants specified by design factors that effect each parameter, number of parameters per participant for each variant (k), and misfit measures (D = deviance, AIC = Akaike Information Criterion and BIC = Bayesian Information Criterion) minus the minimum value in each column. Factors are emphasis (E: speed vs. accuracy) and stimulus (S: high/low/very-low frequency words and non words). Diffusion parameters: a = response caution parameter, distance between response boundaries; accumulation rate parameters, v = mean, sv = standard deviation; start-point (bias) parameters, z = mean relative to lower boundary, sz = uniform distribution width; non-decision time parameters: t_0 = minimum time for stimulus encoding and response production, st = width of uniform distribution of non-decision time. Note that, for example, the notation v ~ E*S means that the v parameter can be affected by the main effects of the E and S factors as well as their interaction

Model type	Model definition	k	D	AIC	BIC
Least constrained	a ~ E, v ~ E*S, sv ~ E*S, z ~ E, sz ~ E, t_0 ~ E, st ~ E	27	0	70	984
AIC selected	a ~ E, v ~ E*S, sv ~ S, z ~ E, sz ~ E, t_0 ~ E, st ~ E	23	66	0	552
BIC selected	a ~ E, v ~ S, sv ~ 1, z ~ 1, sz ~ E, t_0 ~ E, st ~ 1	14	635	263	0
Selective influence	a ~ E, v ~ S, sv ~ S, z ~ E, sz ~ E, t_0 ~ E, st ~ E	19	237	35	225
AIC selected	a ~ E, v ~ S, sv ~ S, z ~ E, sz ~ 1, t_0 ~ E, st ~ E	19	237	35	225
BIC selected	a ~ E, v ~ S, sv ~ 1, z ~ E, sz ~ 1, t_0 ~ E, st ~ 1	14	635	263	0

in fit from the least-constrained to the BIC model as illustrated by the difference of deviances in Table 2.1 (i.e., df $= 17 \times (27 - 19) = 136$, $\chi^2(136) = 237$, $p < 0.001$). In contrast, the 19-parameter variant selected from the overall set by AIC that allows emphasis to affect mean rate does much better in accommodating the effect of emphasis on accuracy. Further, the reduction in fit relative to the least constrained model does not approach significance (i.e., df $= 17 \times (27 - 23) = 68$, $\chi^2(68) = 66$, $p = 0.55$). Finally, parameter estimates for the least constrained model were stable and there was a significant effect of emphasis on mean rate. This overall pattern of results confirmed a failure of traditional selective influence assumption and was consistent with findings from perceptual an mnemonic paradigms reported by Rae and colleagues and others [26].

In closing we return to what is perhaps one of the most difficult questions in cognitive process modeling, absolute fit, which might be couched as a model-selection question: "when should all model variants be rejected"? A model may provide a very accurate account of some conditions or some aspects of data but systematically misfit in other cases. For example, Rae and colleagues found even the least-constrained diffusion model misfit error RT distribution in some conditions. Heathcote and Love [48] showed the same was true in those data, although to a lesser degree, for the LBA model. Should both the diffusion and LBA models be rejected? Clearly some judgement is required to decide, since some misfit can be forgiven. A case in point for evidence-accumulation models is sequential effects. Sequential effects can be quite strong in the sorts of paradigms to which such models are fit [63, 64], but that fitting almost invariably assumes data are independently distributed over trials.

On this issue we think Box's [65] famous dictum that "all models are false but some are useful" is salutary. That is, some misfit can be tolerated, especially when no

better alternatives are available, as long as the model captures theoretically important features of a data set. To the degree this happens, parameter estimates are likely to provide an accurate distillation of the data and a more meaningful characterisation than simple summary statistics (e.g., mean RT or accuracy alone can be confided by speed-accuracy tradeoff). Further, if that distillation is sensible in terms of the underlying rationale of the model, and consistent with conclusions based on alternative analyses that do not reply on the model, then it is likely that the cognitive process model has served a useful role.

2.7 Concluding Remarks

Good standards are important in all areas of science. In cognitive modeling, good standards include not only careful model development and checking but also transparency of method and, ideally, sharing of model code and data. Transparency is obviously of key importance, as it allows interested colleagues to implement the model at hand, whether for teaching, for testing, or for extending the approach to other contexts. Even relatively simple models can sometimes be surprisingly difficult to replicate because crucial information is missing. It is therefore common practice to make available the model code, either on a personal website, archived together with the journal publication as supplementary material, or in a public repository (e.g., the OSU Cognitive modeling Repository, http://cmr.osu.edu/).

It is useful to make the model code available with several example data sets so that the new user can confirm that the code works as it should. Ideally, the entire data set that is modeled is made freely available online as well. This benefits the new user (who may be able to use the data set for a different purpose, or for a test of alternative explanations), but it also benefits the author, as the online data set may easily become a modeling benchmark.

In this introductory chapter we have discussed a suite of standard sanity checks that every modeler should turn to on a regular basis. This holds especially true for cognitive process models that are relatively new and untested. By applying the suite of sanity checks the modeler can gain confidence in the validity of a model, and consequently make a more compelling case that the model yields a reliable connection from observed behavior to unobserved psychological process.

Exercises

1. You fit a model of task switching to some data. The model includes a parameter which reflects how often people actively prepare for the upcoming task, and you find that the best estimate of this parameter is 50 %. What should you also consider, before concluding that task preparation occurs half of the time?
2. If you fit a complex model and a simpler model to some data, and found that the simple model had best BIC but the complex model had the best AIC, which would you expect to give the closest fit to data? And how could you resolve the tension between the two criteria?
3. You examine data plots with panels showing the probability of making different choices and for each choice the median, 10th, and 90th percentiles of the time to make each choice. What characteristics of which plot would tell you about the the average time to make a response and the variability in response times? What measurement derived from the plot tells you about the skew of the response time distributions? What relationship between the plots would be indicative of a speed-accuracy tradeoff?
4. An easy approach to model selection is to construct strong prior assumptions about which experimental effects will influence which parameters, effectively ruling out all other model variants. For example, one might make the prior assumption that a manipulation of stimulus strength can influence only sensitivity in a signal detection model, and not criterion placement. Name one danger of this method.
5. If you conducted a parameter recovery simulation for your new cognitive model, and found that there was unacceptably large variance in the recovered parameters (i.e., large inaccuracies that vary randomly with each new synthetic data set), what might you do?

Further Reading

1. Here are four course books on cognitive modeling, take your pick: Lewandowsky and Farrell [66], Busemeyer and Diederich [67], Hunt [68], and Polk and Seifert [69].
2. A hands-on, Bayesian approach to cognitive modeling is presented in Lee & Wagenmakers [31]; see also www.bayesmodels.com.
3. The series of tutorial articles in the *Journal of Mathematical Psychology* are a good source of information that is relatively easy to digest.

References

1. Brown SD, Heathcote AJ (2008) The simplest complete model of choice reaction time: linear ballistic accumulation. Cognit Psychol 57:153–178
2. Ratcliff R, Thapar A, McKoon G (2006) Aging, practice, and perceptual tasks: a diffusion model analysis. Psychol Aging 21:353–371
3. Riefer DM, Knapp BR, Batchelder WH, Bamber D, Manifold V. (2002) Cognitive psychometrics: assessing storage and retrieval deficits in special populations with multinomial processing tree models. Psychol Assess 14:184–201
4. Mulder M., Wagenmakers EJ, Ratcliff R, Boekel W, Forstmann BU (2012) Bias in the brain: a diffusion model analysis of prior probability and potential payoff. J Neurosci 32:2335–2343
5. Ratcliff R (1978) A theory of memory retrieval. Psychol Rev 85:59–108
6. Ratcliff R, Gomez P, McKoon G (2004) Diffusion model account of lexical decision. Psychol Rev 111:159–182
7. Ratcliff R, Starns JJ (2009) Modeling confidence and response time in recognition memory. Psychol Rev 116:59–83
8. Ratcliff R, van Dongen HPA (2009) Sleep deprivation affects multiple distinct cognitive processes. Psychon Bull Rev 16:742–751
9. Smith PL, Ratcliff R (2004) The psychology and neurobiology of simple decisions. Trends Neurosci 27:161–168
10. Ratcliff R, Tuerlinckx F (2002) Estimating parameters of the diffusion model: approaches to dealing with contaminant reaction times and parameter variability. Psychon Bull Rev 9:438–481
11. Wagenmakers EJ, van der Maas HJL, Grasman RPPP (2007) An EZ–diffusion model for response time and accuracy. Psychon Bull Rev 14:3–22
12. Donkin C, Brown S, Heathcote A, Wagenmakers EJ (2011)Different models but the same conclusions about psychological processes? Psychon Bull Rev 18:61–69
13. Lejuez CW, Read JP, Kahler CW, Richards JB, Ramsey SE, Stuart GL, Strong DR, Brown RA (2002) Evaluation of a behavioral measure of risk taking: the balloon analogue risk task (BART). J Exp Psychol Appl 8:75–84
14. Wallsten TS, Pleskac TJ, Lejuez CW (2005) Modeling behavior in a clinically diagnostic sequential risk–taking task. Psychol Rev 112:862–880
15. van Ravenzwaaij D, Dutilh G, Wagenmakers EJ (2011) Cognitive model decomposition of the BART: assessment and application. J Math Psychol 55:94–105
16. Hintze JL, Nelson RD (1998) Violin plots: a box plot–density trace synergism. Am Stat 52:181–184
17. Rolison JJ, Hanoch Y, Wood S (2012) Risky decision making in younger and older adults: the role of learning. Psychol Aging 27:129–140
18. Parks TE (1966) Signal-detectability theory of recognition-memory performance. Psychol Rev 73(1):44–58
19. Macmillan NA, Creelman CD (2005) Detection theory: a user's guide, 2nd edn. Erlbaum, Mahwah
20. Wixted JT, Stretch V (2000) The case against a criterion-shift account of false memory. Psychol Rev 107:368–376
21. Ratcliff R, Rouder JN (1998) Modeling response times for two-choice decisions. Psychol Sci 9:347–356
22. Voss A, Rothermund K, Voss J (2004) Interpreting the parameters of the diffusion model: an empirical validation. Mem Cognit 32:1206–1220
23. Ho TC, Brown S, Serences JT (2009) Domain general mechanisms of perceptual decision making in human cortex. J Neurosci 29:8675–8687
24. Schwarz G (1978) Estimating the dimension of a model. Annals Stat 6:461–464
25. Lee M, Vandekerckhove J, Navarro DJ, Tuerlinckx F (2007) Presentation at the 40th Annual Meeting of the Society for Mathematical Psychology, Irvine, USA, July 2007

26. Starns JJ, Ratcliff R, McKoon G (2012) Evaluating the unequal-variance and dualprocess explanations of zROC slopes with response time data and the diffusion model. Cogniti Psychol 64:1–34
27. Rae B, Heathcote C, Donkin A, Averell L, Brown SD (in press) The hare and the tortoise: emphasizing speed can change the evidence used to make decisions. J Exp Psychol Learn Mem Cogn
28. Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman & Hall, New York
29. Wagenaar WA, Boer JPA (1987) Misleading postevent information: testing parameterized models of integration in memory. Acta Psychol 66:291–306
30. Vandekerckhove J, Matzke D, Wagenmakers EJ 2013) In Busemeyer J, Townsend J, Wang ZJ, Eidels A (eds) Oxford handbook of computational and mathematical psychology. Oxford University Press
31. Lee MD, Wagenmakers EJ (in press) Bayesian modeling for cognitive science: a practical course. Cambridge University Press
32. Jeffreys H (1961) Theory of probability, 3rd edn. Oxford University Press, Oxford
33. Anscombe FJ (1973) Graphs in statistical analysis. Am Stat 27:17–21
34. McCullagh PN, Nelder JA (1983) Generalized linear models. Chapman & Hall, London
35. Raftery AE (1995) Bayesian model selection in social research. Sociol Methodol 25:111–164
36. Averell L, Heathcote A (2011) The form of the forgetting curve and the fate of memories. J Math Psychol 55:25–35
37. Brown S, Heathcote A (2003) Averaging learning curves across and within participants. Behav Res Methods Instrum Comput 35:11–21
38. Heathcote A, Brown S, Mewhort DJK (2000) The power law repealed: the case for an exponential law of practice. Psychon Bull Rev 7:185–207
39. Pachella RG (1974) The interpretation of reaction time in information–processing research. In: Kantowitz BH (ed) Human information processing: tutorials in performance and cognition. Lawrence Erlbaum Associates, Hillsdale, pp 41–82
40. Audley RJ, Pike AR (1965) Some alternative models of stochastic choice. Br J Math Stat Psychol 207–225
41. Vickers D, Caudrey D, Willson R (1971) Discriminating between the frequency of occurrence of two alternative events. ACTPSY 35:151–172
42. Ratcliff R, Thapar A, McKoon G (2001) The effects of aging on reaction time in a signal detection task. Psychol Aging 16:323
43. Wagenmakers EJ, Ratcliff R, Gómez P, McKoon G (2008) A diffusion model account of criterion shifts in the lexical decision task. J Memory Lang 58:140–159
44. Rae B, Heathcote A, Donkin C, Averell L, Brown SD (2013) The Hare and the tortoise: emphasizing speed can change the evidence used to make decisions. J Exp Psychol Learn Mem Cogn 1–45
45. Wagenmakers EJ, Ratcliff R, Gómez P, Iverson GJ (2004) Assessing model mimicry using the parametric bootstrap. J Math Psychol 48:28–50
46. Ratcliff R, Rouder JN (1998) Modeling response times for two–choice decisions. Psychol Sci 9:347–356
47. Morey RD (2008) Confidence intervals from normalized data: a correction to Cousineau. Tutor Quant Methods Psychol 4:61–64
48. Heathcote A, Love J (2012) Linear deterministic accumulator models of simple choice. Front Psychol 3
49. Starns JJ, Ratcliff R, McKoon G (2012) Evaluating the unequal-variability and dual-process explanations of zroc slopes with response time data and the diffusion model. Cognit Psychol 64:1–34
50. Cleveland WS (1993) Visualizing data. Hobart Press, New Jersey
51. Heathcote A, Brown S, Mewhort DJK (2002) Quantile maximum likelihood estimation of response time distributions. Psychon Bull Rev 9(2):394–401
52. Heathcote A, Brown S (2004) A theoretical basis for QML. Psychon Bull Rev 11:577–578

53. Speckman PL, Rouder JN (2004) A comment on Heathcote, Brown, and Mewhort's QMLE method for response time distributions. Psychon Bull Rev 11:574–576
54. Kass RE, Raftery AE (1995) Bayes factors. J Am Stat Assoc 90:773–795
55. Lodewyckx T, Kim W, Lee MD, Tuerlinckx F, Kuppens P, Wagenmakers EJ (2011) A tutorial on Bayes factor estimation with the product space method. J Math Psychol 55:331–347
56. Shiffrin R, Lee M, Kim W, Wagenmakers EJ (2008) A survey of model evaluation approaches with a tutorial on Hierarchical Bayesian Methods. Cognit Sci A Multidiscip J 32:1248–1284
57. Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A (2002) Bayesian measures of model complexity and fit. J Royal Stat Soc Series B (Stat Methodol) 64:583–639
58. Ando T (2007) Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. Biometrika 94:443–458
59. Myung IJ, Pitt MA (1997) Applying Occam's razor in modeling cognition: a Bayesian approach. Psych Bull Rev 4:79–95
60. Burnham KP, Anderson DR (2004) Multimodel inference: understanding AIC and BIC in model selection. Sociol Methods Res 33:261–304
61. Donkin C, Brown S, Heathcote A (2011) Drawing conclusions from choice response time models: a tutorial using the linear ballistic accumulator. J Math Psychol 55:140–151
62. Turner BM, Sederberg PB, Brown SD, Steyvers M (2013) A method for efficiently sampling from distributions with correlated dimensions. Psychol Methods 18:368–384
63. Farrell S, Wagenmakers EJ, Ratcliff R (2006) 1/f noise in human cognition: is it ubiquitous, and what does it mean? Psych Bull Rev 13:737–741
64. Gilden DL (1997) Fluctuations in the time required for elementary decisions. Psychol Sci 8:296–301
65. Box GEP (1979) Robustness in the strategy of scientific model building. In: Launer RL, Wilkinson GN (ed) Robustness in statistics: proceedings of a workshop. Academic, New York, pp 201–236
66. Lewandowsky S, Farrell S (2010) Computational modeling in cognition: principles and practice. Sage, Thousand Oaks
67. Busemeyer JR, Diederich A (2010) Cognitive modeling. Sage, Thousand Oaks
68. Hunt E (2006) The mathematics of behavior. Cambridge University Press, Cambridge
69. Polk TA, Seifert CM (eds) (2002) Cognitive modeling. MIT Press, Cambridge

Chapter 3
An Introduction to the Diffusion Model of Decision Making

Philip L. Smith and Roger Ratcliff

Abstract The diffusion model assumes that two-choice decisions are made by accumulating successive samples of noisy evidence to a response criterion. The model has a pair of criteria that represent the amounts of evidence needed to make each response. The time taken to reach criterion determines the decision time and the criterion that is reached first determines the response. The model predicts choice probabilities and the distributions of response times for correct responses and errors as a function of experimental conditions such as stimulus discriminability, speed-accuracy instructions, and manipulations of relative stimulus frequency, which affect response bias. This chapter describes the main features of the model, including mathematical methods for obtaining response time predictions, methods for fitting it to experimental data, including alternative fitting criteria, and ways to represent the fit to multiple experimental conditions graphically in a compact way. The chapter concludes with a discussion of recent work in psychology that links evidence accumulation to processes of perception, attention, and memory, and in neuroscience, to neural firing rates in the oculomotor control system in monkeys performing saccade-to-target decision tasks.

3.1 Historical Origins

The human ability to translate perception into action, which we share with nonhuman animals, relies on our ability to make rapid decisions about the contents of our environment. Any form of coordinated, goal-directed action requires that we be able to recognize things in the environment as belonging to particular cognitive categories or classes and to select the appropriate actions to perform in response. To a very significant extent, coordinated action depends on our ability to provide rapid answers to questions of the form: "What is it?" and "What should I do about it?" When viewed in this way, the ability to make rapid decisions—to distinguish

P. L. Smith (✉)
Melbourne School of Psychological Sciences, The University of Melbourne, Melbourne, VIC, Australia

R. Ratcliff
Department of Psychology, The Ohio State University, Columbus, OH, USA

B. U. Forstmann, E.-J. Wagenmakers (eds.), *An Introduction to Model-Based Cognitive Neuroscience,* DOI 10.1007/978-1-4939-2236-9_3

predator from prey, or friend from foe—appears as one of the basic functions of the brain and central nervous system. The purpose of this chapter is to provide an introduction to the mathematical modeling of decisions of this kind.

Historically, the study of decision-making in psychology has been closely connected to the study of sensation and perception—an intellectual tradition with its origins in philosophy and extending back to the nineteenth century. Two strands of this tradition are relevant: psychophysics, defined as the study of the relationship between the physical magnitudes of stimuli and the sensations they produce, and the study of reaction time or response time (RT). Psychophysics, which had its origins in the work of Gustav Fechner in the Netherlands in 1860 on "just noticeable differences," led to the systematic study of decisions about stimuli that are difficult to detect or to discriminate. The study of RT was initiated by Franciscus Donders, also in the Netherlands, in 1868. Donders, inspired by the pioneering work of Hermann von Helmholtz on the speed of nerve conduction, sought to develop methods to measure the speed of mental processes. These two strands of inquiry were motivated by different theoretical concerns, but led to a common realization, namely, that decision-making is inherently variable. People do not always make the same response to repeated presentation of the same stimulus and the time they take to respond to it varies from one presentation to the next.

Trial-to-trial variation in performance is a feature of an important class of models for speeded, two-choice decision-making developed in psychology, known as *sequential-sampling* models. These models regard variation in decision outcomes and decision times as the empirical signature of a noisy evidence accumulation process. They assume that, to make a decision, the decision maker accumulates successive samples of noisy evidence over time, until sufficient evidence for a response is obtained. The samples represent the momentary evidence favoring particular decision alternatives at consecutive time points. The decision time is the time taken to accumulate a sufficient, or criterion, amount of evidence and the decision outcome depends on the alternative for which a criterion amount of evidence is first obtained. The idea that decision processes are noisy was first proposed on theoretical grounds, to explain the trial-to-trial variability in behavioral data, many decades before it was possible to use microelectrodes in awake, behaving animals to record this variability directly. The noise was assumed to reflect the moment-to-moment variability in the cognitive or neural processes that represent the stimulus [1–4].

In this chapter, we describe one such sequential-sampling model, the diffusion model of Ratcliff [5]. Diffusion models, along with random walk models, comprise one of the two main subclasses of sequential-sampling models in psychology; the other subclass comprises accumulator and counter models. For space reasons, we do not consider models of this latter class in this chapter. The interested reader is referred to references [2–4] and [6] for discussions. To distinguish Ratcliff's model from other models that also represent evidence accumulation as a diffusion process, we refer to it as the *standard diffusion model.* Historically, this model was the first model to represent evidence accumulation in two-choice decision making as a diffusion process and it remains, conceptually and mathematically, the benchmark against

which other models can be compared. It is also the model that has been most extensively and successfully applied to empirical data. We restrict our consideration here to two-alternative decision tasks, which historically and theoretically have been the most important class of tasks in psychology.

3.2 Diffusion Processes and Random Walks

Mathematically, diffusion processes are the continuous-time counterparts of random walks, which historically preceded them as models for decision-making. A random walk is defined as the running cumulative sum of a sequence of independent random variables, Z_j, $j = 1, 2, \ldots$. In models of decision-making, the values of these variables are interpreted as the evidence in a sequence of discrete observations of the stimulus. Typically, evidence is assumed to be sampled at a constant rate, which is determined by the minimum time needed to acquire a single sample of perceptual information, denoted Δ. The random variables are assumed to take on positive and negative values, with positive values being evidence for one response, say R_a, and negative values evidence for the other response, R_b. For example, in a brightness discrimination task, R_a might correspond to the response "bright" and R_b correspond to the response "dim." The mean of the random variables is assumed to be positive or negative, depending on the stimulus presented. The cumulative sum of the random variables,

$$X_i = \sum_{j=1}^{i} Z_j,$$

is a random walk. If the Z_j are real-valued, the domain of the walk is the positive integers and the range is the real numbers. To make a decision, the decision-maker sets a pair of evidence criteria, a and b, with $b < 0 < a$ and accumulates evidence until the cumulative evidence total reaches or exceeds one of the criteria, that is, until $X_i \geq a$ or $X_i \leq b$. The time taken for this to occur is the *first passage time* through one of the criteria, defined formally as

$$T_a = \min\{i\Delta : \ X_i \geq a | X_j > b; \ j < i\}$$
$$T_b = \min\{i\Delta : \ X_i \leq b | X_j < a; \ j < i\}.$$

If the first criterion reached is a, the decision maker makes response R_a; if it is b, the decision maker makes response R_b. The decision time, T_D, is the time for this to occur

$$T_D = \min\{T_a, T_b\}.$$

If response R_a is identified as the correct response for the stimulus presented, then the mean, or expected value, of T_a, denoted $E[T_a]$, is the mean decision time for

correct responses; $E[T_b]$ is the mean decision time for errors, and the probability of a correct response, $P(C)$, is the *first passage probability* of the random walk through the criterion a,

$$P(C) = \text{Prob}\{T_a < T_b\}.$$

Although either T_a or T_b may be infinite on a given realization of the process the other will be finite, so T_D will be finite with probability one; that is, the process will terminate with one or other response in finite time [7]. This means that the probability of an error response, $P(E)$, will equal $1 - P(C)$.

Random walk models of decision-making have been proposed by a variety of authors. The earliest of them were influenced by Wald's sequential probability ratio test (SPRT) in statistics [8] and assumed that the random variables Z_j were the log-likelihood ratios that the evidence at each step came from one as opposed to the other stimulus. The most highly-developed of the SPRT models was proposed by Laming [9]. The later *relative judgment theory* of Link and Heath [10] assumed that the decision process accumulates the values of the noisy evidence samples directly rather than their log-likelihood ratios. Evaluation of these models focused primarily on the relationship between mean RT and accuracy and the ordering of mean RTs for correct responses and errors as a function of experimental manipulations [2–4, 9, 10].

3.3 The Standard Diffusion Model

A diffusion process may be thought of as random walk in continuous time. Instead of accumulating evidence at discrete time points, evidence is accumulated continuously. Such a process can be obtained mathematically via a limiting process, in which the sampling interval is allowed to go to zero while constraining the average size of the evidence at each step to ensure the variability of the process in a given, fixed time interval remains constant [7, 11]. The study of diffusion processes was initiated by Albert Einstein, who proposed a diffusion model for the movement of a pollen particle undergoing random Brownian motion [11]. The rigorous study of such processes was initiated by Norbert Wiener [12]. For this reason, the simplest diffusion process is known variously as the Wiener process or the Brownian motion process.

In psychology, Ratcliff [5] proposed a diffusion model of evidence accumulation in two-choice decision-making—in part because it seemed more natural to assume that the brain accumulates information continuously rather than at discrete time points. Ratcliff also emphasized the importance of studying RT distributions as a way to evaluate models. Sequential-sampling models not only predict choice probabilities and mean RTs, they predict entire distributions of RTs for correct responses and errors. This provides for very rich contact between theory and experimental data, allowing for strong empirical tests.

The main elements of the standard diffusion model are shown in Fig. 3.1. We shall denote the accumulating evidence state in the model as X_t, where t denotes time.

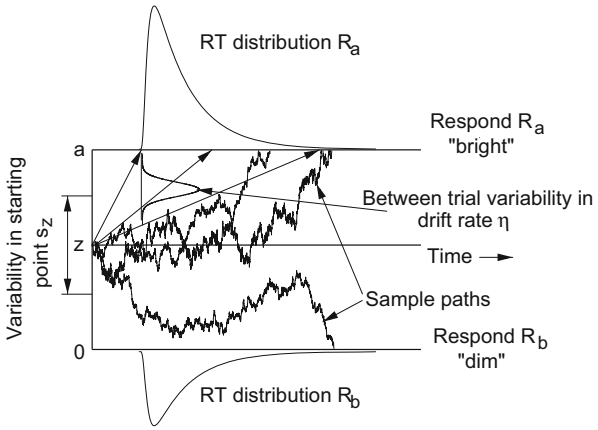


Fig. 3.1 Diffusion model. The process starting at z accumulates evidence between decision criteria at 0 and a. Moment-to-moment variability in the accumulation process means the process can terminate rapidly at the correct response criterion, slowly at the correct response criterion, or at the incorrect response criterion. There is between-trial variability in the drift rate, ξ, with standard deviation η, and between-trial variability in the starting point, z, with range s_z

Before describing the model, we should mention that there are two conventions used in psychology to characterize diffusion models. The convention used in the preceding section assumes the process starts at zero and that the criteria are located at a and b, with $b < 0 < a$. The other is based on Feller's [13] analysis of the so-called gambler's ruin problem and assumes that the process starts at z and that the criteria are located at 0 and a, with $0 < z < a$. As the latter convention was used by Ratcliff in his original presentation of the model [5] and in later work, this is the convention we shall adopt for the remainder of this chapter. The properties of the process are unaltered by translations of the starting point; such processes are called *spatially homogeneous*. For processes of this kind, a change in convention simply represents a relabeling of the y-axis that represents the accumulating evidence state. Other, more complex, diffusion processes, like the Ornstein-Uhlenbeck process [14–16], are not spatially homogeneous and their properties are altered by changes in the assumed placement of the starting point.

As shown in the figure, the process, starting at z, begins accumulating evidence at time $t = 0$. The rate at which evidence accumulates, termed the *drift* of the process and denoted ξ, depends on the stimulus that is presented and its discriminability. The identity of the stimulus determines the direction of drift and the discriminability of the stimulus determines the magnitude. Our convention is that when stimulus s_a is presented the drift is positive and the value of X_t tends to increase with time, making it is more likely to terminate at the upper criterion and result in response R_a. When stimulus s_b is presented the drift is negative and the value of X_t tends to decrease with time, making it is more likely to terminate at the lower boundary with response R_b. In our example brightness discrimination task, bright stimuli lead to positive values of drift and dim stimuli lead to negative values of drift. Highly

discriminable stimuli are associated with larger values of drift, which lead to more rapid information accumulation and faster responding. Because of noise in the process, the accumulating evidence is subject to moment-to-moment perturbations. The time course of evidence accumulation on three different experimental trials, all with the same drift rate, is shown in the figure. These noisy trajectories are termed the *sample paths* of the process. A unique sample path describes the time course of evidence accumulation on a given experimental trial. The sample paths in the figure show some of the different outcomes that are possible for stimuli with the same drift rate. The sample paths in the figure show: (a) a process terminating with a correct response made rapidly; (b) a process terminating with a correct response made slowly, and (c) a process terminating with an error response. In behavioral experiments, only the response and the RT are observables; the paths themselves are not. They are theoretical constructs used to explain the observed behavior.

The noisiness, or variability, in the accumulating evidence is controlled by a second parameter, the *infinitesimal standard deviation*, denoted s. Its square, s^2, is termed the *diffusion coefficient*. The diffusion coefficient determines the variability in the sample paths of the process. Because the parameters of a diffusion model are only identified to the level of a ratio, all the parameters of the model can be multiplied by a constant without affecting any of the predictions. To make the parameters estimable, it is common practice to fix s arbitrarily. The other parameters of the model are then expressed in units of infinitesimal standard deviation, or infinitesimal standard deviation per unit time.

3.4 Components of Processing

As shown in Fig. 3.1, the diffusion model predicts RT distributions for correct responses and errors. Moment-to-moment variability in the sample paths of the process, controlled by the diffusion coefficient, means that on some trials the process will finish rapidly and on others it will finish slowly. The predicted RT distributions have a characteristic unimodal, positively-skewed shape: More of the probability mass in the distribution is located below the mean than above it. As the drift of the process changes with changes in stimulus discriminability, the relative proportions of correct responses and errors change, and the means and standard deviations of the RT distributions also change. However, the shapes of the RT distributions change very little; to a good approximation, RT distributions for low discriminability stimuli are scaled copies of those for high discriminability stimuli [17].

One of the main strengths of the diffusion model is that the shapes of the RT distributions it predicts are precisely those found in empirical data. Many experimental tasks, including low-level perceptual tasks like signal detection and higher-level cognitive tasks like lexical decision and recognition memory, yield families of RT distributions like those predicted by the model [6]. In contrast, other models, particularly those of the accumulator/counter model class, predict distribution shapes that become more symmetrical with reductions in discriminability [6]. Such distributions

tend not to be found empirically, except in situations in which people are forced to respond to an external deadline.

One of the problems with early random walk models of decision-making—which they shared with the simplest form of the diffusion model—is they predicted that mean RTs for correct responses and errors would be equal [2]. Specifically, if $E[R_j|s_i]$, denotes the mean RT for response R_j to stimulus s_i, with $i, j \in \{a, b\}$, then, if the drifts for the two stimuli are equal in magnitude and opposite in sign, as is natural to assume for many perceptual tasks, the models predicted that $E[R_a|s_a] = E[R_a|s_b]$ and $E[R_b|s_a] = E[R_b|s_b]$; that is, the mean time for a given response made correctly is the same as the mean time for that response made incorrectly. They also predicted, when the starting point is located equidistantly between the criteria, $z = a/2$, that $E[R_a|s_a] = E[R_b|s_a]$ and $E[R_a|s_b] = E[R_b|s_b]$; that is, the mean RT for correct responses to a given stimuli is the same as the mean error RT to that same stimulus. This prediction holds regardless of the relative magnitudes of the drifts. Indeed, a stronger prediction holds; the models predicted equality not only of mean RTs, but of the entire distributions of correct responses and errors. These predictions almost never hold empirically. Rather, the typical finding is that when discriminability is high and speed is stressed, error mean times are shorter than correct mean times. When discriminability is low and accuracy is stressed, error mean times are longer than correct mean times [2]. Some studies show a crossover pattern, in which errors are faster than correct responses in some conditions and slower in others [6].

A number of modifications to random walk models were proposed to deal with the problem of the ordering of mean RTs for correct responses and errors, including asymmetry (non-normality) of the distributions of evidence that drive the walk [1, 10], and biasing of an assumed log-likelihood computation on the stimulus information at each step [18], but none of them provided a completely satisfactory account of the full range of experimental findings. The diffusion model attributes inequality of the RTs for correct responses and errors to between-trial variability in the operating characteristics, or "components of processing," of the model. The diffusion model predicts equality of correct and error times only when the sole source of variability in the model is the moment-to-moment variation in the accumulation process. Given the complex interaction of perceptual and cognitive processes involved in decision-making, such an assumption is probably an oversimplification. A more realistic assumption is that there is trial-to-trial variability, both in the quality of information entering the decision process and in the decision-maker's setting of decision criteria or starting points. Trial-to-trial variability in the information entering the decision process would arise either from variability in the efficiency of the perceptual encoding of stimuli or from variation in the quality of the information provided by nominally equivalent stimuli. Trial-to-trial variability in decision criteria or starting points would arise as the result of the decision-maker attempting to optimize the speed and accuracy of responding [4]. Most RT tasks show sequential effects, in which the speed and accuracy of responding depends on the stimuli and/or the responses made on preceding trials, consistent with the idea that there is some kind of adaptive regulation of the settings of the decision process occurring across trials [2, 4].

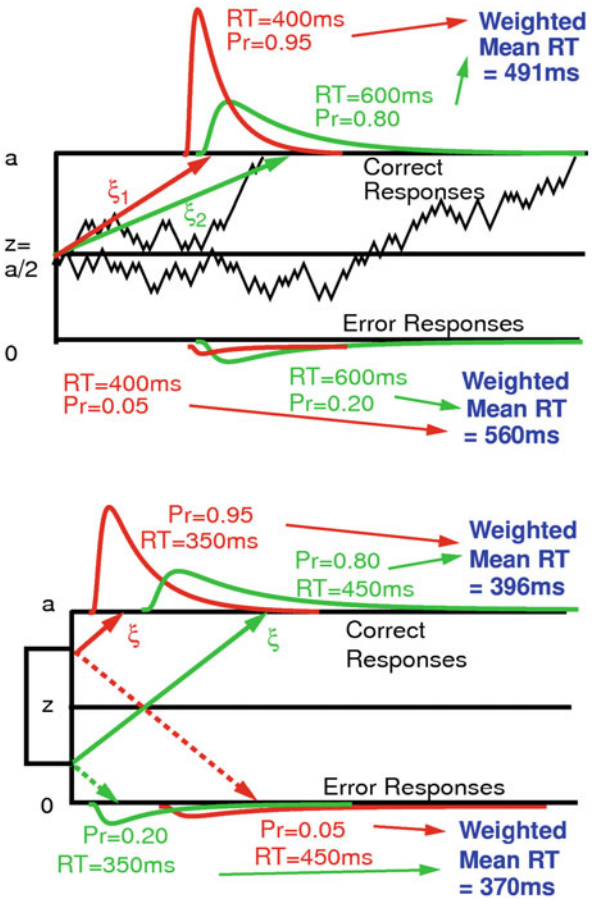


Fig. 3.2 Effects of trial-to-trial variability in drift rates and starting points. The predicted RT distributions are probability mixtures across processes with different drift rates (*top*) or different starting points (*bottom*). Variability in drift rates leads to slow errors; variability in starting points leads to fast errors

The diffusion model assumes that there is trial-to-trial variation in both drift rates and starting points. Ratcliff [5] assumed that the drift rate on any trial, ξ, is drawn from a normal distribution with mean ν and standard deviation η. Subsequently Ratcliff, Van Zandt, and McKoon [19] assumed that there is also trial-to-trial variability in the starting point, z, which they modeled as a rectangular distribution with range s_z. They chose a rectangular distribution mainly on the grounds of convenience, because the predictions of the model are relatively insensitive to the distribution's form. The main requirement is that all of the probability mass of the distribution must lie between the decision criteria, which is satisfied by a rectangular distribution with s_z suitably constrained. The distributions of drift and starting point are shown in Fig. 3.1.

Trial-to-trial variation in drift rates allows the model to predict slow errors; trial-to-trial variation in starting point allows it to predict fast errors. The combination of the two allows it to predict crossover interactions, in which there are fast errors for high discriminability stimuli and slow errors for low discriminability stimuli. Figure 3.2a shows how trial-to-trial variability in drift results in slow errors. The assumption that

drift rates vary across trials means that the predicted RT distributions are probability mixtures, made up of trials with different values of drift. When the drift is small (i.e., near zero), error rates will be high and RTs will be long. When the drift is large, error rates will be low and RTs will be short. Because errors are more likely on trials on which the drift is small, a disproportionate number of the trials in the error distribution will be trials with small drifts and long RTs. Conversely, because errors are less likely on trials on which drift is large, a disproportionate number of the trials in the correct response distribution will be trials with large drifts and short RTs. In either instance, the predicted mean RT will be the weighted mean of the RTs on trials with small drift and large drifts.

Figure 3.2a illustrates how slow errors arise in a simplified case in which there are just two drifts, ξ_1 and ξ_2, with $\xi_1 > \xi_2$. When the drift is ξ_1, the mean RT is 400 ms and the probability of a correct response, $P(C)$, is 0.95. When the drift is ξ_2, the mean RT is 600 and $P(C) = 0.80$. The predicted mean RTs are the weighted means of large drift and small drift trials. The predicted mean RT for correct responses is $(0.95 \times 400 + 0.80 \times 600)/1.75 = 491$ ms. The predicted mean for error responses $(0.05 \times 400 + 0.20 \times 600)/0.25 = 560$ ms. Rather than just two drifts, the diffusion model assumes that the predicted means for correct responses and errors are weighted means across an entire normal distribution of drift. However, the effect is the same: predicted mean RTs errors are longer than those for correct responses.

Figure 3.2b illustrates how fast errors arise as the result of variation in starting point. Again, we have shown a simplified case, in which there are just two starting points, one of which is closer to the lower, error, response criterion and the other of which is closer to the upper, correct, response criterion. In this example, a single value, of drift, ξ, has been assumed for all trials. The model predicts fast errors because the mean time for the process to reach criterion depends on the distance it has to travel and because it is more likely to terminate at a particular criterion if the criterion is near the starting point rather than far from it. When the starting point is close to the lower criterion, errors are faster and also more probable. When the starting point is close to the upper criterion, errors are slower, because the process has to travel further to reach the error criterion, and are less probable. Once again, the predicted distributions of correct responses and errors are probability mixtures across trials with different values of starting point.

In the example shown in Fig. 3.2b, when the process starts near the upper criterion, the mean RT for correct responses is 350 ms and $P(C) = 0.95$. When it starts near the lower criterion, the mean RT for correct responses is 450 ms and $P(C) = 0.80$. The predicted mean RTs for correct responses and errors are again the weighted means across starting points. In this example, the mean RT for correct responses is $(0.95 \times 350 + 0.80 \times 450)/1.75 = 396$ ms; the mean RT for errors is $(0.20 \times 350 + 0.05 \times 450)/0.25 = 370$ ms. Again, the model assumes that the predicted mean times are weighted means across the entire distribution of starting points, but the effect is the same: predicted mean times for errors are faster than those correct responses. When equipped with both variability in drift and starting point, the model can predict both the fast errors and the slow errors that are found experimentally [6].

The final component of processing in the model is the non-decision time, denoted T_{er}. Like many other models in psychology, diffusion model assumes that RT can be additively decomposed into the decision time, T_D, and the time for other processes, T_{er}:

$$RT = T_D + T_{er}.$$

The subscript in the notation means "encoding and responding." In many applications of the model, it suffices to treat T_{er} as a constant. In practice, this is equivalent to assuming that it is an independent random variable whose variance is negligible compared to that of T_D. In other applications, particularly those in which discriminability is high and speed is emphasized and RT distributions have small variances, the data are better described by assuming that T_{er} is rectangularly distributed with range s_t. As with the distribution of starting point, the rectangular distribution is used mainly as a convenience, because when the variance of T_{er} is small compared to that of T_D, the shape of the distribution will be determined almost completely by the shape of the distribution of decision times. The advantage of assuming some variability in T_{er} in these settings is that it allows the model to better capture the leading edge of the empirical RT distributions, which characterizes the fastest 5–10 % of responses, and which tends to be slightly more variable than the model predicts.

3.5 Bias and Speed-Accuracy Tradeoff Effects

Bias effects and speed-accuracy tradeoff effects are ubiquitous in experimental psychology. Bias effects typically arise when the two stimulus alternatives occur with unequal frequency or have unequal rewards attached to them. Speed-accuracy tradeoff effects arise as the result of explicit instructions emphasizing speed or accuracy or as the result of an implicit set on the part of the decision-maker. Such effects can be troublesome in studies that measure only accuracy or only RT, because of the asymmetrical way in which these variables can be traded off. Small changes in accuracy can be traded off against large changes in RT, which can sometimes make it difficult to interpret a single variable in isolation [2].

One of the attractive features of sequential-sampling models like the diffusion model is that they provide a natural account of how speed-accuracy tradeoffs arise. As shown in Fig. 3.3, the models assume that criteria are under the decision-maker's control. Moving the criteria further from the starting point (i.e., increasing a while keeping $z = a/2$) increases the distance the process must travel to reach a criterion and also reduces the probability that it will terminate at the wrong criterion because of the cumulative effects of noise. The effect of increasing criteria will thus be slower and more accurate responding. This is the speed-accuracy tradeoff.

The diffusion model with variation in drift and starting point can account for the interactions with experimental instructions emphasizing speed or accuracy that are found experimentally. When accuracy is emphasized and criteria are set far from the starting point, variations in drift have a greater effect on performance than do

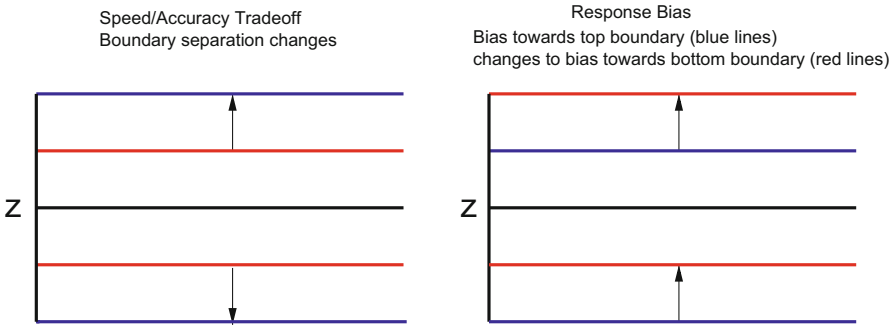


Fig. 3.3 Speed-accuracy tradeoff and response bias. Reducing decision criteria leads to faster and less accurate responding. Shifting the starting point biases the process towards the response associated with the nearer criterion

variations in starting point, and so slow errors are found. When speed is emphasized and criteria are near the starting point, variations in starting point have a greater effect on performance than do variations in drift and fast errors are found.

Like other sequential-sampling models, the diffusion model accounts for bias effects by assuming unequal criteria, represented by a shift in the starting point towards the upper or lower criterion, as shown in Fig. 3.3. Shifting the starting point towards a particular response criterion increases the probability of that response and reduces the average time taken to make it. The probability of making the other response is reduced and the average time to make it is correspondingly increased. The effect of changing the prior probabilities of the two responses, by manipulating the relative stimulus frequencies, is well described by a change in the starting point (unequal decision criteria). In contrast, unequal reward rates not only lead to a bias in decision criteria, they also lead to a bias in the way stimulus information is classified [20]. This can be captured in the idea of a *drift criterion*, which is a criterion on the stimulus information, like the criterion in signal detection theory. The effect of changing the drift criterion is to make the drift rates for the two stimuli unequal. Both kinds of bias effects appear to operate in tasks with unequal reward rates.

3.6 Mathematical Methods For Diffusion Models

Diffusion processes can be defined mathematically either via partial differential equations or by stochastic differential equations. If $f(\tau, y; t, x)$ is the transition density of the process X_t, that is, $f(\tau, y; t, x)\,dx$ is the probability that a process starting at time τ in state y will be found at time t in a small interval $(x, x + dx)$, then the accumulation process X_t, with drift ξ and diffusion coefficient s^2, satisfies the partial differential equation

$$\frac{\partial f}{\partial \tau} = \frac{1}{2} s^2 \frac{\partial^2 f}{\partial y^2} + \xi \frac{\partial f}{\partial y}.$$

This equation is known in the probability literature as Kolmogorov's backward equation, so called because its variables are the starting time τ and the initial state y. The process also satisfies a related equation known as Kolmogorov's forward equation, which is an equation in t and x [7, 11]. The backward equation is used to derive RT distributions; the forward equation is useful for characterizing the accumulated evidence at time t for processes that have not yet terminated at one of the criteria [5].

Alternatively, the process can be defined as satisfying the stochastic differential equation [11]:

$$dX_t = \xi dt + s\, dW_t.$$

The latter equation is useful because it provides a more direct physical intuition about the properties of the accumulation process. Here dX_t is interpreted as the small, random change in the accumulated evidence occurring in a small time interval of duration dt. The equation says that the change in evidence is the sum of a deterministic and a random part. The deterministic part is proportional to the drift rate, ξ; the random part is proportional to the infinitesimal standard deviation, s. The term on the right, dW_t, is the differential of a Brownian motion or Wiener process, W_t. It can be thought of as the random change in the accumulation process during the interval dt when it is subject to the effects of many small, independent random perturbations, described mathematically as a *white noise* process. White noise is a mathematical abstraction, which cannot be realized physically, but it provides a useful approximation to characterize the properties of physical systems that are perturbed by broad-spectrum, Gaussian noise. Stochastic differential equations are usually written in the differential form given here, rather than in the more familiar form involving derivatives, because of the extreme irregularity of the sample paths of diffusion processes, which means that quantities of the form dX_t/dt are not well defined mathematically.

Solution of the backward equation leads to an infinite series expression for the predicted RT distributions and an associated expression for accuracy [5, 7, 11]. The stochastic differential equation approach leads to a class of integral equation methods that were developed in mathematical biology to study the properties of integrate-and-fire neurons. The interested reader is referred to references [6, 16, 21] for details. For a two-boundary process with drift ξ, boundary separation a, starting point z, and infinitesimal standard deviation s, with no variability in any of its parameters, the probability of responding at the lower barrier, $P(\xi, a, z)$, is

$$P(\xi, a, z) = \frac{\exp(-2\xi a/s^2) - \exp(-2\xi z/s^2)}{\exp(-2\xi a/s^2) - 1}.$$

The cumulative distribution of first passage times at the lower boundary is

$$G(t, \xi, a, z) =$$

$$P(\xi, a, z) - \frac{\pi s^2}{a^2} e^{-\xi z/s^2} \sum_{k=1}^{\infty} \frac{2k \sin\left(\frac{k\pi z}{a}\right) \exp\left\{-\frac{1}{2}\left(\frac{\xi^2}{s^2} + \frac{k^2\pi^2 s^2}{a^2}\right) t\right\}}{\left(\frac{\xi^2}{s^2} + \frac{k^2\pi^2 s^2}{a^2}\right)}.$$

The probability of a response and the cumulative distribution of first passage times at the upper boundary are obtained by replacing ξ with $-\xi$ and z with $a - z$ in the preceding expressions. More details can be found in reference [5].

In addition to the partial differential equation and integral equation methods, predictions for diffusion models can also be obtained using finite-state Markov chain methods or by Monte Carlo simulation [22]. The Markov chain approach, developed by Diederich and Busemeyer [23], approximates a continuous-time, continuous-state, diffusion process by a discrete-time, discrete-state, birth-death process [5]. A transition matrix is defined that specifies the probability of an increment or a decrement to the process, conditional on its current state. The entries in the transition matrix express the relationship between the drift and diffusion coefficients of the diffusion process and the transition probabilities of the approximating Markov chain [24]. The transition matrix includes two special entries that represent criterion states, which are set equal to 1.0, expressing the fact that once the process has transitioned into a criterion state, it does not leave it. An initial state vector is defined, which represents the distribution of probability mass at the beginning of the trial, including the effects of any starting point variation. First passage times and probabilities can then be obtained by repeatedly multiplying the state vector by the transition matrix. These alternative methods are useful for more complex models for which an infinite-series solution may not be available. There are now software packages available for fitting the standard diffusion model that avoid the need to implement the model from first principles [25–27].

3.7 The Representation of Empirical Data

The diffusion model predicts accuracy and distributions of RT for correct responses and errors as a function of the experimental variables. In many experimental settings, the discriminability of the stimuli is manipulated as a within-block variable, while instructions, payoffs, or prior probabilities are manipulated as between-block variables. The model assumes that manipulations of discriminability affect drift rates, while manipulations of other variables affect criteria or starting points. Although criteria and starting points can vary from trial to trial, they are assumed to be independent of drift rates, and to have the same average value for all stimuli in a block. This assumption provides an important constraint in model testing.

To show the effects of discriminability variations on accuracy and RT distributions, the data and the predictions of the model are represented in the form of a *quantile-probability plot*, as shown in Fig. 3.4. To construct such a plot, each of the RT distributions is summarized by an equal-area histogram. Each RT distribution is represented by a set of rectangles, each representing 20 % of the probability mass in the distribution, except for the two rectangles at the extremes of the distribution, which together represent the 20 % of mass in the upper and lower tails. The time-axis bounds of the rectangles are distribution quantiles, that is, those values of time that cut off specified proportions of the mass in the distribution. Formally, the pth

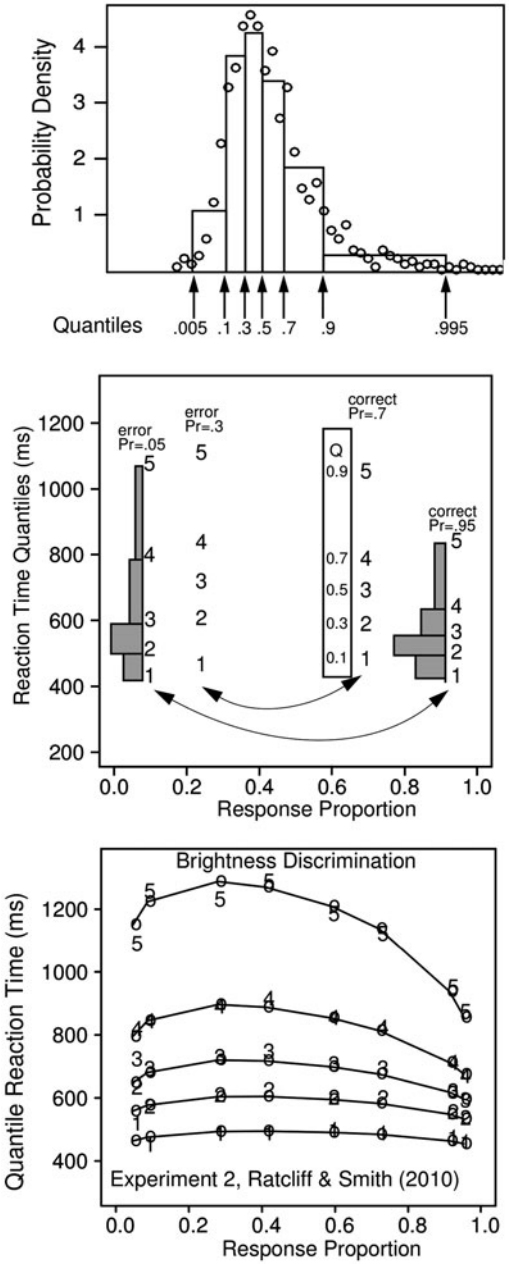


Fig. 3.4 Representing data in a quantile probability plot. *Top panel*: An empirical RT distribution is summarized using an equal-area histogram with bins bounded by the distribution quantiles. *Middle panel*: The quantiles of the RT distributions for correct responses and errors are plotted vertically against the probability of a correct response on the *right* and the probability of an error response on the *left*. *Bottom panel*: Example of an empirical quantile probability plot from a brightness discrimination experiment

quantile, Q_p, is defined to be the value of time such that the proportion of RTs in the distribution that are less than or equal to Q_p is equal to p. The distribution in the figure has been summarized using five quantiles: the 0.1, 0.3, 0.5, 0.7, and 0.9 quantiles. The 0.1 and 0.9 quantiles represent the upper and lower tails of the distribution, that is, the fastest and slowest responses, respectively. The 0.5 quantile is the median and represents the distribution's central tendency. As shown in the figure, the set of five quantiles provides a good summary of the location, variability, and shape of the distribution.

To construct a quantile probability plot, the quantile RTs for correct responses and errors are plotted on the y-axis against the choice probabilities (i.e., accuracy) on the x-axis for each stimulus condition, as shown in the middle panel of the figure. Specifically, if, $Q_{i,p}(C)$ and $Q_{i,p}(E)$ are, respectively, the quantiles of the RT distributions for correct responses and errors in condition i of the experiment, and $P_i(C)$ and $P_i(E)$ are the probabilities of a correct response and an error in that condition, then the values of $Q_{i,p}(C)$ are plotted vertically against $P_i(C)$ for $p = 0.1, 0.3, 0.5, 0.7, 0.9$, and the values of $Q_{i,p}(E)$ are similarly plotted against $P_i(E)$. All of the distribution pairs and choice probabilities from each condition are plotted in a similar way.

The bottom panel of the figure shows data from a brightness discrimination experiment from Ratcliff and Smith [28] in which four different levels of stimulus discriminability were used. Because of the way the plot is constructed, the two outermost distributions in the plot represent performance for the most discriminable stimuli and the two innermost distributions represent performance for the least discriminable stimuli. The value of the quantile-probability plot is that it shows how performance varies parametrically as stimulus discriminability is altered, and how different parts of the RT distributions for correct responses and errors are affected differently. As shown in the figure, most of the change in the RT distribution with changing discriminability occurs in the upper tail of the distribution (e.g., the 0.7 and 0.9 quantiles); there is very little change in the leading edge (the 0.1 quantile). This pattern is found in many perceptual tasks and also in more cognitive tasks like recognition memory. The quantile-probability plot also shows that errors were slower than correct responses in all conditions. This appears as a left-right asymmetry in the plot; if the distributions for correct responses and errors were the same, the plot would be mirror-image symmetrical around its vertical midline. The predicted degree of asymmetry is a function of the standard deviation of the distribution of drift rates, η and, when there are fast errors, of the range of starting points, s_z. The slow-error pattern of data in Fig. 3.4 is typical of difficult discrimination tasks in which accuracy is emphasized.

The pattern of data is Fig. 3.4 is rich and highly-constrained and represents a challenge for any model. The success of the diffusion model is that it has shown repeatedly that it can account for data of this kind. Its ability to do so is not a just a matter of model flexibility. It is not the case that the model is able to account for any pattern of data whatsoever [29]. Rather, as noted previously, the model predicts families of RT distributions that have a specific and quite restricted form. Distributions of this particular form are the ones most often found in experimental data.

3.8 Fitting the Model to Experimental Data

Fitting the model to experimental data requires estimation of its parameters by iterative, nonlinear minimization. A variety of minimization algorithms have been used in the literature, but the Nelder-Mead SIMPLEX algorithm has been popular because of its robustness [30]. Parameters are estimated to minimize a fit statistic, or loss function, that characterizes the discrepancy between the model and the data. A variety of fit statistics have been used in applications, but chi-square-type statistics, either the Pearson chi-square (χ^2) or the likelihood-ratio chi-square (G^2), are common. For an experiment with m stimulus conditions, these are defined as

$$\chi^2 = \sum_{i=1}^{m} n_i \sum_{j=1}^{12} \frac{(p_{ij} - \pi_{ij})^2}{\pi_{ij}}$$

and

$$G^2 = 2 \sum_{i=1}^{m} n_i \sum_{j=1}^{12} p_{ij} \ln\left(\frac{p_{ij}}{\pi_{ij}}\right),$$

respectively These statistics are asymptotically equivalent and yield similar results in most applications. In these equations, the outer summation over i indexes the m conditions in the experiment and the inner summation over j indexes the 12 bins defined by the quantiles of the RT distributions for correct responses and errors. (The use of five quantiles per distribution gives six bins per distribution, or 12 bins per correct and error distribution pair.) The quantities p_{ij} and π_{ij} are the observed and predicted proportions of probability mass in each bin, respectively, and n_i is the number of stimuli in the ith experimental condition. For bins defined by the quantile bounds, the values of p_{ij} will equal 0.2 or 0.1, depending on whether or not the bin is associated with a tail quantile, and the values of π_{ij} are the differences in the probability mass in the cumulative finishing time distributions, evaluated at adjacent quantiles, $G(Q_{i,p}, \nu, a, z) - G(Q_{i,p-1}, \nu, a, z)$. Here we have written the cumulative distribution as a function of the mean drift, ν, rather than the trial-dependent drift, ξ, to emphasize that the cumulative distributions are probability mixtures across a normal distribution of drift values. Because the fit statistics keep track of the distribution of probability mass across the distributions of correct responses and errors, minimizing them fits both RT and accuracy simultaneously.

Fitting the model typically requires estimation of around 8–10 parameters. For an experiment with a single experimental condition and four different stimulus discriminabilities like the one shown in Fig. 3.4, a total of 10 parameters must be estimated to fit the full model. There are four values of the mean drift, ν_i, $i = 1, \ldots, 4$, a boundary separation parameter, a, a starting point, z, a non-decision time, T_{er}, and variability parameters for the drift, starting point, and non-decision time, η, s_z, and s_t, respectively. As noted previously, to make the model estimable, the infinitesimal standard deviation is typically fixed to an arbitrary value (Ratcliff uses $s = 0.1$ in his

work, but $s = 1.0$ has also been used). In experiments in which there is no evidence of response bias, the data can be pooled across the two responses to create one distribution of correct responses and one distribution of errors per stimulus condition. Under these conditions, a symmetrical decision process can be assumed ($z = a/2$) and the number of free parameters reduced by one. Also, as discussed previously, in many applications the non-decision time variability parameter can be set to zero without worsening the fit.

Although the model has a reasonably large number of free parameters, it affords a high degree of data reduction, defined as the number of degrees of freedom in the data divided by the number of free parameters in the model. There are $11m$ degrees of freedom in a data set with m conditions and six bins per distribution (one degree of freedom is lost for each correct-error distribution pair, because the expected and observed masses are constrained to be equal in each pair, giving $12 - 1 = 11$ degrees of freedom per pair). For the experiment in Fig. 3.4, there are 44 degrees of freedom in the data and the model had nine free parameters, which represents a data reduction ratio of almost 5:1. For larger data sets, data reduction ratios of better than 10:1 are common. This represents a high degree of parsimony and explanatory power.

It is possible to fit the diffusion model by maximum likelihood instead of by minimum chi-square. Maximum likelihood defines a fit statistic (a likelihood function) on the set of raw RTs rather than on the probability mass in the set of bins, and maximizes this (i.e., minimizes its negative). Despite the theoretical appeal of maximum likelihood, its disadvantage is that it is vulnerable to the effects of contaminants or outliers in a distribution. Almost all data sets have a small proportion of contaminant responses in them, whether from finger errors or from lapses in vigilance or attention, or other causes. RTs from such trials are not representative of the process of theoretical interest. Because maximum likelihood requires that all RTs be assigned a non-zero likelihood, outliers of this kind can disrupt fitting and estimation, whereas minimum chi-square is much less susceptible to such effects [31].

Many applications of the diffusion model have fitted it to group data, obtained by quantile-averaging the RT distributions across participants. A group data set is created by averaging the corresponding quantiles, $Q_{i,p}$, for each distribution of correct responses and errors in each experimental condition across participants. The choice probabilities in each condition are also averaged across participants. The advantage of group data is that it is less noisy and variable than individual data. A potential concern when working with group data is that quantile averaging may distort the shapes of the individual distributions, but in practice, the model appears to be robust to averaging artifacts. Studies comparing fits of the model to group and individual data have found that both methods lead to similar conclusions. In particular, the averages of the parameters estimated by fitting the model to individual data agree fairly well with the parameters estimated by fitting the model to quantile-averaged group data [32, 33]. Although the effects of averaging have not been formally characterized, the robustness of the model to averaging may be a result of the relative invariance of its families of distribution shapes, discussed previously.

3.9 The Psychophysical Basis of Drift

The diffusion model has been extremely successful in characterizing performance in a wide variety of speeded perceptual and cognitive tasks, but it does so by assuming that all of the information in the stimulus can be represented by a single value of drift, which is a free parameter of the model, and that the time course of the stimulus encoding processes that determine the drift can be subsumed within the non-decision time, T_{er}, which is also a free parameter. Recent work has sought to characterize the perceptual, memory, and attentional processes involved in the computation of drift and how the time course of these processes affects the time course of decision making [34].

Developments in this area have been motivated by recent applications of the diffusion model to psychophysical discrimination tasks, in which stimuli are presented very briefly, often at very low levels of contrast and followed by backward masks to limit stimulus persistence. Surprisingly, performance in these tasks is well described by the standard diffusion model, in which the drift rate is constant for the duration of an experimental trial [35, 36]. The RT distributions found in these tasks resemble those obtained from tasks with response-terminated stimuli, like those in Fig. 3.4, and show no evidence of increasing skewness at low stimulus discriminability, as would be expected if the decision process were driven by a decaying perceptual trace. The most natural interpretation of this finding is that the drift rate in the decision process depends on a durable representation of the stimulus stored in visual short-term memory (VSTM), which preserves the information it contains for the duration of an experimental trial.

This idea was incorporated in the *integrated system model* of Smith and Ratcliff [34], which combines submodels of perceptual encoding, attention, VSTM, and decision-making in a continuous-flow architecture. It assumes that transient stimulus information encoded by early visual filters is transferred to VSTM under the control of spatial attention and the rate at which evidence is accumulated by the decision process depends on the time-varying strength of the VSTM trace. Because the VSTM trace is time-varying, the decision process in the model is *time-inhomogeneous*. Predictions for time-inhomogeneous diffusion processes cannot be obtained using the infinite-series method, but can be obtained using either the integral equation method [16] or the Markov chain approximation [23]. The integrated system model has provided a good account of performance in tasks in which attention is manipulated by spatial cues and discriminability is limited by varying stimulus contrast or backward masks. It has also provided a theoretical link between stimulus contrast and drift rates, and an account of the shifts in RT distributions that occur when stimuli are embedded in dynamic noise, which is one of the situations in which the standard model fails [28, 37]. The main contribution of the model to our understanding of simple decision tasks is to show how performance in these tasks depends on the time course of processes of perception, memory, attention, and decision-making acting in concert.

3.10 Conclusion

Recently, there has been a burgeoning of interest in the diffusion model and related models in psychology and in neuroscience. In psychology, this has come from the realization that the model can provide an account of the effects of stimulus information, response bias, and response caution (speed-accuracy tradeoff) on performance in simple decision tasks, and a way to characterize these components of processing quantitatively in populations and in individuals. In neuroscience, it has come from studies recording from single cells in structures of the oculomotor systems of awake behaving monkeys performing saccade-to-target decision tasks. Neural firing rates in these structures are well-characterized by assuming that they provide an online read-out of the process of accumulating evidence to a response criterion [38]. This interpretation has been supported by the finding that the parameters of a diffusion model estimated from monkeys' RT distributions and choice probabilities can predict firing rates in the interval prior to the overt response [39, 40]. These results linking behavioral and neural levels of analysis have been accompanied by theoretical analyses showing how diffusive evidence accumulation at the behavioral level can arise by aggregating the information carried in individual neurons across the cells in a population [41, 42].

There has also been recent interest in investigating alternative models that exhibit diffusive, or diffusion-like, model properties. Some of these investigations have been motivated by a quest for increased neural realism, and the resulting models have included features like racing evidence totals, decay, and mutual inhibition [43]. Although arguments have been made for the importance of such features in a model, and although these models have had some successes, none has yet been applied as systematically and as successfully to as wide a range of experimental tasks as has the standard diffusion model.

Exercises

Simulate a random walk with normally-distributed increments in Matlab, R, or some other software package. Use your simulation to obtain predicted RT distributions and choice probabilities for a range of different accumulation rates (means of the random variables, Z_i). Use a small time step of, say, 0.001 s to ensure you obtain a good approximation to a diffusion process and simulate 5000 trials or more for each condition. In most experiments to which the diffusion model is applied, decisions are usually made in around a second or less, so try to pick parameters for your simulation that generate RT distributions on the range 0–1.5 s.

1. The drift rate, ξ, and the infinitesimal standard deviation, s, of a diffusion process describe the change occurring in a unit time interval (e.g., during one second). If ξ_{rw} and s_{rw} denote, respectively, the mean and standard deviation of the distribution of increments, Z_i, to the random walk, what values must they be set to

in order to obtain a drift rate of $\xi = 0.2$ and an infinitesimal standard deviation of $s = 0.1$ in the diffusion process? (Hint: The increments to a random walk are independent and the means and variances of sums of independent random variables are both additive).

2. Verify that your simulation yields unimodal, positively-skewed RT distributions like those in Fig. 3.1. What is the relationship between the distribution of correct responses and the distribution of errors? What does this imply about the relationship between the mean RTs for correct responses and errors?
3. Obtain RT distributions for a range of different drift rates. Drift rates of $\xi = \{0.4, 0.3, 0.2, 0.1\}$ with a boundary separation $a = 0.1$ are likely to be good choices with $s = 0.1$. Calculate the 0.1, 0.3, 0.5, 0.7, and 0.9 quantiles of the distributions of RT for each drift rate. Construct a Q-Q (quantile-quantile) plot by plotting the quantiles of the RT distributions for each of the four drift conditions on the y-axis against the quantiles of the largest drift rate (e.g., $\xi = 0.4$) condition on the x-axis. What does a plot of this kind tell you about the families of RT distributions predicted by a model?
4. Compare the Q-Q plot from your simulation to the empirical Q-Q plots reported by Ratcliff and Smith [28] in their Fig. 20. What do you conclude about the relationship?
5. Read Wagenmakers and Brown [17]. How does the relationship they identify between the mean and variance of empirical RT distributions follow from the properties of the model revealed in the Q-Q plot?

Further Reading

Anyone wishing to properly understand the RT literature should begin with Luce's (1986) classic monograph, *Response Times* [2]. Although the field has developed rapidly in the years since it was published, it remains unsurpassed in the depth and breadth of its analysis. Ratcliff's (1978) *Psychological Review* article [5] is the fundamental reference for the diffusion model, while Ratcliff and Smith's (2004) *Psychological Review* article [6] provides a detailed empirical comparison of the diffusion model and other sequential-sampling models. Smith and Ratcliff's (2004) *Trends in Neuroscience* article [38] discusses the emerging link between psychological models of decision-making and neuroscience.

References

1. Link, SW (1992) The wave theory of difference and similarity. Erlbaum, Englewood Cliffs
2. Luce RD (1986) Response times. Oxford University Press, New York
3. Townsend JT, Ashby FG (1983) Stochastic modeling of elementary psychological processes. Cambridge University Press, Cambridge
4. Vickers D (1979) Decision processes in visual perception. Academic, London

5. Ratcliff R (1978) A theory of memory retrieval. Psychol Rev 85:59–108
6. Ratcliff R, Smith PL (2004) A comparison of sequential-sampling models for two choice reaction time. Psychol Rev 111:333–367
7. Cox DR, Miller HD (1965) The theory of stochastic processes. Chapman & Hall, London.
8. Wald A (1947) Sequential analysis. Wiley, New York
9. Laming DRJ (1968) Information theory of choice reaction time. Wiley, New York
10. Link SW, Heath RA (1975) A sequential theory of psychological discrimination. Psychometrika 40:77–105
11. Gardiner CW (2004) Handbook of stochastic methods, 3rd edn. Springer, Berlin
12. Wiener N (1923) Differential space. J Math Phys 2:131–174
13. Feller W (1967) An introduction to probability theory and its applications, 3rd edn. Wiley, New York
14. Busemeyer J, Townsend JT (1992) Fundamental derivations from decision field theory. Math Soc Sci 23:255–282
15. Busemeyer J, Townsend JT (1993) Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. Psychol Rev 100:432–459
16. Smith PL (2000) Stochastic dynamic models of response time and accuracy: a foundational primer. J Math Psychol 44:408–463
17. Wagenmakers E-J, Brown S (2007) On the linear relationship between the mean and standard deviation of a response time distribution. Psychol Rev 114:830–841
18. Ashby FG (1983) A biased random walk model for two choice reaction time. J Math Psychol 27:277–297
19. Ratcliff R, Van Zandt T, McKoon G (1999) Connectionist and diffusion models of reaction time. Psychol Rev 106:261–300
20. Leite FP, Ratcliff R (2011) What cognitive processes drive response biases? A diffusion model analysis. Judgm Decis Mak 6:651–687
21. Buonocore A, Giorno V, Nobile AG, Ricciardi L (1990) On the two-boundary first-crossing-time problem for diffusion processes. J Appl Probab 27:102–114
22. Tuerlinckx F, Maris E, Ratcliff R, De Boeck P (2001) A comparison of four methods for simulating the diffusion process. Behav Res Methods Instrum Comput 33:443–456
23. Diederich A, Busemeyer JR (2003) Simple matrix methods for analyzing diffusion models of choice probability, choice response time, and simple response time. J Math Psychol 47:304–322
24. Bhattacharya RB, Waymire EC (1990) Stochastic processes with applications. Wiley, New York
25. Vandekerckhove J, Tuerlinckx F (2008) Diffusion model analysis with MATLAB: a DMAT primer. Behav Res Methods 40:61–72
26. Wiecki TV, Sofer I, Frank MJ (2013) HDDM: Hierarchical Bayesian estimation of the drift-diffusion model in Python. Front Neuroinformatics 7:1–10
27. Voss A, Voss J (2008) A fast numerical algorithm for the estimation of diffusion model parameters. J Math Psychol 52:1–9
28. Ratcliff R, Smith PL (2010) Perceptual discrimination in static and dynamic noise: the temporal relationship between perceptual encoding and decision making. J Exp Psychol Gen 139:70–94
29. Ratcliff R (2002) A diffusion model account of response time and accuracy in a brightness discrimination task: fitting real data and failing to fit fake but plausible data. Psychon Bull Rev 9:278–291
30. Nelder JA, Mead R (1965) A simplex method for function minimization. Comput J 7:308–313
31. Ratcliff R, Tuerlinckx F (2002) Estimating parameters of the diffusion model: approaches to dealing with contaminant reaction times and parameter variability. Psychon Bull Rev 9:438–481
32. Ratcliff R, Thapar A, McKoon G (2003) A diffusion model analysis of the effects of aging on brightness discrimination. Percept Psychophys 65:523–535
33. Ratcliff R, Thapar A, McKoon G (2004) A diffusion model analysis of the effects of aging on recognition memory. J Mem Lang 50:408–424

34. Smith PL, Ratcliff R (2009) An integrated theory of attention and decision making in visual signal detection. Psychol Rev 116:283–317
35. Ratcliff R, Rouder J (2000) A diffusion model account of masking in two-choice letter identification. J Exp Psychol Hum Percept Perform 26:127–140
36. Smith PL, Ratcliff R, Wolfgang BJ (2004) Attention orienting and the time course of perceptual decisions: response time distributions with masked and unmasked displays. Vis Res 44:1297–1320
37. Smith PL, Ratcliff R, Sewell DK (2014) Modeling perceptual discrimination in dynamic noise: time-changed diffusion and release from inhibition. J Math Psychol 59:95–113
38. Smith PL, Ratcliff R (2004) Psychology and neurobiology of simple decisions. Trends Neurosci 27:161–168
39. Ratcliff R, Cherian A, Segraves M (2003) A comparison of macaque behavior and superior colliculus neuronal activity to predictions from models of simple two-choice decisions. J Neurophysiol 90:1392–1407
40. Ratcliff R, Hasegawa Y, Hasegawa R, Smith PL, Segraves M (2007) A dual diffusion model for single cell recording data from the superior colliculus in a brightness discrimination task. J Neurophysiol 97:1756–1797
41. Smith PL (2010) From Poisson shot noise to the integrated Ornstein-Uhlenbeck process: Neurally-principled models of diffusive evidence accumulation in decision-making and response time. J Math Psychol 54:266–283
42. Smith PL, McKenzie CRL (2011) Diffusive information accumulation by minimal recurrent neural models of decision-making. Neural Comput 23:2000–2031
43. Usher M, McClelland JL (2001) The time course of perceptual choice: the leaky, competing accumulator model. Psychol Rev 108:550–592

Chapter 4
An Introduction to Human Brain Anatomy

Birte U. Forstmann, Max C. Keuken and Anneke Alkemade

If you want to understand function, study structure.
(Swaab [75])

Abstract This tutorial chapter provides an overview of the human brain anatomy. Knowledge of brain anatomy is fundamental to our understanding of cognitive processes in health and disease; moreover, anatomical constraints are vital for neurocomputational models and can be important for psychological theorizing as well. The main challenge in understanding brain anatomy is to integrate the different levels of description ranging from molecules to macroscopic brain networks. This chapter contains three main sections. The first section provides a brief introduction to the neuroanatomical nomenclature. The second section provides an introduction to the different levels of brain anatomy and describes commonly used atlases for the visualization of functional imaging data. The third section provides a concrete example of how human brain structure relates to performance.

4.1 Introduction

The human brain is the most complex and fascinating organ of the human body. Over centuries, it has been studied by philosophers, physicians, anatomists, biologists, engineers, psychologists, and in recent times also by neuroscientists. In the Middle ages, anatomical training played a central role in medical education and knowledge of human anatomy was highly valued [60]. However, during early years of the last century and the rapidly developing field of empirical psychology, knowledge of brain anatomy was considered insufficient to understand cognitive functioning. For

B. U. Forstmann (✉) · M. C. Keuken · A. Alkemade
University of Amsterdam, Cognitive Science Center Amsterdam,
Nieuwe Achtergracht 129, 1018 Amsterdam, The Netherlands
e-mail: buforstmann@gmail.com

M. C. Keuken
e-mail: mckeuken@gmail.com

A. Alkemade
e-mail: jmalkemade@gmail.com

B. U. Forstmann, E.-J. Wagenmakers (eds.), *An Introduction to Model-Based Cognitive Neuroscience,* DOI 10.1007/978-1-4939-2236-9_4

many, brain anatomy became largely irrelevant to the development of psychological models of function and dysfunction [14, 15]. This position was stated succinctly by the American philosopher and cognitive scientist Jerry Fodor: '. . . if the mind happens in space at all, it happens somewhere north of the neck. What exactly turns on knowing how far north?' [16, 32].

In the last decade the advent of ultra-high field 7 Tesla (T) or higher magnetic resonance imaging (MRI) holds the promise to reinstate anatomy to its former important position. In vivo structural and functional brain measurements on the sub-millimeter scale allow researchers to zoom in on fine-grained features such as the layering of cortex [4, 28, 37, 39, 51, 85] and very small nuclei in the subcortex [1, 34, 47] without having to lift the skull. Interest in this new technology is rapidly growing and to date more than 50 ultra-high field MRI scanners are available for human brain research worldwide. However, the importance of studying anatomy with 'expensive' neuroimaging techniques continues to be criticized. Neurosceptics rightly highlight the deficiencies of implausible and theoretically uninspiring research findings (e.g., http://blogs.discovermagazine.com/neuroskeptic/) and one may wonder whether much knowledge has recently been added to the work of the great neuroanatomists of the past centuries.

Recent developments in neuroimaging have provided a powerful tool to the field of the neurosciences, which appears to have regained some of brain anatomy's popularity in neuroscientific research. Traditional postmortem neuroanatomical studies are being replaced, at least partially, by advanced non-invasive techniques that now allow one to study larger numbers, and more importantly, brains of healthy individuals. In its modern form, anatomical studies will continue to contribute to our understanding of brain anatomy and function.

In this chapter, we first provide a brief introduction to the nomenclature of neuroanatomy because it is all too easy to get lost in localization [20]. Next, we introduce descriptive, sectional, and connectional neuroantomy and list grey and white matter atlases that are widely applied to study systematically brain structure. Finally, a concrete example about function-structure relationships zooming in on individual differences will be presented.

4.2 Nomenclature

Nomenclature is a normalized system of exactly defined terms arranged according to specific classification principles and is essential in the field of anatomy [76]. The origins of the anatomical terminology date back to classical Greek and Roman antiquity over 2500 years ago, and explain the current use of Greek and Latin spelling. Due to the early development of terminology and nomenclature, anatomy is considered to be one of the first exact medical fields [42].

The official anatomical nomenclature that is currently used is in Latin and was created by the Federative Committee on Anatomical Terminology (FCAT) and approved by the International Federation of Associations of Anatomists (IFAA). It was

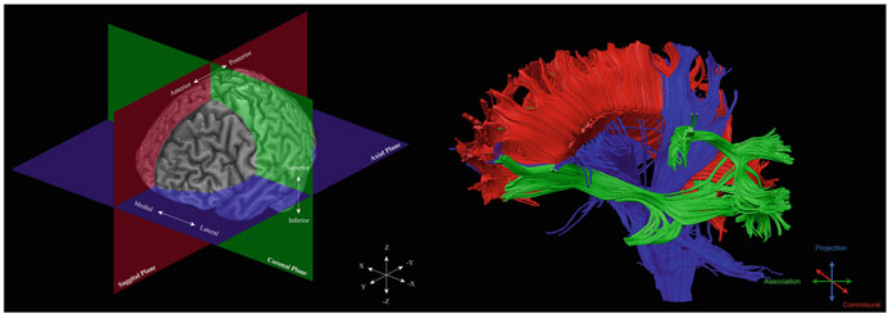


Fig. 4.1 (**a**) Conventions in anatomical terminology; The sagittal plane (*red*) divides the brain in a *left* and *right* part, the axial or transverse plane (*blue*) divides the brain in inferior and superior parts, and the coronal or frontal plane (*green*) is the vertical plane perpendicular to the sagittal plane. (**b**) Connectional neuroanatomy. The corpus callosum and anterior commissure represent the main commissural tracts connecting the *left* and *right* hemisphere (*red*). The associative pathways (*green*) connect specific cortical areas within a hemisphere (*ipsilateral connections*). The ascending and descending projection pathways connect cortical and subcortical structures (*blue*)

published a year later as the Terminologia Anatomica (TA); [30]. Since Latin is taught increasingly less in high schools [17], the TA has been translated in several different languages including English, Spanish, Russian, and Japanese.

Another commonly used way of describing the brain is by differentiating between *descriptive*, *sectional*, and *connectional* aspects of neuroanatomy (see also [16]).

Descriptive neuroanatomy of the nervous system can be defined as the process of identifying and labeling different parts of the brain and spinal cord [16]. The brain can be described from its surface including different views. The surface of the brain can be viewed from above (axial view), from the front (anterior or coronal view), the back (posterior or coronal view), as well as from the side (lateral or sagittal view; Fig. 4.1a). The same terminology is used to indicate different regions of the brain surface (e.g., superior frontal gyrus, Fig. 4.2). The terms distal, proximal, bilateral, unilateral, ipsilateral, and contralateral are used to indicate the location of an area relative to another area. In the following we provide some examples to clarify these terms. Place a finger in your elbow cavity and do the following: Move your finger towards your wrist. This movement is in *distal* direction compared to your elbow cavity. If you move your finger back to the elbow cavity then this movement is in *proximal* direction.

The term *bilateral* indicates that a brain area is represented in both hemispheres. The term *unilateral* refers to only one hemisphere. A well-known example for a unilateral representation is Broca's area, an area essential for language [23, 49]. In the majority of people this area is located in the left hemisphere.

Finally, consider a plane that cuts your body symmetrically into a left and right part. Any connection between two regions that remains on one side is called *ipsilateral*. If the connection crosses from the left to the right side or vice versa, this is called *contralateral*.

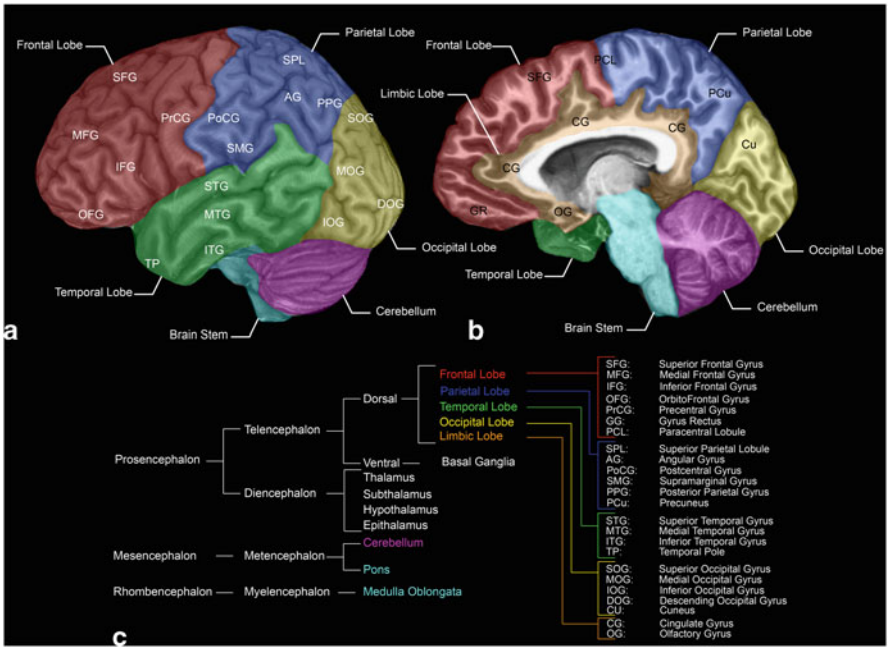


Fig. 4.2 Sectional anatomy. Cortical subdivisions (**a**) *lateral* view (**b**) *medial* view (**c**) brain ontogeny

Sectional neuroanatomy describes the relationship between cortical and subcortical structures, most commonly visualized along orthogonal axial, coronal, and sagittal planes (Fig. 4.1a). Axial planes divide the brain into an upper and lower part. In radiological convention, the axial slices are viewed from the feet upwards to the head. Consequently, on axial slices the left side of the brain is on the right side of the page, whereas in the neurological convention this mapping is inverted. Axial planes are also sometimes indicated as horizontal or transverse planes. To allow generalization and comparison of results, brains in neuroimaging studies are often displayed within a space of reference (e.g., [53, 54, 78] see also the section about grey and white matter atlases). The axial coordinates move in the Z-direction (negative values indicate slices inferior to the anterior commissure, while positive values indicate slices superior to the anterior commissure; Fig. 4.1a).

Coronal planes cut through the brain in the anterior to posterior direction. The coronal planes are conventionally oriented with the left side of the brain on the right side of the page (radiological convention). Slices anterior to the anterior commissure are indicated with positive values on the Y-axis. Finally, the midsagittal plane divides the brain into two hemispheres. Parasagittal slices through the left hemisphere are indicated by negative values on the X-axis.

Connectional neuroanatomy delineates the origin, course, and termination of connecting pathways. Post-mortem dissections of white matter tracts, i.e., the paler tissue of the brain and spinal cord, require special preparation and are particularly difficult

to perform. Recent developments in diffusion MRI tractography have revitalized the field. The tracts are classified according to their course and terminal projections (Fig. 4.1b). So-called commissural pathways run along a horizontal axis and connect the two hemispheres. The majority of projection pathways follow a vertical course along a dorso-ventral (descending) or ventro-dorsal (ascending) axis and connect the cerebral cortex to subcortical nuclei, cerebellum, and the spinal cord. Association tracts run longitudinally along an anterior-posterior axis from front to back and vice versa and connect cortical areas within the same hemisphere.

4.3 Different Levels of Brain Anatomy

For more than five centuries, the nervous system has been divided into grey matter and white matter [82]. Grey matter contains cell bodies, neuronal extensions including axons and dendrites as well as synapses between extensions, glia, and blood vessels (cf. [77]). Grey matter compartments consist of relatively low numbers of cell bodies and are composed of mostly unmyelinated axons, dendrites, and glia cell processes which together form a synaptically dense region which is called the neuropil (cf. [63]). White matter is indicated as such due to its white appearance as a result of mainly myelinated axons, although it also contains unmyelinated axons. White matter is used as a generic term for nervous system volumes where axons predominate connecting grey matter regions. It can also include neurons when it is in close proximity to grey matter [63, 72].

In this section we will describe in more detail sectional neuroanatomy including grey matter regions such as the cerebral cortex and the basal ganglia. Next connectional anatomy will be discussed including white matter compartments and pathways.

4.3.1 Sectional Anatomy

Sectioning of the brain in three orthogonal planes allows us to investigate anatomical structures located deep inside the brain. Deep brain structures include the ventricular system, deep grey nuclei, and a number of white and grey matter structures. The ventricular system is filled with cerebrospinal fluid (CSF) produced by the choroid plexus which provides a highly protective bath with great buffering capacity in which the brain is immersed. The ventricular system consists of a number of brain cavities which are interconnected and allow for a CSF flow towards the submeningeal space. The mediobasal part of the brain and the areas bordered by the lateral ventricles are made up by a number of grey matter nuclei. These nuclei are separated by many fiber tracts which form the cerebral white matter. The outer aspect of the cerebral white matter forms the cerebral cortex. Below we will describe the aforementioned structures in more detail.

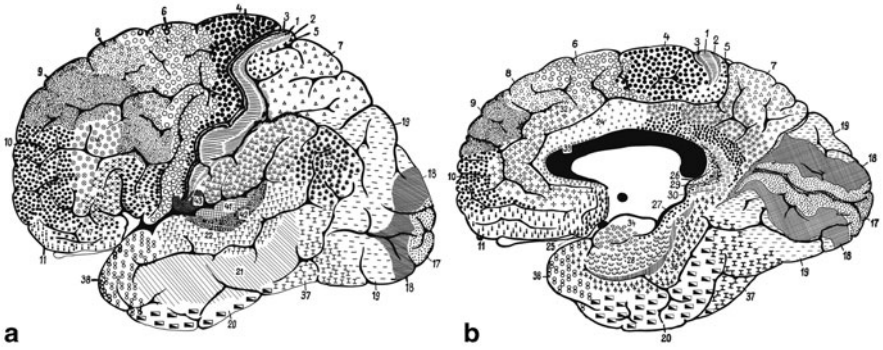


Fig. 4.3 Reproduction of the classical Brodmann map. (**a**) *lateral* view, (**b**) *medial* view. Individual numbers indicate Brodmann area numbers

4.3.1.1 Cerebral Cortex

The outer surface of the cerebral hemispheres is called cerebral cortex which consists of a densely folded strip of grey matter. The cerebral cortex displays a multi-layered organizational pattern. The architecture of the cortical layers varies across the brain. The majority of the cortex is six-layered (neocortex), and can be further parcellated in distinct areas or fields. The cytoarchitecture reveals variations in cell shape, size, and density; characteristics which are used as anatomical landmarks to divide the cortex into distinct areas.

According to the work of individual groups and authors, the number of areas that are distinguished varies from a minimum of 17 [13] to a maximum of 107 [26]. The most widely used maps are provided by Brodmann [9] (Fig. 4.3) and his original atlas contained 52 areas.

According to Brodmann's classification, the frontal lobe consists of eleven fields, which are grouped in five main regions. (1) Area 4 corresponds to the primary motor cortex, containing neuronal bodies, as well as cortico-spinal projection fibers which show a somatotopical organization; (2) Area six contains the premotor region and is subdivided into the lateral premotor cortex (PMC), and the pre-supplementary motor area (pre-SMA); (3) Area 44 and 45 correspond to Broca's area; (4) Area 8–10, and 46 include dorsolateral prefrontal areas, and 47 to the ventrolateral prefrontal corex; (5). Finally, areas 11 and 47 represent the main parts of the orbitofrontal cortex.

The parietal lobe is divided into four regions consisting of a total of nine separate fields by Brodmann. (1) Areas 1–3 correspond to the somatosensory cortex and its cytoarchitecture strongly resembles that of the primary motor cortex; (2) The more laterally located areas 5 and 7 together form the superior polymodal parietal cortex; (3) Areas 39 and 40 are located in the inferior polymodal parietal cortex, corresponding to the Geschwind's territory; (4) The medial parts of areas 31, 5, and 7 form the precuneus, and area 43 is considered a transition region of the fronto-parietal operculum.

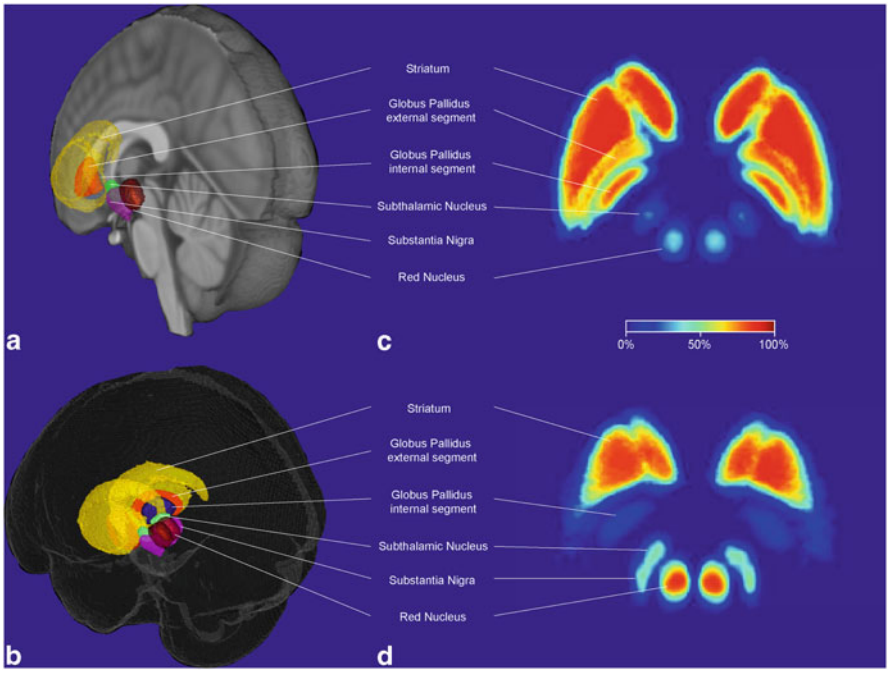


Fig. 4.4 Probabilistic Basal Ganglia atlas. (**a**) Probability maps for the striatum, globus pallidus external segment, globus pallidus internal segment, subthalamic nucleus, substantia nigra, and *red* nucleus of the *left* hemisphere. (**b**) Bilateral representation. (**c**) Axial presentation, level of the globus pallidus. The color intensity reflects the percentage overlap across the 30 participants. (**d**) Axial presentation, level of the *red* nucleus. The color intensity reflects the percentage overlap across the 30 participants. (adapted from [46])

The temporal lobe is also divided into four main regions consisting of seven separate fields. (1) Area 41 is the primary auditory cortex; (2) Adjacent is the auditory association cortex consisting of area 42 and 22, which in part overlays with Wernicke's area; (3) The temporal visual association cortex is formed by areas 20, 21, and 37; (4) Finally, area 38, one of the paralimbic areas, occupies the temporopolar cortex.

The occipital lobe consists of three areas (17–19). The primary visual cortex corresponds to area 17, and area 18 and 19 form the visual association cortex.

The limbic lobe includes as much as 15 fields. Areas 11, 24, 25, 32, 33, 36, and 47 form an olfactocentric group and areas 23, 26–31 form a hippocampocentric group. Additionally, many white matter tracts are present in the cerebral cortex with two main orientations: tangential and radial. Further myeloarchitectonic subdivision within the cerebral cortex reveals over 200 additional areas [84].

Basal Ganglia The basal ganglia are located deep in the white matter of the cerebral hemispheres anterior to the thalamus positioned medial to the lateral ventricles (Fig. 4.4). They consist of two major functional divisions, the striatum and, more medially located, the globus pallidus. The striatum in turn is composed of two highly interconnected masses, the putamen and caudate nucleus.

The caudate nucleus shows a somewhat elongated shape with a large anterior head, which becomes more narrow towards the thalamus and continues to narrow down and forward towards the temporal lobe along the wall of the lateral ventricle. Towards the antero-ventral part of putamen, distinguishing between the caudate nucleus and putamen around the inferior border of the internal capsule is difficult.

The globus pallidus is further divided into an external and internal segment. The putamen and globus pallidus together form the lentiform nucleus and are located lateral to the internal capsule. Finally, two areas that are functionally but not ontologically considered to be part of the basal ganglia are the subthalamic nucleus (STN) and the substantia nigra. The STN forms a small lens-shaped nucleus, which is high in iron content and a main target region for deep-brain stimulation for the treatment of refractory movement disorders, e.g., Parkinson's disease, and is located adjacent in anterior and lateral direction of the substantia nigra. The substantia nigra is a crescent shaped region that is one of the production sites of the neurotransmitter dopamine and the degeneration of this nucleus is a main hallmark of Parkinson's disease. The substantia nigra is named after its pigmented, melanin containing neurons, which allows it to be clearly distinguished in histological sections.

4.3.1.2 Cerebellum

The cerebellum, or "little brain", is located inferior of the occipital lobe and posterior to the brain stem. The functions of the cerebellum range from maintaining balance, coordination of half automatic movements to cognitive functions such as attention, language, and working memory [6, 38, 73]. Similar to the cerebrum, the cerebellum is distributed across the two hemispheres, which are connected through the vermis. Each hemisphere has three lobes: the anterior, posterior, and flocculonodular lobe. The cerebellum is connected to the brain stem via three peduncles namely the superior cerebellar peduncle, the middle cerebellar peduncle, and the inferior cerebellar peduncle connecting respectively to the midbrain, pons, and medulla oblongata. While the cerebellum occupies only 10 % of the total volume of the entire human brain, it contains many more neurons than the rest of the brain combined, such that for each neuron in the cerebral cortex there are on average 3.6 neurons in the cerebellum [40, 43].

4.3.2 Connectional Neuroanatomy

White matter tracts can be classified in different groups: tracts in the brainstem and projections, association, and commissural tracts in the cerebral hemispheres. Projection fibers connect cortical and subcortical gray matter, association tracts connect two cortical areas, and commissural tracts connect the brain's hemispheres.

Brain regions are connected by white matter tracts which vary across subjects in terms of their position, extent, and course [12]. Even though the regional pattern of connectivity is largely preserved across individuals and these afferent, i.e., conducting incoming, and efferent, i.e., conducting outgoing, pathways strongly influence

the information processing properties of individual brain regions [61]. Consequently, it is often desirable to understand regional activation patterns in terms of an underlying system of neural regions and their interactions [35, 36]. To this end, advances in diffusion weighted tractography, a technique that allows the measurement of white matter fibers, offer the potential for explicitly relating functional and connectional anatomy. For instance, regional connectivity patterns can be used to reliably identify both cortical [3, 41] and subcortical regions [5, 22, 50] even when these areas lack clear macroanatomical borders. Consequently, by collecting connectivity information along with functional data, it is possible to directly relate structure to function in individuals. The final section of this chapter provides a concrete example relating white matter connectivity to function by zooming in on interindividual differences.

Finally, the *brainstem* consists of five major white matter tracts: the superior, middle, and inferior cerebellar peduncles, the corticospinal tract, and the medial lemniscus.

Projection fibers connect the cortex with distant regions located in the lower parts of the brain and spinal cord. These fibers can be differentiated into two classes: the corticothalamic/thalamocortical fibers (collectively called thalamic radiations) and the long corticofugal (corticoefferent) fibers. The corticofugal fibers include such fibers as the corticopontine, corticoreticular, corticobulbar, and corticospinal tracts. These fibers all penetrate the internal capsule either between the thalamus and the putamen or between the caudate nucleus and the putamen. When approaching the cortex, they fan out forming the corona radiata. The thalamus is known to have reciprocal connections to many areas in cortex.

Association fibers connect different ipsilateral cortical areas and are classified into short and long association fibers. The former connect areas within the same cerebral hemisphere and include the fibers connecting adjacent gyri, so-called U-fibers. The long association fibers connect different lobes, forming prominent fiber bundles. Some of the main association fibers are (1) the superior longitudinal or arcuate fasciculus connecting the frontal, temporal, and parietal lobe; (2) the inferior longitudinal fasciculus connecting the temporal and occipital lobe; (3) the superior fronto-occipital fasciculus, connecting the frontal and parietal lobe; (4) the inferior fronto-occipital fasciculus connecting the orbital cortex to the ventral occipital lobe, and; (5) the uncinate fasciculus connecting the anterior temporal lobe and lateral orbital frontal regions. Three major fibers connecting the limbic system (hippocampus), namely (1) the cingulum connecting the prefrontal, parietal and occipital cortex to the temporal lobe and hippocampus; (2) the stria terminalis connecting the hypothalamus to the amygdala, and; (3) the fornix, which is a projection fiber, connecting the medial temporal lobe to the mamillary bodies and hypothalamus.

Commissural fibers interconnect the two cerebral hemispheres and contain more than 300 million axons [58]. Most of the commissures interconnect homologous cortical areas in roughly mirror-image sites, but a substantial number have heterotopic connections ending in asymmetrical areas. The largest of the commissural fiber tracts is the corpus callosum, which connects the neocortical hemispheres. Inferior to the corpus callosum are two other commissures, namely the anterior and the posterior commissure (AC and PC respectively). The anterior commissure connects the bilateral anterior and ventral temporal lobes. An imaginary line between connecting these

two points, the AC-PC line, is often used for MRI-analyses. See Catani and de Schotten [15] for an in vivo atlas of the projection, association, and commissural fibers.

4.4 Neuroanatomical Atlases

Neuroanatomical atlases are a useful tool because they provide a common reference framework for structural and functional MRI studies. This is important in light of the substantial variability between individual healthy brains [44, 79], even in identical twins [80]. For example, all humans have a transverse gyrus (known as Heschl's gyrus) in the superior temporal lobe that is associated with the primary auditory cortex, but the number and size of these transverse gyri varies across individuals [62, 64]. Similar variability is observed in other cortical regions [2, 19, 59, 65, 66]. The shape and relative locations of subcortical structures also differ between individuals and change with age [24, 47, 48]. The same can be observed for commissural fibers. This variability in brain structure across individuals is of critical importance because it implicates that no two twin brains are identical at a macroscopic level and therefore no single brain can be considered representative of a population. Consequently, any atlas based on a single 'template' brain will necessarily provide limited accuracy. In theory, it may be possible to determine a topological transformation that could morph one brain into precise correspondence with another, although current spatial normalization procedures tend to correct only for gross anatomical variability.

4.4.1 Grey Matter and White Matter Atlases

Prior to the development of MRI scanners, atlases were solely based on the analysis of post-mortem tissue. Classical examples include for instance the work by Brodmann [9]. As described earlier, Brodmann's atlas is based on cyotarchitectonics, i.e., the density and types of cells present in different cortical layers. Other cytoarchitectonic atlases have been published by von Economo and Koskinas [26], Ngowyang [57], Sarkisov and colleagues [69], and Braak [8]. Strikingly, all these atlases vary substantially in the number of areas, the size of the areas, and the exact location of delineated areas. The most recent cytoarchitectonic atlas is provided by Duvernoy which covers both cortical and subcortical brain areas including the brain stem [25].

A different class of atlases is based on the myeloarchitecture, i.e., the distribution of myelinated axons in the brain. The most prominent myeloarchitecture atlases were published by Reil [67, 68], Burdach [11], Smith [71], Flechsig [31], Vogt [83], and Strasburger [74].

Other classes of atlases that are used in the cognitive neurosciences are based on non-human primate data. An example of such an atlas is provided by Schmahmann and Pandya [70] which includes tracer sections. Postmortem sections were analyzed after the in vivo injection with chemical tracers.

4.4.2 Standard Space Atlases

With the advent of MRI, the need for a common reference space increased because a common reference allows the comparison across individuals' functional and structural brain maps. A first attempt at a standard 3D space was developed by Talairach and Tournoux [78] which was based on a single post-mortem brain from a 60-year old woman. This brain template, together with the anatomical labels, provided the opportunity to compare individuals' functional brain maps in a common reference space. However, while this approach was considered state of the art for many years, further progression of the research field now allows for better alternatives. In the beginning of the 1990s, a template was developed which has now become the standard space in MRI research: the Montreal NeuroImaging (MNI) template. Several versions of this template exist with the most recent one including 452 subjects' anatomical MRI scans [29] specifying 142 brain areas. Other MRI atlases are based on macroscopic delineations of cortical and subcortical areas, such as the Harvard-Oxford atlas which includes 96 brain areas and is created based on 37 subjects [21, 52]. Eickhoff and colleagues [27] created an atlas which is based on cytoarchitectonic properties and includes ten ex vivo brains. More recently, specialized atlases partly relying on ultra-high resolution 7 T MRI have been developed including the STN [34, 47] (http://www.nitrc.org/projects/atag), the locus coeruleus [46], and the thalamus [56].

4.5 Structure-Function Relationships

In the final section of this chapter, we will discuss a concrete example how structure-function relationships can be established in a model-based framework. This example focuses on interindividual differences in the speed-accuracy tradeoff (SAT), both in behavior and in structural features of the human brain. The SAT describes a behavioral regularity that explains why the temporal benefits of responding more quickly come at the cost of making more errors. A possible mechanism for the SAT is that the response threshold can be adjusted depending on the strategy. If the response needs to be very accurate, a high response threshold is adopted so that sufficient information can be accumulated. If a speeded response is required, a low response threshold ensures responses that are quick but error-prone.

Non-human primate and human studies have provided handles on how the mechanism of flexibly adjusting response thresholds could be implemented in the human brain (for a review see [7]). Specifically, the anatomical substrate that could subserve the SAT mechanism is the interplay between cortex and the basal ganglia. There are two main pathways that connect the cortex to the basal ganglia namely the cortical-striatal pathway and the cortical-subthalamic pathway. The striatal hypothesis poses that an emphasis on speed promotes excitatory input from cortex to striatum; the increased baseline activation of the striatum acts to decrease the inhibitory control that the output nuclei of the basal ganglia exert over the brain, thereby facilitating faster but possibly premature responses. Alternatively, the STN hypothesis posits

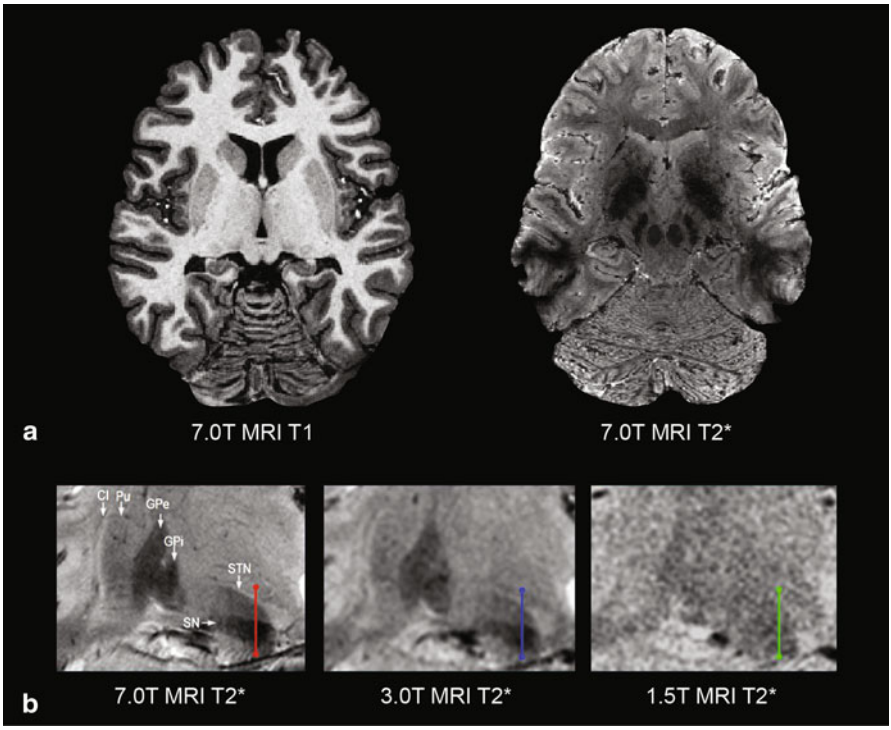


Fig. 4.5 Visualization of the STN with different MRI contrasts at different field strengths. (**a**) The *left* panel shows a 7 T T1 weighted anatomical MRI scan. Note the low contrast in the midbrain. The *right* panel shows a 7 T T2* weighted anatomical scan. Note the improved contrast in the subcortical regions. (**b**) The *lower left* panel shows a 7 T T2* weighted anatomical scan. The use of high field strength yields excellent contrast allowing differentiation between individual subcortical nuclei including the STN and SN. The *middle* panel shows a 3 T T2* weighted anatomical scan and the *lower right* panel shows a 1.5 T T2* weighted anatomical scan for comparison. (adapted from [18])

that an emphasis on accuracy promotes excitatory input from cortex (e.g., anterior cingulate cortex) to the STN; increased STN activity may lead to slower and more accurate choices.

These two hypotheses result in competing anatomical predictions. In particular, the striatal hypothesis predicts that participants who display better control of the SAT have stronger white matter tract connections between cortex and striatum, whereas the STN hypothesis predicts stronger white matter connections between cortex and STN. Testing of these distinct hypotheses has been technically challenging, but a first step has been made by visualization of the STN using 7 T MRI [33, 34, 47]. Since the STN is a very small lens-shaped nucleus in the subcortex, it is barely visible on 3 T anatomical MRI scans (Fig. 4.5). From an anatomical point of view, the technical development of ultra-high resolution MRI has therefore already been proven crucial.

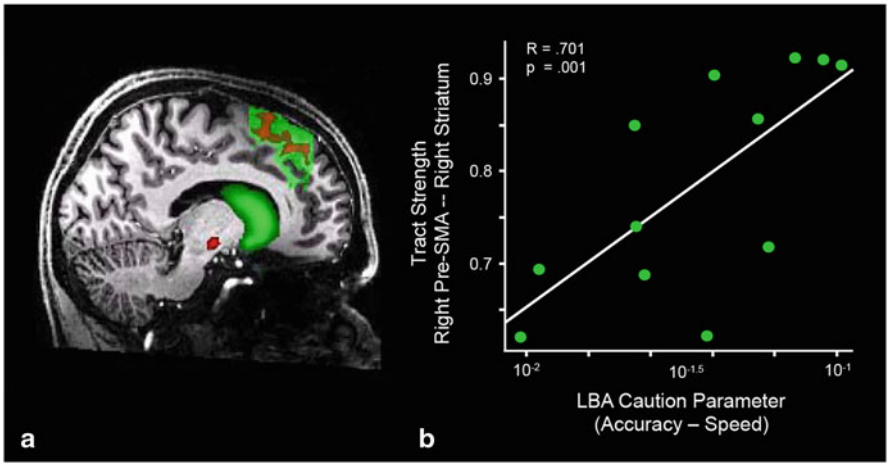


Fig. 4.6 Structural differences in brain connectivity predict individual differences in decision making. (**a**) Connectivity-based seed classification for the pre-SMA projecting into the striatum (*green*) and STN (*red*). (**b**) Individual differences in tract strength between right pre-SMA and right striatum are associated with flexible adjustments of the speed-accuracy tradeoff. (adapted from [33])

The STN contains iron, yielding hypointensities which in turn results in a loss of the MR signal. Moreover, there are large interindividual differences in the STN's location [47]. (Fig. 4.5)

Next, diffusion weighted imaging (DWI) was applied to calculate the tract strength between cortex and striatum and cortex and STN, respectively. Additionally, behavioral data from a moving dots kinematogram with a SAT manipulation was analyzed. The behavioral data were then modeled using the linear ballistic accumulator (LBA) model to account for latent cognitive processes such as individuals' response caution ([10] see also the chapter 1B Cognitive models in action). Statistical model selection techniques confirmed that the LBA explained the data best when only the response caution parameter was free to vary between the speed and the accuracy condition. Interindividual differences in efficacy of changing response caution were quantified by the differences in LBA threshold estimates between both conditions. These differences were then related to participants' tract strength measures between pre-SMA and striatum and pre-SMA and STN, respectively.

Results showed that individual tract strength between pre-SMA and striatum translate to individual differences in the efficacy with which people adjust their response thresholds (Fig. 4.6). This supports the striatal hypothesis of how the brain regulates the competing demands for speed vs. accuracy and show that individual differences in brain connectivity affect decision making even in simple perceptual tasks. The findings also show that, inconsistent with the STN hypothesis of SAT, there is little or no evidence that strong connections from any cortical region to STN lead to more flexibility in threshold settings (see [33] for more details regarding these findings).

Importantly, this research could only be performed using ultra-high resolution 7 T MRI, which allowed accurate localization of the STN.

In sum, this example highlights structure-function relationships as revealed through individual differences in brain connectivity. This approach provides a window to human cognition that exploits individual anatomical variability and complements popular methods such as functional MRI.

4.6 Concluding Comments

In contemporary neuroscience, the anatomy of the central nervous system has essentially become the anatomy of the connections between brain regions. Understanding patterns of connectivity and how neurons communicate is probably one of the next frontiers in the neurosciences [55]. It is important to delineate fine-grained cerebral and subcortical structures using ultra-high resolution MRI [81]. Ultimately, by linking information from large-scale networks with macro- and microanatomical knowledge we can reach a deeper understanding of descriptive, sectional, and connectional neuroanatomy.

Finally, capturing the interindividual variability in the form of probabilistic atlases and maps both from in vivo and postmortem brains holds a great promise. It will facilitate our understanding of functional neuroanatomy including changes associated with development, aging, learning, and disease. Informing mathematical and neurocomputational models of the human brain with the anatomical knowledge gained by using ultra-high resolution MRI promises exciting new avenues of understanding human cognition.

Exercises

Q(1)

Step (1) Imagine a human brain floating in front of you with the eyes pointing towards the cutter.

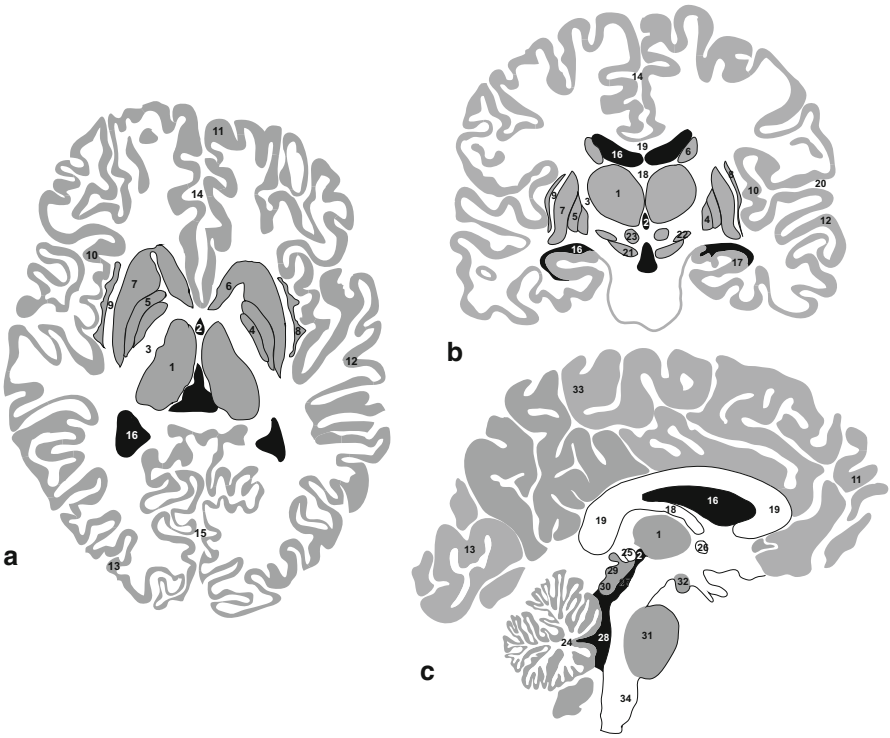
Step (2) Take out your imaginary knife and make the following cuts:

a. Cut the midline.
b. Take the left part and cut it in half again so that you have a superior and inferior part.
c. Take the inferior part and cut it again in half so that you have an anterior and posterior part.

Q1(a) Which two parts did you have after step 2a?
Q1(b) In which part is the cerebellum located after step 2b?
Q1(c) In which of the two parts is the temporal pole located after step 2c?

Q(2)

Q2(a) Which panel shows the brain in an axial view?
Q2(b) Which panel shows the brain in a sagittal view?
Q2(c) Name all the different anatomical structures in the following figure:



Q(3) What is the difference between myeloarchitecture and cytoarchitecture?

Further Reading

1. For a more in depth textbook on anatomy see Catani and de Schotten [16]. *Atlas of human brain connections* (1st ed., pp. 1–515). Oxford University Press.
2. See http://brancusi.usc.edu for an extensive resource for information about neural circuitry.
3. See www.brain-map.org for a large online collection of neuroanatomical atlases, gene expression and cross-species comparisons.
4. See http://www.unifr.ch/ifaa/Public/EntryPage/ViewTAOnLine.html for the anatomical nomenclature.
5. See https://bigbrain.loris.ca/main.php for an ultra-high resolution image of a post-mortem stained brain.
6. See http://www.appliedneuroscience.com/Brodmann.pdf for the English translation of the seminal work by Korbinian Brodmann.

References

1. Alkemade A, Keuken MC, Forstmann BU (2013) A perspective on terra incognita: uncovering the neuroanatomy of the human subcortex. Front Neuroanat 3:7–40
2. Amunts K, Zilles K (2012) Architecture and organizational principles of Broca's region. Trend Cogn Sci 16(8):418–426
3. Anwander A, Tittgemeyer M, Cramon von D, Friederici A, Knosche T (2006) Connectivity-based parcellation of Broca's area. Cereb Cortex 17(4):816–825
4. Bazin P-L, Weiss M, Dinse J, Schäfer A, Trampel R, Turner R (2013) A computational framework for ultra-high resolution cortical segmentation at 7 Tesla. Neuroimage 2:201–209
5. Behrens T, Johansen-Berg H, Woolrich MW, Smith SM, Wheeler-Kingshott C, Boulby PA et al (2003) Non-invasive mapping of connections between human thalamus and cortex using diffusion imaging. Nat Neurosci 6(7):750–757
6. Berquin PC, Giedd JN, Jacobsen LK, Hamburger SD, Krain AL, Rapoport JL, Castellanos FX (1998) Cerebellum in attention-deficit hyperactivity disorder: a morphometric MRI study. Neurology 50(4):1087–1093
7. Bogacz R, Wagenmakers E-J, Forstmann BU, Nieuwenhuis S (2010) The neural basis of the speed-accuracy tradeoff. Trend Neurosci 33(1):10–16
8. Braak H (1980) Architectonics of the human telencephalic cortex. Springer, Berlin
9. Brodmann K (1909) Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt aufgrund des Zellenbaues. Johann Ambrosius Barth, Leipzig
10. Brown SD, Heathcote A (2008) The simplest complete model of choice response time: linear ballistic accumulation. Cogn Psychol 57(3):153–178
11. Burdach KF (1819) Vom Baue und Leben des Gehirns. Dyk'schen Buchhandlung, Leipzig, pp 1–285
12. Bürgel U, Amunts K, Hoemke L, Mohlberg H, Gilsbach JM, Zilles K (2006) White matter fiber tracts of the human brain: three-dimensional mapping at microscopic resolution, topography and intersubject variability. Neuroimage 29(4):1092–1105
13. Campbell AW (1905) Histological studies on the localisation of cerebral function. Cambridge University Press, Cambridge
14. Catani M, Ffytche DH (2010) On the study of the nervous system and behaviour. Cortex 46(1):106–109
15. Catani M, de Schotten MT (2008) A diffusion tensor imaging tractography atlas for virtual in vivo dissections. Cortex 44(8):1105–1132
16. Catani M, de Schotten MT (2012) Atlas of human brain connections, 1st edn. Oxford University Press, Oxford, pp 1–515
17. Cha Y-K (1991) Effect of the global system on language instruction, 1850–1986. Sociol Educ 64(1):19–32
18. Cho ZH, Min HK, Oh SH, Han JY, Park CW, Chi JG et al (2010) Direct visualization of deep brain stimulation targets in Parkinson disease with the use of 7-tesla magnetic resonance imaging. J Neurosurg 113:1–9
19. Choi H-J, Zilles K, Mohlberg H, Schleicher A, Fink GR, Armstrong E, Amunts K (2006) Cytoarchitectonic identification and probabilistic mapping of two distinct areas within the anterior ventral bank of the human intraparietal sulcus. J Compar Neurol 495(1):53–69
20. Derrfuss J, Mar RA (2009) Lost in localization: the need for a universal coordinate database. Neuroimage 48(1):1–7
21. Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D et al (2006) An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 31(3):968–980
22. Devlin JT, Sillery EL, Hall DA, Hobden P, Behrens TEJ, Nunes RG et al (2006) Reliable identification of the auditory thalamus using multi-modal structural analyses. Neuroimage 30(4):1112–1120
23. Dronkers NF, Plaisant O, Iba-Zizen MT, Cabanis EA (2007) Paul Broca's historic cases: high resolution MR imaging of the brains of Leborgne and Lelong. Brain 130(5):1432–1441

24. Dunnen DWF, Staal MJ (2005) Anatomical alterations of the subthalamic nucleus in relation to age: a postmortem study. Mov Disord 20(7):893–898
25. Duvernoy MH (1999) The human brain, 2nd edn. Springer, Wien
26. Economo C, Koskinas GN (1925) Die Cytoarchitektonik der Hirnrinde des erwachsenen Menschen. Springer, Wien
27. Eickhoff SB, Stephan KE, Mohlberg H, Grefkes C, Fink GR, Amunts K, Zilles K (2005a) A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. Neuroimage 25(4):1325–1335
28. Eickhoff S, Walters NB, Schleicher A, Kril J, Egan GF, Zilles K et al (2005b) High-resolution MRI reflects myeloarchitecture and cytoarchitecture of human cerebral cortex. Hum Brain Mapp 24(3):206–215
29. Evans AC, Janke AL, Collins DL, Baillet S (2012) Brain templates and atlases. Neuroimage 62(2):911–922
30. Federative Committee on Anatomical Terminology (1998) Terminologia Anatomica. (Federative Committee on Anatomical Terminology, Ed.). Thieme, New York, pp 1–292
31. Flechsig P (1920) Anatomie des menschlichen Gehirns und Rückenmarks auf myelogenetischer Grundlage. Thieme, Leipzig, pp 1–121
32. Fodor J (1999) Diary. London Rev Books 21(9):68–69. http://www.lrb.co.uk/v21/n19/jerry-fodor/diary. Accessed 1 April 2013
33. Forstmann BU, Anwander A, Schafer A, Neumann J, Brown S, Wagenmakers E-J et al (2010) Cortico-striatal connections predict control over speed and accuracy in perceptual decision making. Proc Natl Acad Sci U S A 107(36):15916–15920
34. Forstmann BU, Keuken MC, Jahfari S, Bazin PL, Neumann N, Schafer A et al (2012) Cortico-subthalamic white matter tract strength predict interindividual efficacy in stopping a motor response. Neuroimage 60:370–375
35. Friston K (2002a) Beyond phrenology: what can neuroimaging tell us about distributed circuitry? Ann Rev Neurosci 25(1):221–250
36. Friston KJ (2002b) Bayesian estimation of dynamical systems: an application to fMRI. Neuroimage 16(2):513–530
37. Geyer S, Weiss M, Reimann K, Lohmann G, Turner R (2011) Microstructural parcellation of the human cerebral cortex-from Brodmann's post-mortem map to in vivo mapping with high-field magnetic resonance imaging. Front Hum Neurosci. doi:10.3389/fnhum.2011.00019
38. Gottwald B, Mihajlovic Z, Wilde B, Mehdorn HM (2003) Does the cerebellum contribute to specific aspects of attention? Neuropsychologia 41(11):1452–1460
39. Heidemann RM, Ivanov D, Trampel R, Fasano F, Meyer H, Pfeuffer J, Turner R (2012) Isotropic submillimeter fMRI in the human brain at 7 T: combining reduced fieldâŁłofâŁłview imaging and partially parallel acquisitions. Magn Reson Med 68(5):1506–1516
40. Herculano-Houzel S (2010) Coordinated scaling of cortical and cerebellar numbers of neurons. Front Neuroanat. doi:10.3389/fnana.2010.00012
41. Johansen-Berg H, Behrens T, Robson MD, Drobnjak I, Rushworth M, Brady JM et al (2004) Changes in connectivity profiles define functionally distinct regions in human medial frontal cortex. Proc Natl Acad Sci U S A 101(36):13335–13340
42. Kachlik D, Baca V, Bozdechova I, Cech P, Musil V (2008) Anatomical terminology and nomenclature: past, present and highlights. Surg Radiol Anat 30(6):459–466
43. Kandel ER, Schwartz JH, Jessell T (2000) Principles of neural science, 4th edn. McGraw-Hill, New York, pp 1–1414
44. Kennedy DN, Lange N, Makris N, Bates J, Meyer J, Caviness VS (1998) Gyri of the human neocortex: an MRI-based analysis of volume and variance. Cerebral Cortex 8(4):372–384 (New York: 1991)
45. Keren NI, Lozar CT, Harris KC, Morgan PS, Eckert MA (2009) In vivo mapping of the human locus coeruleus. Neuroimage 47(4):1261–1267
46. Keuken MC, Bazin P-L, Crown L, Hootsmans J, Laufer A, Muller-Axt C, Sier R, van der Putten EJ, Schafer A, Turner R, Forstmann BU (2014) Quantifying inter-individual anatomical variability in the subcortex using 7T structural MRI. Neuroimage 94:40–46

47. Keuken MC, Bazin PL, Schafer A, Neumann J, Turner R, Forstmann BU (2013) Ultra-High 7 T MRI of structural age-related changes of the subthalamic nucleus. J Neurosci 33(11):4896–4900
48. Kitajima M, Korogi Y, Kakeda S, Moriya J, Ohnari N, Sato T et al (2008) Human subthalamic nucleus: evaluation with high-resolution MR imaging at 3.0 T. Neuroradiology 50(8):675–681
49. Knecht S, Dräger B, Deppe M, Bobe L, Lohmann H, Flöel A et al (2000) Handedness and hemispheric language dominance in healthy humans. Brain 123(12):2512–2518
50. Lehéricy S (2004) 3-D diffusion tensor axonal tracking shows distinct SMA and Pre-SMA projections to the human striatum. Cereb Cortex 14(12):1302–1309
51. Leuze CWU, Anwander A, Bazin PL, Dhital B, Stuber C, Reimann K et al (2012) Layer-specific intracortical connectivity revealed with diffusion MRI. Cerebral Cortex. doi:10.1093/cercor/bhs311
52. Makris N, Goldstein JM, Kennedy D, Hodge SM, Caviness VS, Faraone SV et al (2006) Decreased volume of left and total anterior insular lobule in schizophrenia. Schizophr Res 83(2–3):155–171
53. Mazziotta JC, Toga AW, Evans A, Fox P, Lancaster J (1995) A probabilistic atlas of the human brain: theory and rationale for its development the international consortium for brain mapping (ICBM). Neuroimage 2(2PA):89–101
54. Mazziotta J, Toga A, Evans A, Fox P, Lancaster J, Zilles K et al (2001) A probabilistic atlas and reference system for the human brain: international consortium for brain mapping (ICBM). Philos Trans R Soc B Biol Sci 356(1412):1293–1322
55. Mesulam M (2005) Imaging connectivity in the human cerebral cortex: the next frontier? Ann Neurol 57(1):5–7
56. Morel A, Magnin M, Jeanmonod D (1997) Multiarchitectonic and stereotactic atlas of the human thalamus. J Compar Neurol 387(4):588–630
57. Ngowyang G (1934) Die Cytoarchitektonik des menschlichen Stirnhirns I. Cytoarchitektonische Felderung der Regio granularis und Regio dysgranularis. Monogr Natl Res Inst Psychol Acad Sin (Shanghai) 7:1–68
58. Oishi K, Faria AV, van Zijl PC, Mori S (2010) MRI atlas of human white matter, 2nd ed. Academic, Waltham, pp 1–266
59. Ono M, Kubik S, Abernathey CD (1990) Atlas of the cerebral sulci. Thieme, New York, pp 1–232
60. Paluzzi A, Belli A, Bain P, Viva L (2007) Brain "imaging"in the Renaissance. J R Soc Med 100(12):540–543
61. Passingham RE, Stephan KE, Kötter R (2002) The anatomical basis of functional localization in the cortex. Nat Rev Neurosci 3(8):606–616
62. Penhune VB, Zatorre RJ, MacDonald JD, Evans AC (1996) Interhemispheric anatomical differences in human primary auditory cortex: probabilistic mapping and volume measurement from magnetic resonance scans. Cerebral Cortex 6(5):661–672 (New York: 1991)
63. Purves D, Augustine GJ, Fitzpatrick D, Hall WC, LaMantia AS, White LE (2012) Neuroscience, 5th edn. Sinauer Associates Inc, Massachusetts, pp 1–833
64. Rademacher J, Morosan P, Schormann T, Schleicher A, Werner C, Freund HJ, Zilles K (2001) Probabilistic mapping and volume measurement of human primary auditory cortex. Neuroimage 13(4):669–683
65. Rajkowska G, Goldman-Rakic PS (1995a) Cytoarchitectonic definition of prefrontal areas in the normal human cortex: I. Remapping of areas 9 and 46 using quantitative criteria. Cerebral Cortex 5(4):307–322 (New York: 1991)
66. Rajkowska G, Goldman-Rakic PS (1995b) Cytoarchitectonic definition of prefrontal areas in the normal human cortex: II. Variability in locations of areas 9 and 46 and relationship to the Talairach coordinate system. Cerebral Cortex 5(4):323–337 (New York: 1991)
67. Reil JC (1809) Die Sylvische Grube oder das Thal, das gestreifte große Hirnganglium, dessen Kapsel und die Seitentheile des großen Gehirns. Arch Physiol 9:195–208
68. Reil JC (1812) Die vördere Commissur im großen Gehirn. Arch Physiol 11:89–100

69. Sarkisov SA, Filimonoff IN, Kononowa EP, Preobraschenskaja IS, Kukuew EA (1955) Atlas of the cytoarchitectonics of the human cerebral cortex. Medgiz, Moskow
70. Schmahmann JD, Pandya DN (2009) Fiber pathways of the brain. Oxford University Press, Oxford, pp 1–654
71. Smith G (1907) A new topographical survey of the human cerebral cortex, being an account of the distribution of the anatomically distinct cortical areas and their relationship to the cerebral sulci. J Anat Physiol 41(Pt 4):237
72. Standring S (2008) Gray's anatomy, 40 edn. Elsevier, Amsterdam, pp 1–1576
73. Stoodley CJ, Schmahmann JD (2009) Functional topography in the human cerebellum: a meta-analysis of neuroimaging studies. Neuroimage 44(2):489–501
74. Strasburger EH (1937) Die myeloarchitektonische Gliederung des Stirnhirns beim Menschen und Schimpansen. Journal fur psychologie und neurologie 47(6):565–606
75. Swaab DF (2003) The human hypothalamus: basic and clinical aspects. Part 1: nuclei of the human hypothalamus. In: Aminoff MJ, Boller F, Swaab DF (eds) Handbook of clinical neurology, vol 79. Elsevier, Amsterdam, p 9
76. Swanson LW (2000) What is the brain? Trends Neurosci 23(11):519–527
77. Swanson LW, Bota M (2010) Foundational model of structural connectivity in the nervous system with a schema for wiring diagrams, connectome, and basic plan architecture. Proc Natl Acad Sci U S A 107(48):20610–20617
78. Talairach J, Tournoux P (1988) Co-planar stereotaxic atlas of the human brain. Thieme, New York, pp 1–122
79. Thompson PM, Schwartz C, Lin RT, Khan AA, Toga AW (1996) Three-dimensional statistical analysis of sulcal variability in the human brain. J Neurosci 16(13):4261–4274
80. Thompson PM, Cannon TD, Narr KL, Van Erp T, Poutanen V-P, Huttunen M et al (2001) Genetic influences on brain structure. Nat Neurosci 4(12):1253–1258
81. Turner R (2012) Neuroscientific applications of high-field MRI in humans. In: Hennig J, Speck O (eds) High-Field MR imaging. Springer, Berlin
82. Vesalius A (1543) De humani corporis fabrica libri septem. School of medicine, Padua
83. Vogt O (1910) Die myeloarchitektonische Felderung des menschlichen Stirnhirns. J Psychol Neurol 15(4/5):221–232
84. Vogt C, Vogt O (1926) Die vergleichend-architektonische und die vergleichend-reizphysiologische Felderung der Großhirnrinde unter besonderer Berücksichtigung der menschlichen. Naturwissenschaften 14(50):1190–1194
85. Waehnert MD, Dinse J, Weiss M, Streicher MN, Waehnert P, Geyer S et al (2013) Anatomically motivated modeling of cortical laminae. NeuroImage. doi:10.1016/j.neuroimage.2013.03.078

Chapter 5
An Introduction to fMRI

F. Gregory Ashby

Abstract Functional magnetic resonance imaging (fMRI) provides an opportunity to indirectly observe neural activity noninvasively in the human brain as it changes in near real time. Most fMRI experiments measure the blood oxygen-level dependent (BOLD) signal, which rises to a peak several seconds after a brain area becomes active. Several experimental designs are common in fMRI research. Block designs alternate periods in which subjects perform some task with periods of rest, whereas event-related designs present the subject with a set of discrete trials. After the fMRI experiment is complete, pre-processing analyses prepare the data for task-related analyses. The most popular task-related analysis uses the General Linear Model to correlate a predicted BOLD response with the observed activity in each brain region. Regions where this correlation is high are identified as task related. Connectivity analysis then tries to identify active regions that belong to the same functional network. In contrast, multivariate methods, such as independent component analysis and multi-voxel pattern analysis identify networks of event-related regions, rather than single regions, so they simultaneously address questions of functional connectivity.

5.1 Introduction

Functional magnetic resonance imaging (fMRI) provides researchers an opportunity to observe neural activity noninvasively in the human brain, albeit indirectly, as it changes in near real time. This exciting technology has revolutionized the scientific study of the mind. For example, largely because of fMRI, there are now emerging new fields of Social Neuroscience, Developmental Neuroscience, Neuroeconomics, and even Neuromarketing.

This chapter provides a brief overview of fMRI and fMRI data analysis. This is a complex topic that includes many difficult subtopics, such as (1) MR physics, (2) a description of the complex machinery and equipment one finds in a typical

F. G. Ashby (✉)
Department of Psychological & Brain Sciences, University of California,
Santa Barbara, CA 93106, USA
Tel.: 805-893-7909
e-mail: ashby@psych.ucsb.edu

B. U. Forstmann, E.-J. Wagenmakers (eds.), *An Introduction to Model-Based Cognitive Neuroscience,* DOI 10.1007/978-1-4939-2236-9_5

brain-imaging center, (3) how to run this equipment effectively (e.g., set the many parameters that control the scanner; spot and avoid artifacts that can corrupt the data), (4) experimental design, and (5) fMRI data analysis. Obviously, covering all this material in depth is far beyond the scope of any single chapter, or even any single book. The reader interested in learning more about these topics is urged to consult any of the books listed at the end of this chapter under "Further Reading."

5.2 What Can Be Learned from fMRI?

Currently, the typical fMRI experiment records a sluggish, indirect measure of neural activity with a temporal resolution of 1–3 s and a spatial resolution of 25–30 mm^3. Nevertheless, as the thousands of fMRI publications attest, this highly imperfect technology has dramatically influenced the study of mind and brain. Because of its poor temporal resolution, fMRI is not appropriate for resolving small timing differences between different cognitive stages or processes. And although the spatial resolution is typically good enough to localize brain activity at the level of major brain structures, it is not good enough to localize activity, for example, at the level of the cortical column. Partly for these reasons, the use of fMRI is not without controversy. Although the mind sciences have been generally enthusiastic about fMRI, a smaller group of scientists remain skeptical. For example, fMRI has been labeled by some as "the new phrenology" [1].

Because of this controversy, before examining fMRI in more detail, it is worth considering what this relatively new technology has to contribute to the mind sciences. Because of its limited temporal and spatial resolution, fMRI is most appropriate for answering questions about gross neural architecture, rather than about neural process. But it can be very effective at addressing such questions. For example, consider the fMRI results shown in Fig. 5.1. The four panels show areas of significant activation within prefrontal cortex on four different days of training as subjects practiced a difficult perceptual categorization task [2]. Note that as automaticity develops in this task, prefrontal activation reduces significantly. In fact, by the 20th practice session (i.e., after approximately 11,000 trials of practice) there is no evidence of any task-related activity in prefrontal cortex. Therefore, these fMRI data show clearly that the neural architecture mediating this categorization behavior changes qualitatively with practice. This same question could be addressed without fMRI, and it seems possible that a similar conclusion might be reached after many clever behavioral experiments. But Fig. 5.1 paints a clear and compelling picture that one rarely sees with purely behavioral approaches.

FMRI can be used in a similar way to test purely cognitive theories that make no neuroscience assumptions. For example, suppose some cognitive theory predicts that the same perceptual and cognitive processes mediate performance in two different tasks. Then this theory should predict similar patterns of activation in an fMRI study of the two tasks, even if the theory makes no predictions about what those activation patterns should look like. If qualitatively different activation patterns are found in

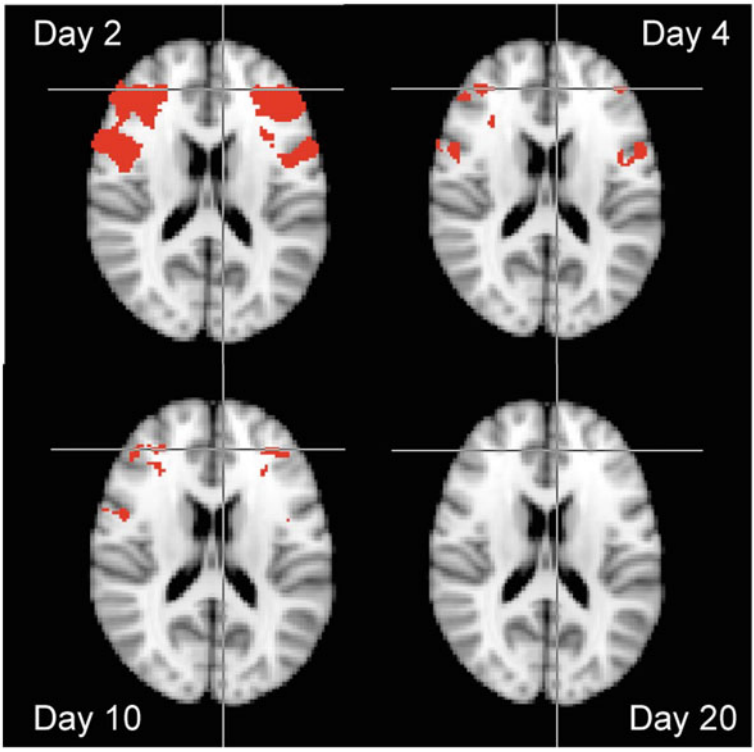


Fig. 5.1 Significant BOLD activation in prefrontal cortex on four different days of training on the same categorization task [8]

the tasks, then the theory probably needs some serious re-thinking. Applications like this are the primary reason that fMRI is popular even with cognitive scientists who have no fundamental interest in neuroscience.

As a final illustration of the benefits of fMRI, consider the following example. Unstructured categories are categories in which the stimuli are assigned to each category randomly. As a result, there is no similarity-based or logical rule for determining category membership. An example might be the category of numbers that have personal meaning to you (e.g., social security number, phone number, etc.). I had hypothesized in print that unstructured categories must be learned by explicit memorization, even when feedback-based training is provided [3]. But then Seger and her colleagues published several fMRI papers showing that feedback-based unstructured category learning elicits elevated task-related activation in the striatum, not the hippocampus [4–6]. These results suggested that unstructured category learning might be mediated by procedural memory, not declarative memory. Because of these papers, my colleagues and I decided to look for behavioral evidence that unstructured category learning is mediated by procedural memory. In fact, we found that switching the locations of the response buttons interfered with the expression of

unstructured category learning, but not with a (rule-based) version of this task that used similar stimuli and was known to depend on declarative memory. This sensitivity to response location is a hallmark of procedural memory, so our behavioral results were in agreement with the fMRI results. The important point here is that this behavioral experiment would not even have been run if the fMRI experiments had not identified a neural network that previously had been associated with procedural memory.

In summary, fMRI is a powerful method for studying neural and cognitive architecture, but it is not as effective at addressing questions about process. For this reason, it is not some magic method that will supplant all others in the scientific study of the mind. But it does provide an important new tool for mind scientists. When used in a converging operations approach that includes more traditional behavioral methodologies, it has the potential to dramatically improve our understanding of the human mind.

5.3 MR Physics and BOLD Imaging

The MR scanner uses superconducting electromagnets to produce a static, uniform magnetic field of high strength. Ten years ago, the standard field strength used in fMRI research was 1.5 Tesla (T), whereas the standard today is 3 T. Even so, a number of research centers have scanners considerably stronger than this (e.g., above 10 T). Some of these are used with human subjects, but many are only used for non-human animal research.

The static field, by itself, does not produce an MR signal. An MR signal requires radiofrequency coils that generate magnetic pulses. Turning a pulse on changes the magnetization alignment of protons (typically within water molecules) within the magnetic field. When the pulse is turned off, the protons relax to their original equilibrium alignment, which releases energy detected by the coils as the raw MR signal. Spatial resolution is provided by additional magnetic fields known as gradients. The strength of each gradient changes linearly along a single spatial dimension. Thus, three mutually orthogonal gradients are used to localize a signal in three spatial dimensions. The software that controls all these magnetic fields is typically called the pulse sequence.

The pulse sequence is run on the main computer that controls the scanner. In most fMRI experiments, a second computer creates the stimuli that are presented to the subject and records the subject's behavioral responses. This second computer is synchronized with the first, so that the onset of each stimulus presentation occurs at a precisely controlled moment during image acquisition. Visual stimuli are most often presented by directing a computer-controlled projector at a mirror directly above the subject's face, and responses are collected on some device held in the subject's hands (e.g., that has buttons or a joystick).

Two general types of pulse sequences are common, depending on whether the goal is structural or functional imaging. The goal of structural MR is usually to measure

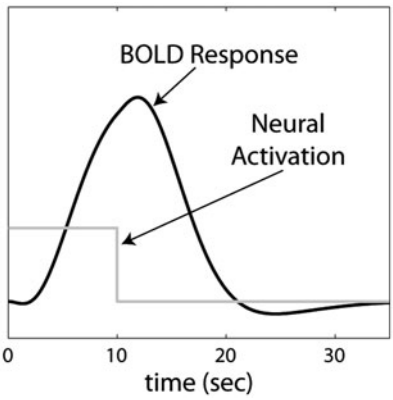


Fig. 5.2 A hypothetical BOLD response (*black curve*) to a constant 10 s neural activation (*gray curve*)

the density of water molecules, which differs, for example in bone, gray matter, cerebrospinal fluid, and tumors. The vast majority of functional MR (fMRI) experiments measure the blood oxygen-level dependent (BOLD) signal. The physics of this process is complex and far beyond the scope of this chapter. For our purposes, it suffices to know that the BOLD signal is a measure of the ratio of oxygenated to deoxygenated hemoglobin. Hemoglobin is a molecule in the blood that carries oxygen from the lungs to all parts of the body. It has sites to bind up to four oxygen molecules. A key discovery that led eventually to BOLD fMRI was that hemoglobin molecules fully loaded with oxygen have different magnetic properties than hemoglobin molecules with empty binding sites [7].

The theory, which is not yet fully worked out, is that active brain areas consume more oxygen than inactive areas. When neural activity increases in an area, metabolic demands rise and, as a result, the vascular system rushes oxygenated hemoglobin into the area. An idealized example of this process is shown in Fig. 5.2. The rush of oxygenated hemoglobin into the area causes the ratio of oxygenated to deoxygenated hemoglobin (i.e., the BOLD signal) to rise quickly. As it happens, the vascular system over compensates, in the sense that the BOLD signal actually rises well above baseline to a peak at around 6 s after the end of the neural activity that elicited these responses. Following this peak, the BOLD signal gradually decays back to baseline over a period of 20–25 s.

5.4 The Scanning Session

An experimental session that collects fMRI data also commonly includes a variety of other types of scans. At least four different types of scans are commonly acquired. Typically, the first scan completed in each session is the localizer. This is a quick structural scan (1–2 min) of low spatial resolution and is used only to locate the subject's brain in 3-dimensional space. This knowledge is needed to optimize the

location of the slices that will be taken through the brain in the high-resolution structural scan and in the functional scans that follow.

The ordering of the other scans that are commonly done is not critical. Frequently, however, the second type of scan completed is the high-resolution structural scan. Depending on the resolution of this scan and on the exact nature of the pulse sequence that is used to control the scanner during acquisition, it may take 8–10 min to collect these data. The structural scan plays a key role in the analysis of the functional data. Because speed is a high priority in fMRI (i.e., to maximize temporal resolution), spatial resolution is sacrificed when collecting functional data. The high-resolution structural scan can compensate somewhat for this loss of spatial information. This is done during preprocessing when the functional data are aligned with the structural image. After this mapping is complete, the spatial coordinates of activation observed during fMRI can be determined by examining the aligned coordinates in the structural image.

The third step is often to collect the functional data. This can be done in one long run that might take 20–30 min to complete, or it can be broken down into 2 or 3 shorter runs, with brief rests in between. There are many parameter choices to make here, but two are especially important for the subsequent fMRI data analysis. One choice is the time between successive whole brain scans, which is called the repetition time and abbreviated as the TR. If the whole brain is scanned, typical TRs range from 2–3 s, but TRs as low as 1 s are possible on many machines, especially if some parts of the brain are excluded from the scanning.

Another important choice is voxel size, which determines the spatial resolution of the functional data. When a subject lies in the scanner, his or her brain occupies a certain volume. If we assign a coordinate system to the bore of the magnet, then we could identify any point in the subject's brain by a set of three coordinate values (*x, y, z*). By convention, the *z* direction runs down the length of the bore (from the feet to the head), and the *x* and *y* directions reference the plane that is created by taking a cut perpendicular to the *z* axis. The brain, of course, is a continuous medium, in the sense that neurons exist at (almost) every set of coordinate values inside the brain. FMRI data, however, are discrete. The analog-to-digital conversion is performed by dividing the brain into a set of cubes (or more accurately, rectangular right prisms). These cubes are called voxels because they are three-dimensional analogues of pixels – that is, they could be considered as volume pixels.

A typical voxel size in functional imaging might be 3 mm $\times$ 3 mm $\times$ 3.5 mm. In this case, in a typical human brain, 33 separate slices might be acquired each containing a 64 $\times$ 64 array of voxels for a whole brain total of 135,168 voxels. In each fMRI run, a BOLD response is recorded every TR seconds in each voxel. Thus, for example, in a 30 min run with a TR of 2 s, 135,168 BOLD responses could be recorded 900 separate times (i.e., 30 times per minute $\times$ 30 min), for a total of 121,651,200 BOLD values. This is an immense amount of data, and its sheer volume greatly contributes to the difficulties in data analysis.

Many studies stop when the functional data acquisition is complete, but some other types of scans are also common. A fourth common type of scan is the field map. The ideal scanner has a completely uniform magnetic field across its entire

bore. Even if this were true, placing a human subject inside the bore distorts this field to some extent. After the subject is inside the scanner, all inhomogeneities in the magnetic field are corrected via a process known as shimming. If shimming is successful, the magnetic field will be uniform at the start of scanning. Sometimes, however, especially in less reliable machines, distortions in the magnetic field will appear in the middle of the session. The field map, which takes only a minute or two to collect, measures the homogeneity of the magnetic field at the moment when the map is created. Thus, the field map can be used during later data analysis to correct for possible nonlinear distortions in the strength of the magnetic field that develop during the course of the scanning session.

5.5 Experimental Design

Almost all fMRI experiments use a block design, an event-related design, or a free-running design. In a block design, the functional run consists of a series of blocks, each of which may last for somewhere between 30 s to a couple of minutes. Within each block, subjects are instructed to perform the same cognitive, perceptual, or motor task continuously from the beginning of the block until the end. In almost all block-design experiments, subjects will simply rest on some blocks. For example, a researcher interested in studying the neural network that mediates rhythmic finger tapping might use a block design in which blocks where the subject is resting alternate with blocks in which the subject taps his or her finger according to some certain rhythm.

Event-related designs are run more like standard psychological experiments, in the sense that the functional run is broken down into a set of discrete trials. Usually each trial is one of several types, and each type is repeated at least 20 times over the course of the experiment. As in a standard experiment, however, the presentation order of the trial types within each run is often random. When analyzing data from an event-related design, it is critical to know exactly when the presentation of each stimulus occurred, relative to TR onset. A common practice is to synchronize stimulus presentation with TR onset. The first event-related designs included long rests between each pair of successive trials. In these slow event-related designs, rests of 30 s are typical. These are included so that the BOLD response in brain regions that participate in event processing can decay back to baseline before the presentation of the next stimulus. This makes statistical sense, but it is expensive since it greatly reduces the number of trials a subject can complete in any given functional run. Another problem is that because subjects have so much time with nothing to do, they might think about something during these long rests, and any such uncontrolled cognition would generate an unwanted BOLD response that might contaminate the stimulus-induced BOLD response.

Most current event-related designs use much shorter delays. These rapid event-related designs became possible because statistical methods were developed for dealing with the overlapping BOLD responses that will occur anytime the BOLD

response in a brain region has not decayed to baseline by the time another stimulus is presented. It is important to realize however, that even in rapid event-related designs the delay between trials is still significantly longer than in standard laboratory experiments. For example, a typical rapid event-related design might use random delays between successive trials that cover a range between 2 and 16 s. There are several reasons for this. First, because of the need to synchronize stimulus presentation with the TR, it is often necessary to delay stimulus presentation until the onset of the next TR. Second, in order to get unique estimates of the parameters of the standard statistical models that are used to analyze fMRI data, delays of random duration must be used. The process of adding such random delays between events is called *jittering*.

Finally, in free-running designs, events are presented to the subject continuously in time and typically discrete events are impossible to define. For example, subjects might watch a movie in the scanner, or simply lay there passively. The activities that subjects perform in free-running designs are often more natural than is possible with more structured designs, but this increased freedom comes at a cost because the data that result are more challenging to analyze than the data collected from block or event-related designs.

5.6 Data Analyses

A number of features of fMRI data greatly complicate its analysis. First, as mentioned above, a typical scanning session generates a huge amount of data. Second, fMRI data are characterized by substantial spatial and temporal correlations. For example, the sluggish nature of the BOLD response means that if the BOLD response in some voxel is greater than average on one TR then it is also likely to be greater than average on the ensuing TR. Similarly, because brain tissue in neighboring voxels will be supplied by a similar vasculature, a large response in one voxel increases the likelihood that a large response will also be observed at neighboring voxels. A third significant challenge to fMRI data analysis is the noisy nature of fMRI data. Typically the signal that the data analysis techniques are trying to find is less than 2 or 3 % of the total BOLD response.

The analysis of fMRI BOLD data is broken down into two general stages—preprocessing and task-related analysis. Preprocessing includes a number of steps that are required to prepare the data for task-related analysis. This includes, for example, aligning the functional and structural scans, correcting for any possible head movements that might have occurred during the functional run, and various types of smoothing (to reduce noise). Typically, the same preprocessing steps are always completed, regardless of the particular research questions that the study was designed to address. In contrast, the task-related analyses include all analyses that are directed at these questions.

A wide variety of software packages are available for fMRI data analysis. Many of these are free, and they each have their own advantages and disadvantages. The most widely used package is SPM (Statistical Parametric Mapping), which is written

and maintained by the Wellcome Trust Centre for Neuroimaging at the University College London. SPM is freely available at http://www.fil.ion.ucl.ac.uk/spm/. SPM is a collection of Matlab functions and routines with some externally compiled C code that is included to increase processing speed. A thorough description of the statistical foundations of SPM was provided by Friston, Ashburner, Kiebel, Nichols, and Penny [8].

Another widely used fMRI data analysis software package is called FSL, which is an acronym for the FMRIB Software Library. FSL is produced and maintained by the FMRIB Analysis Group at the University of Oxford in England. FSL is also freely available and can be downloaded at http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FSL. Descriptions of the statistical foundations of the FSL routines were provided by Smith et al. [9] and by Woolrich et al. [10].

5.6.1 Modeling the BOLD Response

The goal of almost all fMRI experiments is to learn something about neural activity. Unfortunately however, the BOLD response measured in most fMRI experiments provides only an indirect measure of neural activation [11, 12]. Although it is commonly assumed that the BOLD signal increases with neural activation, it is known that the BOLD response is much more sluggish than the neural activation that is presumed to drive it. As a result, for example, the peak of the BOLD signal lags considerably behind the peak neural activation (e.g., see Fig. 5.2).

Logothetis and colleagues have presented evidence that the BOLD response is more closely related to local field potentials than to the spiking output of individual neurons [13, 14]. Local field potentials integrate the field potentials produced by small populations of cells over a sub-millimeter range, and they vary continuously over time. Most applications of fMRI make no attempt to model neural activation at such a detailed biophysical level. Rather, neural activation is typically treated as a rather abstract latent (i.e., unobservable) variable. It is assumed to increase when a brain region is active and to decrease during periods of inactivity. As with any latent variable, however, to make inferences about neural activation from observable BOLD responses requires a model of how these two variables are related.

Almost all current applications of fMRI assume that the transformation from neural activation to BOLD response can be modeled as a linear, time-invariant system. Although it is becoming increasingly clear that the transformation is, in fact, nonlinear (e.g., [15–17]), it also appears that these departures from linearity are not severe so long as events are well separated in time (e.g., at least a few seconds apart) and brief exposure durations are avoided [17]. These two conditions are commonly met in fMRI studies of high-level cognition.

In the linear systems approach, one can conceive of the vascular system that responds to a sudden oxygen debt as a black box. The input is neural activation and the output is the BOLD response. Suppose we present a stimulus event E_i to a subject at time 0. Let $N_i(t)$ denote the neural activation induced by this event at

time t and let $B_i(t)$ denote the corresponding BOLD response. Then from the systems theory perspective, the box represents the set of all mathematical transformations that convert the neural activation $N_i(t)$ into the BOLD response $B_i(t)$. For convenience, we will express this mathematical relationship as

$$B_i(t) = f[N_i(t)]$$

where the operator f symbolizes the workings of the black box.

A system of this type is said to be linear and time-invariant if and only if it satisfies the superposition principle, which is stated as follows:

If $f[N_1(t)] = B_1(t)$ and $f[N_2(t)] = B_2(t)$, then it must be true that

$f[a_1 N_1(t) + a_2 N_2(t)] = a_1 B_1(t) + a_2 B_2(t)$, for any constants a_1 and a_2.

In other words, if we know what the BOLD response is to neural activation $N_1(t)$ and to neural activation $N_2(t)$, then we can determine exactly what the BOLD response will be to any weighted sum of these two neural activations by computing the same weighted sum of the component BOLD responses.

If the superposition principle holds then there is a straightforward way to determine the BOLD response to *any* neural activation from the results of one simple experiment. All we need to do is to measure the BOLD response that occurs when the neural activation is an impulse—that is, when it instantly increases from zero to some large value then instantly drops back to zero. Denote the BOLD response in this idealized experiment by $h(t)$. In linear systems theory the function $h(t)$ is called the impulse response function because it describes the response of the system to an impulse. In the fMRI literature, however, $h(t)$ is known as the hemodynamic response function, often abbreviated as the hrf. Note that "hemodynamic response function" is not a synonym for "BOLD response". Rather the hrf is the hypothetical BOLD response to an idealized impulse of neural activation.

If the relationship between neural activation and the BOLD response satisfies superposition, then once we know the hrf, the BOLD response to any neural activation $N(t)$, no matter how complex, can be computed exactly from the so-called convolution integral:

$$B(t) = \int_0^t N(\tau)h(t-\tau)d\tau \tag{5.1}$$

The convolution integral massively simplifies the analysis of fMRI data, and as a result it forms the basis for the most popular methods of fMRI data analysis.

Given that the hrf plays such a critical role in analyzing fMRI data, the natural next question to ask is: how can we determine numerical values of the hrf? The most obvious method for determining the hrf, which is suggested by the name "impulse response function", is simply to input an impulse to the system and record the output. If the system is linear and time-invariant, then the output will exactly equal $h(t)$. With traditional fMRI experiments, of course, we cannot directly input a neural activation, so using this method to estimate the hrf is highly problematic. Even so, this method has been used to estimate the hrf in primary visual cortex (e.g., [18, 19]).

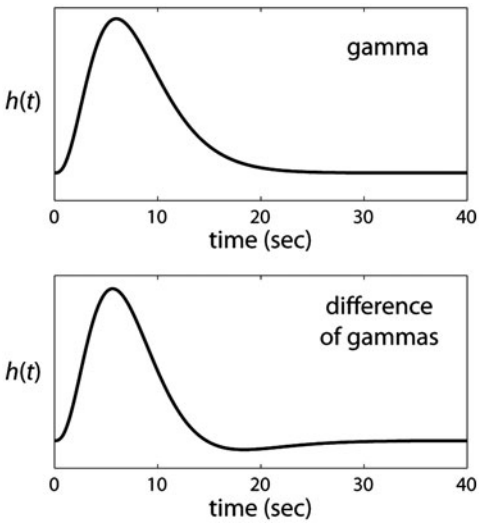


Fig. 5.3 Two popular models of the hrf

A much more popular method is to select a specific mathematical function for the hrf based on our knowledge of what we think this function should look like. For example, we know the hrf should peak at roughly 6 s and then slowly decay back to baseline. So we could select a mathematical function with these properties and then just assume that this is a good model of the hrf. In fact, this is, by far, the most popular method for determining the hrf in fMRI data analysis. The most popular choices are a gamma function or the difference of two gamma functions. Examples of both of these models are shown in Fig. 5.3.

5.6.1.1 Preprocessing

The most common goal of fMRI research is to identify brain areas activated by the task under study. The data that come directly out of the scanner, however, are poorly suited to this goal. The preprocessing of fMRI data includes all transformations that are needed to prepare the data for the more interesting task-related analyses. Preprocessing steps typically are the same for all experiments, so any analyses that do not depend on the specific hypotheses that the experiment was designed to test are typically called preprocessing.

The variability in raw fMRI data is so great that it easily can swamp out the small changes in the BOLD response induced by most cognitive tasks. Some of this variability is unavoidable in the sense that it is due to factors that we cannot control or even measure (e.g., thermal and system noise). But other sources of variability are systematic. For example, when a subject moves his or her head, the BOLD response sampled from each spatial position within the scanner suddenly changes in a predictable manner. The analyses done during preprocessing remove as many of these systematic non-task-related sources of variability as possible.

Typically the first preprocessing step is slice-time correction. Almost all fMRI data are collected in slices. If the TR is 2.5 s, then the time between the acquisition of the first and last slice will be almost this long. Slice-time correction corrects for these differences in the time when the slices are acquired.

The second step is to correct for variability due to head movement. Arguably, this is probably the most important preprocessing step. Even small, almost imperceptible head movements can badly corrupt fMRI data. Huettel et al. [20] give an example where a head movement of 5 mm increases activation values in a voxel by a factor of 5. When a subject moves his or her head, brain regions will move to new spatial locations within the scanner, and as a result, activation in those regions will be recorded in different voxels than they were before the movement occurred. Mathematical methods for correcting for head movements depend heavily on the assumption that when a subject moves his or her head, the brain does not change shape or size and therefore can be treated as a rigid body. Head movement correction then becomes a problem of rigid body registration (e.g., [21]). The BOLD responses from one TR are taken as the standard and then rigid body movements are performed separately on the data from every other TR until each of these data sets agrees as closely as possible with the data from the standard.

The third step, called coregistration, is to align the structural and functional data. This is critical because the spatial resolution of the functional data is poor. For example, with functional data a voxel size of $3 \times 3 \times 3.5$ mm is common. With structural images, however, the voxel size might be $.86 \times .86 \times .89$ mm, which is an improvement in resolution by a factor of almost 50.

The fourth step, normalization, warps the subject's structural image to a standard brain atlas. There are huge individual differences in the sizes and shapes of individual brains, and these differences extend to virtually every identifiable brain region. These differences make it difficult to assign a task-related activation observed in some cluster of voxels to a specific neuroanatomical brain structure. A researcher particularly skilled in neuroanatomy could coregister the functional activation onto the structural image and then look for landmarks in the structural scan that would allow the neuroanatomical locus of the cluster to be identified. An alternative is to register the structural scan of each subject separately to some standard brain where the coordinates of all major brain structures have already been identified and published in an atlas. Then we could determine the coordinates of a significant cluster within this standard brain, look these coordinates up in the atlas, and thereby determine which brain region the cluster is in. The process of registering a structural scan to the structural scan from some standard brain is called normalization [22].

Among the earliest and still most widely used brain atlases is the Talairach atlas [23], which is based entirely on the detailed dissection of one hemisphere of the brain of a 60-year old French woman. The atlas is essentially a look-up table containing major brain areas and their anatomical (x, y, z) coordinates. For many years, the Talairach atlas was almost universally used in neuroimaging, primarily because of the lack of any reasonable alternatives. But there has always been widespread dissatisfaction with this atlas because it is based on one hemisphere of a single, rather

unrepresentative brain. More recently, an atlas produced by the Montreal Neurological Institute (MNI) has become popular. The MNI atlas was created by averaging the results of high resolution structural scans that were taken from 152 different brains. The coordinate system was constructed to match the Talairach system, in the sense that it uses the same axes and origin. Whichever atlas is used, it is important to note that the registration problem in normalization is considerably more complex than in head motion correction or coregistration. This is because normalization requires more than rigid body registration. Not only will there be rigid body differences between the standard brain and the brain of typical subjects, but there will also be size and shape differences. Size differences can be accommodated via a linear transformation, but a nonlinear transformation is almost always required to alter the shape of a subject's brain to match either the Talairach or MNI standards.

Step five spatially smoothes the data with the goal of reducing nonsystematic high frequency spatial noise. In this step, the BOLD value in each voxel is replaced by a weighted average of the BOLD responses in neighboring voxels. The weight is greatest at the voxel being smoothed and decreases with distance. There are a number of advantages to spatially smoothing fMRI data. Most of these are due to the effects of the smoothing process on noise in the data. First, because smoothing is essentially an averaging operation, it makes the distribution of the BOLD responses more normal (i.e., because of the central limit theorem). Because the statistical models that dominate fMRI data analysis assume normally distributed noise, smoothing therefore transforms the data in a way that makes it more likely to satisfy the assumptions of our statistical models. A second benefit is that smoothing is required by a number of popular methods for solving the multiple comparisons problem (i.e., those that depend on Gaussian random field theory). A third benefit of smoothing, which is the most important of all, is that it can reduce noise and therefore increase signal-to-noise ratio.

Finally, in step six, temporal filtering is done primarily to reduce the effects of slow fluctuations in the local magnetic field properties of the scanner.

5.6.1.2 Task-Related Data Analyses

After pre-processing is complete, the next step is to try to identify brain regions that were activated by the task under study. The most popular approach to this problem is a correlation-based technique that is the foundation of most fMRI software packages [24, 25]. The idea is to first predict as accurately as possible what the BOLD response should look like in task-sensitive voxels. Next, the observed BOLD response in each voxel is correlated with this predicted signal. Voxels where this correlation is high are then identified as task related.

The first step in predicting the BOLD response to each stimulus event is to make an assumption about how long the neural activation will last in brain regions that process this event. A common assumption is that the neural activation induced by the event onset will persist for as long as the stimulus is visible to the subject. Another possibility is that the neural activation persists until the subject responds (so the

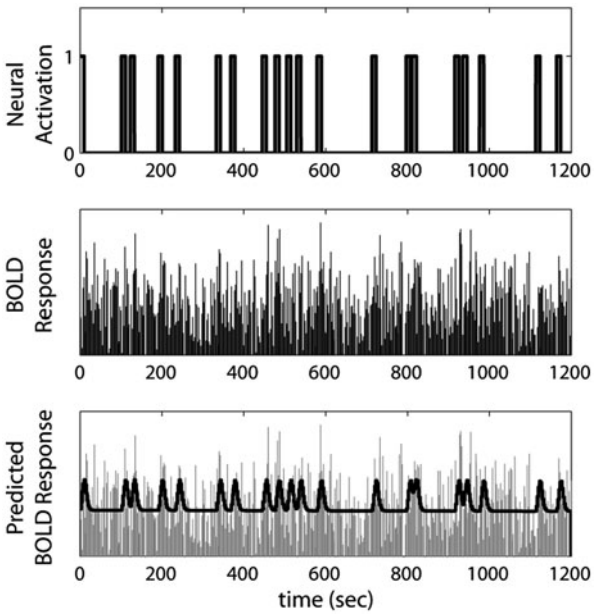


Fig. 5.4 A hypothetical example of the standard correlation-based analysis of fMRI data. The *top* panel shows the boxcar function that models the presumed neural activation elicited by the presentation of 20 separate stimuli. The *middle* panel depicts the hypothetical BOLD response in this experiment in a voxel with task-related activity. The *bottom* panel shows the best-fitting predicted BOLD response that is generated by convolving an hrf with the boxcar function shown in the *top* panel and then adding a constant baseline activation level

duration of neural activation equals the subject's response time). The second step is to model all presumed neural activations via a boxcar function. This is simply a function that persists for the duration of fMRI data acquisition and equals 1 when neural activation is assumed to be present and 0 when neural activation is absent. The top panel of Fig. 5.4 shows a hypothetical example of a boxcar function that describes the presumed neural response to the presentation of 20 separate stimuli. The stimulus presentations were spaced irregularly in time (i.e., jittered) to improve the statistical properties of the analysis. The middle panel of Fig. 5.4 shows the hypothetical BOLD response recorded in this experiment from one task-related voxel. Note that from visual inspection alone, it is not at all obvious that this is a task-related voxel.

The correlation method assumes linearity, so the third step in the analysis is to choose a model of the hrf. As mentioned earlier, the most popular choice is to select a specific mathematical function for the hrf that has no free parameters (e.g., either function shown in Fig. 5.3). Step four is to compute the predicted BOLD response by convolving the neural boxcar function with the hrf (using Eq. 5.1). Because there are no free parameters in either the boxcar function or the hrf, this integral can be evaluated numerically. In other words, for every TR in the experiment, a numerical value of the predicted BOLD response can be computed from Eq. 5.1. The bottom panel in Fig. 5.4 shows the predicted BOLD response that results when the boxcar function in the top panel is convolved with a gamma function hrf (and an estimated baseline activation is added).

The final step is to correlate these predicted BOLD values with the observed BOLD response in every voxel. Voxels where this correlation is high are presumed

to show task-related activity. The correlation is typically done within the context of the familiar General Linear Model (GLM) that is the basis of both multiple regression and analysis-of-variance. The outcome of this analysis in each voxel is the value of a statistic—most often a z or t value—that tests the null hypothesis that activation in that voxel is not correlated with the predicted BOLD response, or in other words, that activity in the voxel is not task related. Extreme values of the statistic are therefore evidence for task-related activity. In Fig. 5.4, the t statistic that results from this correlation has a numerical value of 7.78.

The correlation method applies to data from a single voxel at a time. Thus, if an experiment collects data from the whole brain, this analysis could easily be repeated more than 100,000 times to analyze all of the data collected in the experiment. The result of all these analyses is a value of the test statistic in every voxel that was analyzed. The resulting collection of statistics is often called a statistical parametric map, which motivated the name of the well-known fMRI data analysis software package, SPM.

A more recent variant of this correlation-based approach, called model-based fMRI, uses an independent computational model of the behavior under study to improve and refine the predicted BOLD signal [26]. In typical applications, the model is first fit to the behavioral data collected during the functional run separately for each subject. Next, parameter estimates from the model fits are used to build a model of the neural activation that is unique for every subject. From here, the analysis proceeds exactly as in the standard correlation-based method—that is, the predicted neural activation is convolved with an hrf to generate a predicted BOLD signal and then the GLM is used to generate a statistical parametric map. Model-based fMRI can be used to account for individual differences in fMRI data, but if the computational model is good, it can also be used to identify brain regions that respond selectively to components or sub-processes of the task. In particular, if the model has different parameters that describe different perceptual or cognitive processes that are presumed to mediate the behavior under study, then different regressors can be created that make specific predictions about each of these processes. For example, O'Doherty et al. [27] used this approach to identify separate brain regions associated with the actor versus the critic in actor-critic models of reinforcement learning.

Once a statistical map is constructed, the next problem is to determine which voxels show task-related activity. Of course if we only ask this question about a single voxel, then the answer is taught in every introductory statistics course. We simply decide what type 1 error rate we are willing to accept, find the threshold value of the statistic (e.g., the z or t value) that yields this error rate, and then decide that the voxel shows task-related activity if the value of this statistic exceeds this threshold. However, if the type 1 error rate equals 0.05 for each test, then with 100,000 independent tests, we would expect 5000 false positives if none of these voxels were task sensitive. This is clearly unacceptable. As a result, the criterion for significance must somehow be adjusted on each test to reduce the total number of false positives to some acceptable value. In statistics, this is called the multiple comparisons problem.

If the tests are all statistically independent then it is well known that the exact solution to this problem is to apply the Sidak or Bonferroni corrections. For example, if we want to limit the probability of a type 1 error to 0.05 in an overall collection of 100,000 z-tests (i.e., so that 95 % of the time there are no false positives in the 100,000 tests), then the Bonferroni correction sets the critical value on each test to approximately 0.0000005, which translates to a z-threshold for determining significance of 4.89. In the Fig. 5.4 example, the t value (i.e., 7.78) is so large that it would (correctly) be judged as significant, even if the Bonferroni correction is used, but in general the Bonferroni correction is so conservative that its use will typically cause us to miss true task-related activity in many voxels. The good news is that for fMRI data, the Bonferroni correction is much too conservative. The Bonferroni correction is exact when all the tests are statistically independent. With fMRI data, however, spatial correlations guarantee a positive correlation between test statistics in neighboring voxels. Thus, if significance is found in one voxel then the probability of obtaining significance in neighboring voxels is above chance. As a result, the critical value specified by the Bonferroni correction is too extreme. The bad news is that no exact solution to this problem is known. Even so, many different solutions to this problem have been proposed. Different methods are popular because they make different assumptions and have different goals. Most of the parametric methods rely on the theory of Gaussian random fields [28, 29]. Included in this list are all of the most popular cluster-based methods. These methods all require that spatial smoothing is performed during preprocessing. The most popular nonparametric methods include permutation methods [30] and methods that attempt to control the false discovery rate (FDR), rather than the false positive rate. The idea behind the FDR approach is that with many tests, a few false positives should not be feared [31]. So instead of trying to control the experiment-wise probability of a false positive, the FDR approach argues that a more important goal should be to limit the proportion of significant results that are false positives. In other words, consider the set of all voxels for which the null hypothesis of no signal is rejected. The goal of the FDR approach is to limit the proportion of these voxels for which the null hypothesis was incorrectly rejected.

In the standard correlation analysis so far considered, the GLM is applied separately to every voxel in the whole brain or region of interest. After the multiple comparisons problem is solved, a considerable challenge still remains to interpret the results of all these analyses. For example, suppose the analysis reveals strong task-related activation in the dorsolateral prefrontal cortex and in the dorsal striatum. Because these two significance decisions were based on independent applications of the GLM, we have no basis to conclude that these areas are functionally connected in the task we are studying. It could be that they are both part of independent neural networks that just happened to both be activated at similar times. So an important next step in the data analysis process is to identify functionally connected neural networks that are mediating performance in the task under study. This phase of the data analysis is known as connectivity analysis.

The idea underlying connectivity analysis is that a standard GLM analysis identifies clusters (or voxels) that show task-related activation, but it does not specify whether any pair of these clusters is part of the same or different neural networks.

If two clusters are part of the same network then they should be functionally connected in the sense that activation in one might cause activation in the other, or at least the separate activations in the two clusters should be correlated. If instead, the clusters are in separate neural networks then we would not expect either of these two conditions to hold. Connectivity analysis then, is done after a GLM analysis, with the goal of determining which brain regions in the task-related activation map are functionally connected.

An obvious method for testing whether two brain regions are functionally connected in a particular cognitive task is to measure the correlation between the BOLD responses in the two regions across TRs in an experiment where the task is performed. Regions that work together to mediate performance in the task should have correlated neural activations – that is, they should both be active while the task is being performed and they should both be inactive during rest periods. So one approach to connectivity analysis is to look for voxels or groups of voxels in different brain regions that show correlated BOLD responses (i.e., correlated across TRs). A simple solution to this problem is to compute the standard Pearson correlation between voxels or regions. A more sophisticated, yet similar approach uses Granger causality, which is a conceptually simple method that originated in the economics literature [32]. The idea is that if activation in region X causes activation in region Y, then knowledge of earlier activations in region X should improve our prediction of the current activation in region Y. Granger causality tests for such prediction using autoregressive models that are applied via the GLM [33].

One weakness of both Pearson correlation and Granger causality is that they both can fail because of high-frequency noise and/or because the hrf in the two brain regions might be different. A popular alternative is to compute correlations in the frequency domain using coherence analysis, rather than in the time domain, as is done with the Pearson correlation and with Granger causality [34]. Coherence analysis has an advantage over methods that compute correlations in the time domain because the coherence between BOLD responses in two brain regions is unaffected by hrf differences across the regions or by high-frequency noise.

The GLM-based methods of data analysis that compute correlations between predicted and observed BOLD responses require that we specify exactly when neural activation turns on and off. With free-running designs, this is often impossible. For this reason, data analysis choices are severely limited with free-running designs. Perhaps the most popular approach is to compute inter-subject correlations (ISCs; [35]). This method assumes that every subject experiences the same stimulation during the functional imaging. For example, if subjects are watching a movie, then every subject must see the same movie and the onset of the movie must begin at the same time for every subject. Under these conditions, the idea is to correlate the BOLD responses across TRs for every pair of subjects. If a brain region is responding to the movie then its activity should modulate up and down as the movie unfolds in similar ways in different subjects. So the ISC method identifies as task relevant, those voxels where the mean correlation across subjects is high.

Correlation-based analyses that use the GLM are univariate. This means that they analyze the data one voxel at a time. Univariate methods assign completely separate

parameters to every voxel, which means they assume that the data in neighboring voxels have no relationship to each other. *Post hoc* methods are then applied to try to overcome this obviously incorrect assumption. Included in this list are methods for correcting for the multiple comparisons problem that arises from this univariate approach, and the various connectivity analyses that attempt to recover information about spatial correlations from the many independent tests. An alternative approach is to perform multivariate data analyses that attempt to answer the significance and functional connectivity questions at the same time while also completely avoiding the multiple comparisons problem. The trick is that multivariate approaches identify task-related networks, rather than task-related clusters.

Two multivariate methods for analyzing fMRI data are especially popular. One is independent components analysis (ICA; [36, 37]). ICA, like its relative, principal components analysis, decomposes the observed BOLD data into a set of independent components. It operates on data from the whole brain at once, so the components it identifies are functionally independent neural networks that are simultaneously active during some fMRI experiment. For example, one network might mediate the processing of task-relevant stimulus information, one might mediate processing of any feedback provided during the course of the experiment, one might monitor the auditory stimulation provided by the scanning environment, and one could be the so-called default mode network, which can be seen when subjects lie quietly in the scanner with no task to perform (e.g., [38]). Each network defines a spatial pattern of activation across all voxels in the brain. When the network is active, voxels in brain regions that are part of the network will be active, and voxels that are not part of the network will be inactive. On any TR, ICA assumes that the observable BOLD response is a mixture of the activation patterns associated with each of these networks plus (perhaps) some noise. As the TRs change, the amount that each network is activated could change. For example, the network that processes the stimulus should become more active on TRs during and immediately after stimulus presentation and become less active during rest periods. So, on every TR, ICA estimates a weight for each neural network that measures how active that network is on that TR.

ICA has some significant advantages over standard, univariate GLM-based methods of data analysis. First, because it operates on all data simultaneously, ICA largely avoids the intractable multiple comparisons problem that plagues univariate analyses. Second, ICA identifies networks of event-related voxels, rather than single voxels, so it simultaneously addresses questions of functional connectivity. Third, GLM approaches assume that every time an event occurs during an experimental session, it elicits exactly the same BOLD response. In contrast, ICA allows the weights to differ on every TR so it allows the gradual ramping up or down of a network across TRs that might be seen during learning or habituation. Fourth, ICA makes no assumptions about the hrf or the nature of the BOLD response. In particular, it does not assume linearity between neural activation and the BOLD response. On the other hand, ICA does have weaknesses. First, it is time and resource consuming to run. Second, it is a purely exploratory data-analytic technique in the sense that it provides no straightforward method of testing specific *a priori* hypotheses about any of the components. Finally, it provides no foolproof method of identifying task-related components. In

many cases, an ICA analysis might identify several hundred components, only a few of which are likely related to the task under study. Finding these few components of interest among the hundreds identified can be a difficult challenge.

Recently, another multivariate method for analyzing fMRI data has become popular. This method, called multi-voxel pattern analysis (MVPA), applies machine learning classification methods to BOLD response data [39–41]. The idea is that if some brain region is responding differently to two different event types then it should be possible to find a classification scheme that can look at the responses of all voxels in that region and correctly identify which event triggered the response. This is the technique that is used in all of the well publicized claims that fMRI can be used to read one's mind. The first step in MVPA is to create a vector that represents the BOLD response to a specific event in a region of interest. In the simplest case, for every voxel in the region, the vector might have one entry that measures the BOLD response in that voxel to a single specific event [42]. For example, suppose we are interested in whether a region in ventral temporal cortex that includes 2000 voxels responds differently to pictures of shoes versus chairs [43]. For each picture presented to the subject, we estimate the BOLD response in every voxel in this region. Now imagine a 2000 dimensional space with a coordinate axis for each of the 2000 voxels. Each vector specifies a numerical value on each of these dimensions, and therefore we could plot the entries in each vector as a single point in this high dimensional space. The idea is that vectors from trials when a shoe is presented should cluster in a different part of the space compared to vectors from trials when a chair is presented if this brain region responds differently to these two stimulus classes. On the other hand, if this region does not discriminate between shoes and chairs then the points should all fall in the same region. Of course, it would be impossible to decide whether the points fall in the same or different regions by visual inspection. Instead, machine learning classification techniques are used (e.g., the support vector machine or the naïve Bayes classifier).

5.7 The Future

Despite its limitations, the future for fMRI is bright. It is likely to play an enduring role in psychology, cognitive neuroscience, and the mind sciences in general—at least until it is replaced by some similar, but more powerful technology (e.g., just as fMRI has now largely replaced PET scanning). There are several reasons for this optimism. Perhaps the most obvious is that fMRI allows scientists to investigate questions that before seemed unapproachable. But another reason that fMRI is likely to maintain its popularity is that it is rapidly improving on almost all fronts. New scanners and head coils are more reliable and produce cleaner data with higher signal-to-noise ratio. New pulse sequences allow for innovative types of scanning (e.g., as when diffusion spectrum imaging was developed as a superior alternative to diffusion tensor imaging). New methods of data analysis allow researchers to draw unique conclusions even from data collected using common pulse sequences on older

established machines. All these improvements also increase the flexibility of fMRI. Researchers continue to have more choices than ever when designing and running fMRI experiments and when analyzing the resulting data. As a result, potential applications of fMRI are largely limited by the creativity of fMRI researchers.

Exercises

1. Think of an experiment that is best addressed using fMRI instead of EEG or MEG. What are the key advantages of fMRI?
2. In an fMRI experiment with a TR of 2 s, the temporal resolution is considerably better than 2 s. How can the temporal resolution in fMRI be better than the TR?
3. Suppose we measure the height of 15 subjects, and then run each of them in an fMRI experiment. During data analysis we compute a whole-brain t-map for each subject. Next, for every voxel, suppose we correlate (across subjects) subject height with the value of the t statistic in that voxel. What would you conclude, if after correcting for multiple comparisons, we find a set of voxels where the correlation is significant? Does this outcome seem likely? Did the experiment identify a network that thinks about the subject's height?

Further Reading

Several books provide overviews of the whole fMRI field [20, 44], while others provide more depth on certain subtopics. For example, Hashemi, Bradley, and Lisanti [45] give a mostly nontechnical description of MR physics, whereas Haacke, Brown, Thompson, and Venkatesan [46] provide a much more rigorous treatment. In contrast, Ashby [33] and Poldrack, Mumford, and Nichols [47] focus exclusively on experimental design and fMRI data analysis.

Acknowledgments This research was supported in part by AFOSR grant FA9550-12-1-0355 and by the U.S. Army Research Office through the Institute for Collaborative Biotechnologies under grant W911NF-07-1-0072.

References

1. Dobbs D (2005) Fact or phrenology? Scientific American Mind
2. Waldschmidt JG, Ashby FG (2011) Cortical and striatal contributions to automaticity in information-integration categorization. *Neuroimage 56*:1791 -1802
3. Ashby FG, O'Brien JB (2005) Category learning and multiple memory systems. Trends Cogn Sci 2:83–89
4. Lopez-Paniagua D, Seger CA (2011) Interactions within and between corticostriatal loops during component processes of category learning. J Cogn Neurosci 23:3068 -3083

5. Seger CA, Cincotta CM (2005) The roles of the caudate nucleus in human classification learning. J Neurosci 25:2941–2951
6. Seger CA, Peterson EJ, Cincotta CM, Lopez-Paniagua D, Anderson CW (2010) Dissociating the contributions of independent corticostriatal systems to visual categorization learning through the use of reinforcement learning modeling and Granger causality modeling. NeuroImage 50:644–656
7. Pauling L, Coryell CD (1936) The magnetic properties and structure of hemoglobin, oxygenated hemoglobin, and carbonmonoxygenated hemoglobin. Proc Nat Acad Sci U S A 22:210–236
8. Friston KJ, Ashburner JT, Kiebel SJ, Nichols TE, Penny WD (eds) (2007) Statistical parametric mapping: the analysis of functional brain images. Academic, London
9. Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TEJ, Johansen-Berg H, Bannister PR, De Luca M, Drobnjak I, Flitney DE, Niazy R, Saunders J, Vickers J, Zhang Y, De Stefano N, Brady JM, Matthews PM (2004) Advances in functional and structural MR image analysis and implementation as FSL. NeuroImage 23:208–219
10. Woolrich MW, Jbabdi S, Patenaude B, Chappell M, Makni S, Behrens T, Beckmann C, Jenkinson M, Smith SM (2009) Bayesian analysis of neuroimaging data in FSL. NeuroImage 45:173–186
11. Ogawa S, Lee TM, Kay AR, Tank DW (1990) Brain magnetic resonance imaging with contrast dependent on blood oxygenation. Proc Nat Acad Sci 87:9868–9872
12. Ogawa S, Lee TM, Nayak AS, Glynn P (1990) Oxygenation-sensitive contrast in magnetic resonance imaging of rodent brain at high magnetic fields. Magn Reson Med 16:9–18
13. Logothetis NK (2003) The underpinnings of the BOLD functional magnetic resonance imaging signal. J Neurosci 23:3963–3971
14. Logothetis NK, Pauls J, Augath M, Trinath T, Oeltermann A (2001) Neurophysiological investigation of the basis of the fMRI signal. Nature 412:150–157
15. Boynton GM, Engel SA, Glover GH, Heeger DJ (1996) Linear systems analysis of functional magnetic resonance imaging in human V1. J Neurosci 16:4207–4221
16. Buxton RB, Frank LR (1998) A model for coupling between cerebral blood flow and oxygen metabolism during neural stimulation. J Cerebral Blood Flow Metab 17:64–72
17. Vazquez AL, Noll DC (1998) Non-linear aspects of the blood oxygenation response in functional MRI. NeuroImage 8:108-118
18. Huettel SA, Singerman JD, McCarthy G (2001) The effects of aging upon the hemodynamic response measured by functional MRI. NeuroImage 13:161–175
19. Richter W, Richter M (2003) The shape of the fMRI BOLD response in children and adults changes systematically with age. NeuroImage 20:1122–1131
20. Huettel SA, Song AW, McCarthy G (2004) Functional magnetic resonance imaging. Sinauer, Sunderland
21. Ashburner J, Friston K (2007) Rigid body registration. In: Friston KJ, Ashburner JT, Kiebel SJ, Nichols TE, Penny WD (eds) Statistical parametric mapping: the analysis of functional brain images. Academic, London, pp 49–62
22. Klein A, Andersson J, Ardekani BA, Ashburner J, Avants B, Chiang M-C, Christensen GE, Collins DL, Gee J, Hellier P, Song JH, Jenkinson M, Lepage C, Rueckert D, Thompson P, Vercauteren T, Woods RP, Mann JJ, Parsey RV (2009) Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. NeuroImage 46:786 -802
23. Talairach J, Tournoux P (1988) Co-planar stereotaxic atlas of the human brain: 3-Dimensional proportional system—an approach to cerebral imaging. Thieme Medical Publishers, New York
24. Friston KJ, Frith CD, Liddle PF, Frackowiak RS (1991) Comparing functional (PET) images: the assessment of significant change. J Cerebral Blood Flow Metab 11:690–699
25. Friston K, Holmes A, Worsley K, Poline J, Frith C, Frackowiak R (1995) Statistical parametric maps in functional imaging: a general linear approach. Hum Brain Mapp 2:189–210
26. O'Doherty JP, Hampton A, Kim H (2007) Model-based fMRI and its application to reward learning and decision making. Ann NY Acad Sci 1104:35–53
27. O'Doherty J, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ (2004) Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science 304:452-454

28. Worsley KJ (1995) Estimating the number of peaks in a random field using the Hadwiger characteristic of excursion sets with applications to medical images. Ann Stat 23:640–669
29. Worsley KJ, Marrett S, Neelin P, Vandal AC, Friston KJ, Evans AC (1996) A unified statistical approach for determining significant signals in images of cerebral activation. Hum Brain Mapp 4:58–73
30. Nichols TE, Holmes AP (2001) Nonparametric permutation tests for functional neuroimaging: A primer with examples. Hum Brain Mapp 15:1–25
31. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Methodol 57:289–300
32. Granger CWJ (1969) Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37:424–438
33. Ashby FG (2011) Statistical analysis of fMRI data. MIT Press, Boston
34. Sun FT, Miller LM, D'Esposito M (2004) Measuring interregional functional connectivity using coherence and partial coherence analyses of fMRI data. NeuroImage 21:647-658
35. Hasson U, Nir Y, Levy I, Fuhrmann G, Malach R (2004) Intersubject synchronization of cortical activity during natural vision. Science 303:1634–1640
36. Calhoun VD, Adali T, Pearlson GD, Pekar JJ (2001) Spatial and temporal independent component analysis of functional MRI data containing a pair of task-related waveforms. Hum Brain Mapp 13:43–53
37. McKeown MJ, Makeig S, Brown GG, Jung T-P, Kindermann SS, Bell AJ, Sejnowski TJ (1998) Analysis of fMRI data by blind separation into independent spatial components. Hum Brain Mapp 6:160–188
38. Buckner RL, Andrews-Hanna JR, Schacter DL (2008) The brain's default network: Anatomy, function, and relevance to disease. Ann NY Acad Sci 1124:1–38
39. Haynes J, Rees G (2006) Decoding mental states from brain activity in humans. Nat Rev Neurosci 7:523–534
40. Norman K, Polyn SM, Detre G, Haxby JV (2006) Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. Trends Cogn Sci 10:424–430
41. Pereira F, Mitchell T, Botvinick M (2009) Machine learning classifiers and fMRI: a tutorial overview. NeuroImage 45(1 Suppl):S199–209
42. Mumford JA, Turner BO, Ashby FG, Poldrack RA (2012) Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. NeuroImage 59:2636–2643
43. Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P (2001) Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293:2425–2430
44. Buxton RB (2002) Introduction to functional magnetic resonance imaging: principles and techniques. Cambridge University Press, New York
45. Hashemi RH, Bradley WG Jr, Lisanti CJ (2004) MRI: the basics, 2nd Ed. Lippincott Williams & Wilkins, Philadelphia
46. Haacke EM, Brown RW, Thompson MR, Venkatesan R (1999) Magnetic resonance imaging: physical principles and sequence design. Wiley, New York
47. Poldrack RA, Mumford JA, Nichols TE (2011) Handbook of fMRI data analysis. Cambridge University Press, New York

Chapter 6
An Introduction to Neuroscientific Methods: Single-cell Recordings

Veit Stuphorn and Xiaomo Chen

Abstract This chapter describes the role of single-cell recordings in understanding the mechanisms underlying human cognition. Cognition is a function of the brain, a complex computational network, whose most elementary nodes are made up out of individual neurons. These neurons encode information and influence each other through a dynamically changing pattern of action potentials. For this reason, the activity of neurons in the awake, behaving brain constitutes the most fundamental form of neural data for cognitive neuroscience. This chapter discusses a number of technical issues and challenges of single-cell neurophysiology using a recent project of the authors as an example. We discuss issues such as the choice of an appropriate animal model, the role of psychophysics, technical challenges surrounding the simultaneous recording of multiple neurons, and various methods for perturbation experiments. The chapter closes with a consideration of the challenge that the brain's complexity poses for fully understanding any realistic nervous circuit, and of the importance of conceptual insights and mathematical models in the interpretation of single-cell recordings.

6.1 Introduction

The fundamental goal of cognitive neuroscience is the explanation of psychological processes by their underlying neural mechanisms. This explanatory goal is reductionist and operates under the assumption that some form of identity hypothesis is correct, i.e., that specific mental events or processes are identical or intimately linked to specific neuronal events and processes. An explanation therefore only starts with a description of the processes on the neuronal level that give rise to the processes on the psychological level. A full explanation also requires a specification of the

V. Stuphorn (✉) · X. Chen
Department of Psychological and Brain Sciences, Johns Hopkins University, 338 Krieger Hall, 3400 N. Charles St., Baltimore, MD 21218, USA
Tel.: (410) 516-7964
e-mail: veit@jhu.edu

V. Stuphorn
Department of Neuroscience, Johns Hopkins University School of Medicine and Zanvyl Krieger Mind/Brain Institute, Baltimore, MD, USA

B. U. Forstmann, E.-J. Wagenmakers (eds.), *An Introduction to Model-Based Cognitive Neuroscience,* DOI 10.1007/978-1-4939-2236-9_6

exact causal link between the two levels, i.e., a hypothesis about which of the many physical phenomena in the brain is thought to correspond with a specific mental phenomenon.

Thus, the first question is the adequate level of description of the brain at which this causal (or explanatory) link can be established. In general, there are three broad levels at which brain activity can be described, which relate to three different sets of measurement technologies that are currently available. The first level encompasses all the subcellular, molecular processes that explain the behavior of a neuron. This level of description includes for example biochemical and biophysical investigations of receptors, G-proteins, ion channels, and other building blocks that determine the internal organization and workings of neurons.

The second level encompasses the electrophysiological activity of individual neurons or circuits of individual neurons. This level includes experiments in which the temporal pattern of action potentials of individual neurons is recorded, while behaviorally relevant sensory, motor, or cognitive variables are changed, and is the primary topic of this chapter. Experiments on individual neurons allow one to investigate whether neuronal activity (the temporal pattern of action potentials, or spikes) represents (is correlated with) behaviorally relevant information. This level of description also includes the connection and interaction between individual neurons across different brain areas. Importantly, perturbation experiments in which neuronal activity is either suppressed or enhanced, allow one to go a step further and to establish causal links between spiking activity and behavioral functions.

The third level encompasses experiments aimed at recording mass action of large numbers of neurons. Human imaging experiments (fMRI, PET) fall into this category, as does electrophysiological recordings of field potentials at varying scales (LFP, ECoG, EEG). This third level somewhat overlaps with the second level, insomuch as the second level of description includes simultaneous recordings of many individual neurons within a local circuit, or across different parts of the brain. The main distinction is essentially methodological; level two descriptions are of identified, individual neurons, while level three descriptions are of unidentified, averaged neurons. This summing up over many neurons is due to technical constraints of the measurement techniques used, and leads to lower spatial and temporal resolution. Consequently, recordings of mass activity are likely to be most accurate in cases in which most neurons in a particular location have similar functions and activity patterns. However, recent research has shown that even in primary sensory areas, but particularly in associate brain regions, such as frontal and parietal cortex, individual neurons with different functional roles are often located in close vicinity to each other. We will see an example of this later in this chapter (see Fig. 6.2). Here mass action recordings of brain activity will likely have a lower resolution in the identification of functionally relevant signals. These disadvantages are balanced by two great advantages: the ability to record activity from many—or, indeed, all—parts of the brain, and the non-invasive nature of the measurement methods, which permits their routine use in humans.

6.2 Single Neurons Provide the Critical Link Between Brain and Cognition

It is clear that we can learn from all available techniques and their respective usefulness will depend on the specific question at hand, as well as technical constraints. Thus, pragmatic considerations will lead scientists always towards using all sources of information. Nevertheless, we can ask from a theoretical point of view at which of the three levels we can best articulate the relationship between particular mental (or cognitive) states and neural states [1–3]. Ever since the pioneering work of Adrian and Hartline [4, 5], individual neurons are seen as the elementary units of the nervous system that represent information and perform computations on these representations [2, 3, 6–10]. There is general agreement that the temporal structure of action potentials encodes the information. However, the exact nature of this code is still under active investigation [11–14]. Mountcastle was the first to formulate a research program centered on a systematic comparison of psychophysical measures in conscious subjects and recordings of individual neurons [7, 15]. At present, this program has resulted in multiple examples of individual neurons, whose firing patterns match to a stunning degree with mental states, such as perceptions or decisions, as measured using psychophysical methods [16–18].

A particularly impressive example of such a match between neural activity and perception are results from recordings in single somatosensory nerve fibers in humans during stimulation of the skin [19]. Near-threshold skin indentations resulted in a variable response of the nerve fiber. During some trials, an action potential was generated, while on other trials no electrophysiological response was observed. Astonishingly, on trials in which an action potential was registered, the human subjects reported the subjective experience of a light touch. On trials with an identical mechanical stimulation, but without an action potential, the humans reported no touch perception. This finding implies a very strong linking hypothesis, according to which a single action potential in a peripheral nerve elicits a particular mental state. Similar findings linking changes in the activity of individual neurons to changes in mental state have been observed in cortical neurons of animals [20, 21].

All of these findings point towards the spiking activity of single cells in awake, behaving animals (including humans) as the key level for understanding how physical events in the brain underlie mental events and cognition [2, 3]. So, what is then the best way in which we can get these critical experimental data, and what are the technical requirements? Some of the technical requirements are described by Crist and Lebedev [22]. They include the choice of an appropriate animal model, useful behavioral tasks, methods for electrophysiological recordings of one or more single neurons, methods for perturbing spiking activity in the brain, and data analysis. In the following section of this chapter we will go, one by one, over these different requirements, using an ongoing research project from our laboratory as an example. We chose our own work mainly because we are most familiar with it. Wherever appropriate, we will refer to the works of others to illustrate different approaches than the ones we used.

6.3 Choice of Animal Model

Traditionally, the majority of the electrophysiological investigations of the sensory, motor and cognitive functions of the brain have been done in primates. Techniques for recording from individual neurons in awake, behaving primates were pioneered by Evarts [23, 24] and then further developed by Mountcastle, Wurtz and others [25, 26]. This is in contrast to the majority of modern biomedical research in which rodents, in particular mice and rats, are the dominant animal models. The reasons for this preference are the greater number of genetic and other molecular biological tools that are available in these animals, because of the much shorter generational span of rodents compared to primates. More recently, rodents have been used increasingly to study the neural mechanisms of cognitive functions, such as decision-making under uncertainty [27, 28]. The fact that rats can be trained in sophisticated behavioral tasks opens up the question, to what extent they might not be a superior animal model relative to monkeys. This is a particularly pressing question, since neuroscience is at the moment in the middle of a technical revolution. New tools for observing neural activity of large numbers of neurons optically, such as two photon imaging [29], and the automation of anatomical methods [30] allows for an unprecedented level of insight into the activity of large numbers of neurons, and their internal connection. The functional relevance of identified types of neurons can be probed using optogenetic tools [31]. All of these new tools have been developed in rodents, in particular mice.

We used macaque monkeys in our study, and we feel that there are still strong reasons that support the continued use of this animal model, in particular in cognitive neuroscience. The most important reason is the fact that there are radical anatomical and structural differences between the brains of rodents and primates [32]. This is particularly true for the frontal cortex, which is generally believed to be essential for higher cognitive function in humans and other mammals [33, 34].

Based on cytoarchitectonical and structural differences between different areas in the frontal lobe of rodents and primates, Wise suggested that primates have evolved certain new areas that do not exist in rodents [32, 35]. Recent support for this hypothesis comes from fMRI experiments in humans that show a regional specialization in the representations of primary and secondary, abstract reward in the orbitofrontal cortex [36]. Whereas the anterior lateral orbitofrontal cortex, a phylogenetically recent structure (only present in primates), processes monetary gains, the posterior lateral orbitofrontal cortex, phylogenetically and ontogenetically older (and shared with rodents), processes erotic stimuli, a more basic reward. Interestingly, the phylogenetically newer parts make up the majority of the frontal cortex in primates [32].

These differences in frontal architecture and their unknown functional consequences can lead to difficulties in the interpretation of neuroscientific findings. For example, reports in monkeys have claimed that the activity of certain neurons in the orbitofrontal cortex represents uncertainty and risk (defined as outcome variability) [37]. This finding is in agreement with human neuroimaging studies [38]. Recently,

a very clever study in rats suggests that this neuronal activity pattern might not represent risk per se, but instead acquired salience [28]. However, while this finding is intriguing, it will need to be replicated in primates, simply to make sure that the functional differences revealed in these two studies are not the result of differences in the functional architecture and overall function of orbitofrontal cortex in monkeys and rats.

Another reason to use primates is related to the potential for what might be called 'behavioral mimicry'. Organisms with completely different internal architectures can generate behavior that looks similar, but is produced for entirely different reasons. The formal mathematical proof of this possibility was derived in the theory of finite automata [2, 39]. In such a case of mimicry, the behavioral similarity is likely to be only superficial and strongly context-dependent. A real-world example is the response of rodents, macaques and humans in reward-reversal tasks. In such tasks, one of two options is consistently rewarded and, therefore, almost exclusively chosen. If, however, the reward contingencies are unexpectedly switched without notice, so that the previously unrewarded option will now lead consistently to reward, rodents, monkeys, and humans will all learn to switch their preferences to the newly rewarded option. From this qualitative similarity, one might conclude that very similar, perhaps even identical, choice and learning mechanisms underlie this behavior in all of these organisms. However, such a conclusion does not take into account some intriguing differences in the time course of the switch. Human subjects need typically only one error trial to switch [40]. In contrast, rodents switch their behavior only after 20–40 trials [41]. This is (at least from a human point of view) a staggering amount of exposure to a clear-cut, non-probabilistic change in reward contingencies. This seems to imply that at least humans represent the task contingencies in a different way and might use different learning or choice mechanisms than rodents. Thus, the picture that emerges is complex. Monkeys need at least 10–15 trials [42], which is still different from humans, but closer to them than the behavior of rodents.

Obviously, primates are not superior animal models with regards to all possible research questions. In general, the choice of rodents as models for human behavioral, neural, and even mental processes is likely to be most appropriate if the object of study is an aspect of behavior and the brain that is common among all mammals. An example is the role of the neural circuits in the hypothalamus in the control of hunger and food consumption [43]. However, even for something as seemingly primitive as appetite and food consumption there exist important behavioral differences between humans and other mammals with less complex brains. For instance, humans show reliable behavioral and neural differences while consuming the same wine, when given different information about its price [44]. Thus, the choice of appropriate animal model ultimately depends on the research question. Within the domain of cognitive neuroscience, it seems to us that non-human primates are still the obvious choice, given their overall similarity with humans, and the fact that many of the new techniques first developed in rodents are now applied to primates [45–49].

Of course, there are also large differences between humans and other non-human primates [50]. These differences will likely forever preclude the study of certain human abilities, such as language, in animal models. This is important, because

language, and the abstract, symbol-operating cognitive abilities that go along with it, pervade every other aspect of human mental and cognitive life, from memory to decision-making. In addition, there are likely to be other, potentially more subtle differences in the way cognitive mechanisms operate in humans and other primates. It is therefore of great interest to use every opportunity to study single-unit responses in awake humans [51]. This approach has already led to some insights into language [52], representation of objects [53], and cognitive control [54]. In addition, single-unit recordings can provide new insights into mental diseases, such as obsessive-compulsive disorders [55]. This resource should be used more widely by cognitive neuroscientists.

6.4 Behavioral Tasks and Psychophysics

The ability to link neuronal activity and cognitive function depends critically on our ability to vary the cognitive signals of interest in a controlled and measurable fashion. This, in turn, depends entirely on the behavioral task design, and the psychophysical methods used to analyze behavior and deduce cognitive states from it. Thus, single unit electrophysiology, and mathematical psychology and modeling are critically linked [Forstmann, Wagenmakers, chapter of this book]. Mathematical psychology provides formal models of cognitive processes, which afford quantifiable variables that are related to behavior in an operational manner and that can be compared to measures of neuronal activity. Of course in practice, the hypothesized link might turn out not to exist, because the model might not have been an appropriate description of the underlying cognitive and neuronal mechanism. However, this form of model falsification is exactly how science progresses.

In the case of the specific project that we chose as an example, we were interested in the neuronal mechanism underlying value-based decision making. Decision-making involves the selection of a particular behavioral response from a set of two or more possible responses. Our understanding of the neural processes that underlie this selection process is most advanced in the case of perceptual decisions [56]. These decisions are guided by external sensory stimuli and reflect the beliefs of the decision maker about the external world. Value-based decisions, on the other hand, are much less well understood.

Value-based decision making is the process of selecting an action among several alternatives based on the subjective value of their outcomes. This requires the brain to first estimate the value of the outcome of each possible response, and then to select one of them on the basis of those values [57–60]. This raises two fundamental questions: (1) Where in the brain are the values of different types of outcomes and actions represented and how are these value signals computed? and (2) How and where does the brain compare those value signals to generate a choice?

With respect to the first question, a rapidly growing number of studies have found neural responses that are correlated with some form of value signals [57, 61–63]. Several studies found orbitofrontal cortex (OFC) and amygdala encoding the value

of different goals [62, 64–67]. These signals are stimulus-based and independent of the actions required to obtain them. These *option value* signals represent therefore predicted future states of the world. To allow the selection of an appropriate action, these goal representations need to be associated with the actions that are most likely to bring them about. This type of value signal is known as *action value*. Action-value signals for hand and eye movements have been found in the striatum [68, 69], in the dorsolateral prefrontal cortex (DLPFC) [70], and in the medial frontal cortex [71, 72], including the supplementary eye field (SEF) [73].

With respect to the second question, there are currently two major theories of how the brain compares value signals and uses them to select an action [74]. One theory is the goods- or goal-based theory of decision making [62], according to which the brain computes the subjective value of each offer, selects between these option value signals, and then prepares the appropriate action plan. This theory in its purest form predicts that motor areas should only represent the chosen action. The other theory is the action-based theory of decision making [18, 75–78], according to which potential actions are simultaneously represented in the brain and compete against each other. This competition is biased by a variety of factors including the subjective value of each offer (i.e., their action values). This theory in its purest form predicts that option value signals should not predict the chosen option, before an action is chosen, since these signals are only precursors to the decision. A third alternative is that competition occurs at both levels in parallel [74].

In order to study value-based decision making, we needed to design a task in which the monkey was forced to select actions based on its internal estimation about the worth of various options. In our case, we kept the number of alternatives binary to start with the simplest condition. In addition, we were also interested in creating a task in which identical task conditions would elicit different choices. This would allow us to differentiate between the representation of the decision process itself that should co-vary with the behavioral choice, and the representation of other factors that should stay invariant across trials (e.g., the representation of a particular option and its attributes, independent of whether it is chosen or not).

Both of these conditions were fulfilled in a gambling task, in which the monkeys had to choose between gambles with different probabilities to win reward of varying amounts (Fig. 6.1a, b). We used targets that consisted of two colors corresponding to the two possible reward amounts. The portion of a color within the target corresponded to the probability of receiving that reward amount (Fig. 6.1b). The minimum reward amount for the gamble option was always 1 unit of water, while the maximum reward amount varied between 3, 5 and 9 units, with three different probabilities of receiving the larger reward (20, 40, and 80 %). This resulted in a set of 7 gambles. The colors and differences in area size were easy to discriminate for the monkey. Thus, in presenting two targets to the monkey, the problem for the animal was not one of perceptual uncertainty. Instead, the problem of selecting the better of the two options was related to the uncertainty about the actual outcome that would follow from each choice. A decision-maker that is indifferent to risk should base his decision on the sum of values of the various outcomes weighted by their probabilities, i.e., the expected value of the gamble. However, humans and animals

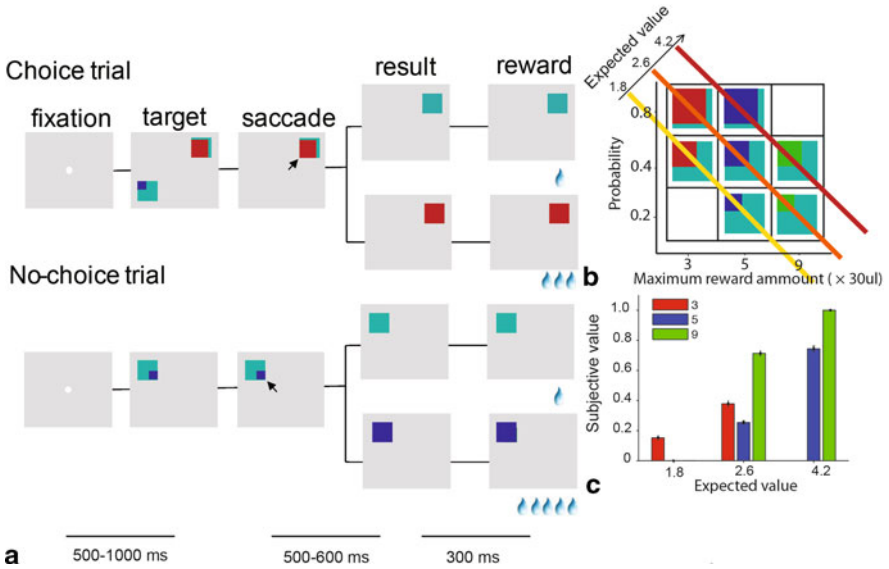


Fig. 6.1 Gambling task and estimate of subjective value. **a** The gambling task consisted of two types of trials, choice trials and no-choice trials. All the trials started with the appearance of a fixation point at the center of the screen, which the monkeys were required to fixate for 500–1000 ms. After that, in choice trial, two targets appeared on two locations that were randomly chosen among the four quadrants. Simultaneously, the fixation point disappeared and within 1000 ms the monkeys had to choose between the gambles by making a saccade toward one the targets. Following the choice, the nonchosen target disappeared from the screen. The monkeys were required to keep fixating the chosen target for 500–600 ms, after which the target changed color. The two-colored square then changed into a single-colored square associated with the final reward amount. This indicated the result of the gamble to the monkeys. The monkeys were required to continue to fixate the target for another 300 ms until the reward was delivered. In the choice trial, each gamble option was paired with all other six gamble options. The sequence of events in no-choice trial was the same as in choice trial except that only one target was presented. In those trials, the monkeys were forced to make a saccade to the given target. All 7 gamble options were presented during no-choice trials. We presented no-choice and choice trials interleaved in blocks of trials that consisted of all twenty one different choice trials and eight onset different trials and seven different no-choice trials. Within a block, the order of trials was randomized. The locations of the targets in each trial were also randomized, which prevented the monkeys from preparing a movement toward a certain direction before the target appearance. **b** Four different colors indicated four different reward amounts (increasing from 1, 3, 5 to 9 units of water, where 1 unit equaled 30 μl). Note that the expected value of the gambles along the diagonal axis was the same. **c** The mean subjective value of the 7 gamble options for one of the monkeys. The subjective value ranges between 0 for the least and 1 for the most valuable option

are not indifferent to risk and their actual decisions deviate from this prediction in a systematic fashion. Thus, the subjective value of a gamble depends on the risk attitude of a decision-maker.

In addition, there is another interesting feature that can be seen in everyday life, as well as in our laboratory task. For certain combinations of gambles, the monkeys

were very certain which one they preferred. However, for a large range of other combinations they varied in their choice, even after they were exposed to the different gamble options daily for many months. One of the reasons for this persistent uncertainty about the value of the gamble options might be the fact that gambles vary across two independent dimensions, reward amount and probability, both of which affect the overall value. Options can be attractive for different reasons, e.g., either because of low risk or high payoff. Assessing the value of a gamble option requires, therefore, a trade-off between the different attributes that have to be integrated in a weighted fashion in order to generate a one-dimensional decision variable, the subjective value of the option. This process has no obvious best solution and agents can remain ambivalent with respect to which of the options is optimal. In addition, there might also be other sources of variance, such as changes in attention or motivation or recent outcome history. In any case, the monkeys showed behavioral variance in our gambling task, which was important for us. Combined, the two features of our task produce a situation that is almost a perfect inversion of the classic perceptual decision-making task, in which sensory stimuli are very ambiguous, but the correct response can easily be selected, given a particular belief about the state of the world [16, 18]. In contrast, in our gambling task, the state of the world is easy to perceive, but the appropriate response is unclear.

In terms of the link between the neural and the mental level, we are faced in our research with the problem of comparing a subjective, internal variable (the subjective value of the various options) to an objective, measurable variable (the firing rate of neurons). Here the behavioral variance is likewise of great importance, since it allows us to use psychophysical scaling techniques [79, 80] to estimate the subjective value of different targets (Fig. 6.1c). These techniques go back to Fechner, who was the first to suggest that a psychophysical experiment could be conducted in which an observer makes judgments along a psychological dimension having no obvious physical correlate (here, subjective value) [81, 82]. Thurstone further developed the theoretical method for analyzing such data from paired comparison judgments [83]. In his model, Thurstone assumed that the judgment process was based on the comparison of noisy internal representations. The differences across the compared distributions lead to systematic differences in the probability of particular judgments. Similar ideas within economics lead to the development of random utility models [84], in which it is presumed that choices are based on noisy internal utility estimations [85]. Importantly, this scaling method gave us an estimate of the subjective value that went beyond a mere ordinal ranking. Instead, we could order the subjective value of the various gambles on an interval scale [82, 86] that allowed us to estimate not only which gamble is preferred over another, but also by how much (Fig. 6.1c).

6.5 Electrophysiological Recordings of One or More Identified Neurons

In our experiment, we used an in-house built system of electrode microdrives that allowed us to independently control the position of up to 6 different electrodes. Our recording setup required us to advance electrodes acutely during each recording session into the brain. This allowed us to cover a wide range of different cortical locations and, more importantly, to position the electrode close to neurons, whose activity was task-related. In this regard, apart from the fact that we did this with multiple electrodes, our approach was very similar to the traditional single electrode recording approach. However, both approaches introduce mechanical disturbances within the brain tissue during advancement of the electrode. These mechanical instabilities, together with pulsations introduced by heart rate and breathing generated instabilities in the position of the neuron with respect to the tip of the electrode. These instabilities influenced our long-term recording stability. To keep the signal to noise ratio of spike identification stable required constant monitoring and minute adjustments of the electrode position by the researcher. This is a well-known problem for single unit electrophysiology, but it is exacerbated in the case of multiple electrodes. Ultimately, for human researchers one reaches very soon an attentional bottleneck. Overcoming these limitations would be a major breakthrough that would allow us to record simultaneously from large numbers of neurons [87].

There are a number of ways to achieve this goal. One possibility is the use of a series of electrode drives that can move electrodes independently operated by multiple researchers working in conjunction [88]. Each researcher is responsible for a few electrodes. This acute recording approach is in some sense the most conservative, insomuch as it requires the least dramatic change relative to traditional methods of single-unit recordings. Because of this, it is easy to implement in principle. However, it is not clear how scalable this approach is, given the increasing demands in well-trained man power.

Another possibility is the use of microelectrode arrays [89–91]. In this approach the electrodes are not advanced acutely for each recording session. Instead, an array consisting of multiple electrodes (with as many recording contacts as desired and technically feasible) is chronically implanted into the brain [92–94]. Due to its better mechanical stability, neuronal spikes can be recorded typically for extended time periods. The signal to noise ratio of microelectrode arrays and of conventional electrodes is comparable [91]. One disadvantage of the microarray recording setup is the inability to actively search for task-relevant neuronal activity. Once implanted, the electrodes cannot be moved and the researcher has to be content with whatever signal he or she can get. To some extent, this disadvantage can even be seen as strength, since pre-selection of ‘interesting’ neurons introduces severe sampling biases in traditional recording studies. This has made it very hard to directly compare the results of single unit studies in different brain areas. Since microarrays sample neurons in different parts of the brain in a more random fashion, they allow a more unbiased comparison [95]. A straightforward and unbiased way to increase the likelihood to

record from task-relevant neurons using chronic microarrays would be a strategy of recording from as many parts of the electrode as possible to increase the number and extent of neurons that can be sampled [90, 94]. This strategy relies on the use of modern lithographic techniques to fabricate electrodes. There has been a lot of progress in the manufacture and use of such polytrodes [96] and there exists a large design space that can be explored for further improvements [97].

A third approach that combines aspects of the acute and chronic recording methods is semichronic recording [98, 99]. In this approach, miniature mechanical microdrives are implanted into the brain, each containing a number of independently movable microelectrodes. Recordings are made by slowly advancing a subset of electrodes in each chamber each day. This procedure has been used very successfully during the investigation of neurons in the rodent hippocampus [100]. Semichronic recording provides the ability to move electrodes into brain areas that are of particular interest, and the possibility of recording from many individual neurons simultaneously. However, this type of recording device is still not commonly used in primate experiments and requires further development.

6.6 Relationship Between Neural Activity and Decision Variables

Decision-making under risk is very common in everyday life, where practically every action can have more than one possible outcome. Value-based decision making requires the translation of internal value signals related to the different options into the selection of unique motor signals necessary to obtain the desired option. Where and how this is achieved is still debated [101]. Therefore, we concentrated our initial research on brain areas that receive input from motivational and cognitive systems and provide output to the motor system. One such region is the supplementary eye field (SEF). SEF receives input from areas that represent option value, such as the orbitofrontal cortex and the amygdala [102, 103]. SEF forms a cortico-basal ganglia loop with the caudate nucleus, which is known to contain saccadic action value signals [104, 105]. SEF projects to oculomotor areas, such as frontal eye field, lateral intraparietal cortex, and superior colliculus [102]. Neurons in SEF become active before value-based saccades, much earlier than neurons in frontal eye field and lateral intraparietal cortex [106]. SEF might therefore participate in the process of value-based decision making in the case of eye movements.

Our initial recordings confirmed that SEF neurons represent three major functional signals during decision-making [107]. One group of neurons encoded the value of reward options, but not the type of eye movement necessary to obtain it. Such option value signals are similar to signals found in the orbitofrontal cortex. These signals appeared first in the SEF. Next, a group of neurons became active that combined information about the value of an option with information about the direction of the saccade necessary to receive the reward. Such action value signals are ideally suited to select the action that will most likely maximize reward. Lastly,

pure motor-related neurons became active that only carried eye movement related signals. This succession of value- to motor-related signals is in line with our working hypothesis that SEF serves as a bridge between the value and the motor systems. We presume that it takes value information and translates it into action value representations. The learning of appropriate action value signals requires a system that can evaluate the outcome of actions that are taken. Interestingly, SEF neurons also carry various monitoring signals [108]. Altogether, these earlier findings suggested that SEF participates in value-based decision making by computing an action value map of the existing choice options. Competition within this map could select the action associated with the highest value, which in turn could be used to guide the selection and execution of the appropriate eye movement.

In order to test this hypothesis, we recorded from SEF neurons, while monkeys chose between gambles of varying subjective values. The histograms in Fig. 6.2 show the activity of multiple identified single neurons during saccades to four different targets recorded in one of these sessions. The activity differences indicate the preferred direction and the strength of the tuning of the different cells. This directional tuning (or lack thereof in the time period preceding the saccade) is of course only one of the functional dimensions along which the SEF neurons can vary [73]. The other major dimension, sensitivity to subjective value of the target, is not shown in Fig. 6.2. However, even while ignoring this other potential source of functional variability; a comparison of the neurons is enough to make clear, why it is important to record from individual neurons. The SEF neurons that were recorded from each of the three electrodes were, as a group, in very close anatomical proximity to each other. Otherwise, it would not have been possible to separate their action potentials from the background modulation of all the other neurons surrounding the electrode tip. Nevertheless, there is a marked difference in directional tuning among these neurons. In particular, one of the neurons recorded by the second electrode (middle column, first row) is most active for saccades to target T3 and least active for saccades to target T4. This is in contrast to two other neurons recorded by the same electrode (middle column, second and fourth row) that show an exactly opposite pattern of activity: these are most active for saccades to target T4 and least active for saccades to target T3. Any form of mass-activity recording would have simply averaged over these differences. In the best case, this would have increased the noise of the recording, and in the worst case it would have led to a failure to detect an important functional difference among the neurons forming the local network.

If a framework based on anatomical proximity is inadequate to functionally understand the SEF neurons, what kind of alternative works better? In our case, it turns out that the functional framework of the action value map works well to give us some insights on the pattern of activity in SEF during decision-making. Figure 6.3 shows the population activity in SEF as a time-direction map of neuronal activity. Here, we sort the neurons according to their preferred direction relative to the position of the chosen and unchosen target. Since the monkey made saccades in four different directions, each neuron contributed four different activity traces to the time-direction map. To avoid a bias introduced by neurons with higher activity levels, we normalized the activity of each neuron across all conditions. To smooth over inevitable

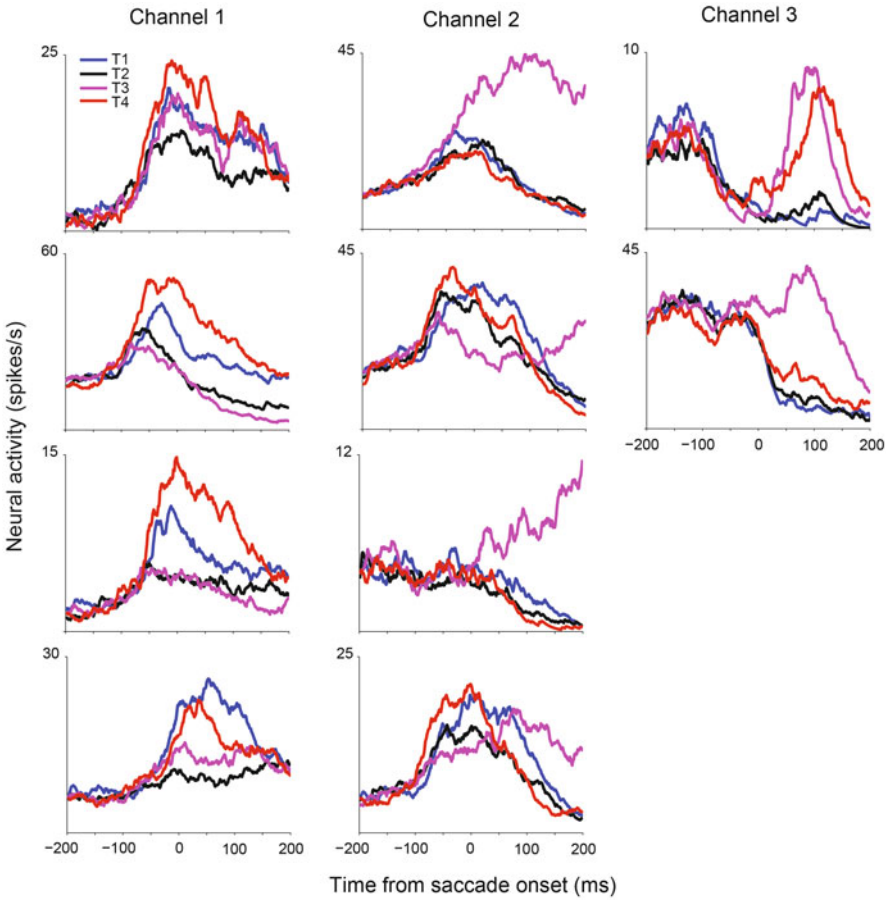


Fig. 6.2 Recording of multiple identified single neurons. An example of numerous individual neurons recorded simultaneously from three different electrodes inserted into different parts of SEF during one recording session. Each panel shows the the average spike rate of one neuron aligned on saccade onset for movements towards each of the four different target locations (*T1*: blue, *T2*: black, *T3*: violet, *T4*: red line). The panel in each of the three columns represents the activity of one individual neuron that was recorded from one of the three electrodes., We were able to isolate four different neurons in the first two electrodes, and two more neurons from the third electrode

differences with which the preferred directions were represented in our neuronal sample, we binned the neuronal activity. It should be understood that this simple act itself represents a form of interpretation or hypothesis regarding the function of the neurons. We presume that each neuron represents the action value of saccades directed towards its preferred direction. Thus, as a whole the activity distribution across the entire neuronal population encodes the combined estimation of the relative values of the various saccades that the monkey can make. Each vertical line in the map represents the state of this activity distribution in the action value map at

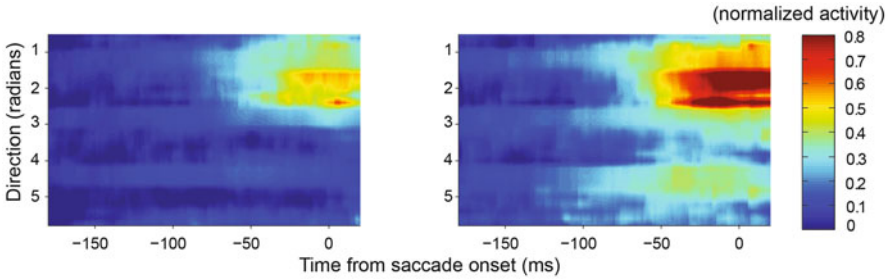


Fig. 6.3 Time-direction map of normalized neuronal activity in SEF. Each of the two maps shows the normalized population activity distribution of directionally SEF neurons as a function of time. We averaged neuronal activity over all different subjective target values. The vertical axis shows the activity distribution of neurons sorted by the orientation of their preferred direction relative to the direction of the chosen target (at 1.57 radians) and the non-chosen target (at 4.71 radians). The horizontal axis shows the change of this activity distribution across time relative to saccade onset. The time-direction map on the left shows the SEF activity during no-choice trials, in which only one target was presented. The time-direction map on the right shows the activity of the same neurons during choice trials. The upper band of activated neurons corresponds to neurons representing the chosen target direction, while the lower band of activated neurons represents the non-chosen target location

one moment in time. Since in our experiment, all targets were presented with the same distance from the center, we can presume that our map here is one-dimensional. Thus, the time-direction map shows the development of action value-related activity over the course of decision-making.

The map on the left shows the simple no-choice case, in which only one target is presented. In response to the target presentation, activity in a broad set of neurons increases. Activity centered on the target direction reaches a maximum around the time of saccade initiation. The map on the right shows the more complex case, in which two targets are presented. There are a number of differences. First, activity starts to rise in two parts of the map. One is centered on the target that will be chosen, while the other is centered on the non-chosen target. The initial rise in activity relative to saccade onset starts earlier, in keeping with the fact that reaction times are longer when the monkey has to choose between two response options. In the beginning, the activity associated with both possible targets is of similar strength, but around 50 ms before saccade onset, a difference develops between these two different groups of cells. The activity centered on the chosen target becomes much larger than the one centered on the non-chosen target, and increases until saccade onset. In fact, the peak activity associated with the chosen target is much larger than the peak activity associated with the same saccade during no-choice trials. This increased activity for increased number of targets is very unusual and differs from the behavior of neurons in other oculomotor regions. However, it might allow SEF to reliably encode the best action value even in the face of distractors. In sum, the SEF population activity seems to represent first both equally possible alternatives, before in a second step reflecting the selection of the chosen target.

These observations then beg the question, whether SEF activity actively takes part in the decision process. To answer this, we have to establish at least two links [2, 3]. The first link is between the variations of neuronal activity and the behavioral choices of the monkey. To establish this link, we have to show that we can decode (i.e., predict) the behavioral choice from the neuronal activity to some statistically significant degree on any given trial. Traditionally, with single units, such a link was established using techniques derived from signal detection theory [109, 110]. These techniques rely on a comparison of the neuron's activity across trials that never occurred simultaneously. For example, all trials, in which the two options A and B are presented, are divided into trials in which option A was chosen and trials in which option B was chosen. By comparing the activity of a neuron across these two types of trials one can hope to see if there are differences related to behavioral choice. However, this entails that the trials recorded from an individual neuron are treated as if they belonged to a pair of neurons. Underlying this analysis is therefore the assumption of a fictitious 'antineuron' that behaves as the mirror image of the recorded neuron, but that was never actually recorded (and most likely does not exist). Apart from these issues, there is the deeper question about the extent to which the different trial repetitions are actually identical. In light of these conceptual problems, it would be better to use the activity of many neurons on a single trial to do what, in the traditional approach, is done with the activity of a single neuron on many trials [111].

A promising new technique for decoding neural activity is the use of modern pattern classification methods to analyze activity recorded simultaneously from multiple neurons [112–115]. An interesting new approach for visualizing the pattern of activity within a large number of neurons is the state space representation [116]. In this general framework, the activity of a group of neurons is represented as a particular point in an N-dimensional space, where N is equal to the number of neurons. The activity of each neuron at a given moment in time is represented numerically along one of the dimensions forming the space. Thus, the entire population forms a vector pointing to the part of the state space that represents the momentary state of the set of neurons. Changes in neuronal activity lead to shifts in the state space that form a trajectory. The direct visualization of this state space is obviously not possible for groups of neurons larger than three. However, it is possible to visualize the main changes in state space following dimensionality-reduction through methods, such as principal component analysis. The mean trajectories describing the shifts of population activity in SEF during decision making are shown in Fig. 6.4a for the set of 10 neurons depicted in Fig. 6.2. The trajectories associated with the choice of the four saccade directions all start in the same part of the state space projection spanned by the first two principal components (PC1, PC2), before moving in four different directions in this state space. The trajectories shown in Fig. 6.4a indicate the mean positions of the state vector. The state vectors associated with the individual trials form a cloud around the mean trajectory.

We can now ask if there is a linear boundary that optimally divides the state space, so that we can distinguish between the state vectors that are associated with particular choices (of options or saccade direction). Next, we can ask how well we

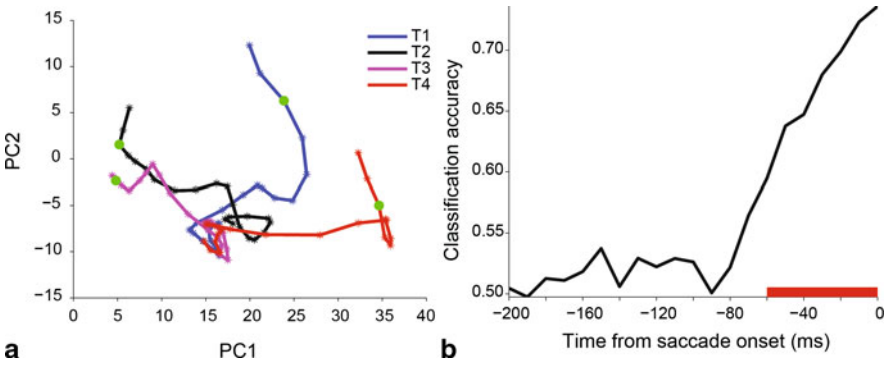


Fig. 6.4 Neuronal dynamics during decision-making within a neuronal state space and results of linear classifier. **a** The activity of the simultaneously recorded neurons shown in Fig. 6.2 defined a 10-dimensional state space. A projection of this state space onto a 2-dimensional subspace is shown. The subspace is defined by the first two principal components (PC1, PC2) explaining variance in the neuronal state vector distribution. The temporal succession of the mean state vector location between target and saccade onset is shown separately for trials in which one of four saccade directions was chosen by the monkey. The mean state vector locations form a trajectory (*T1*: blue, *T2*: black, *T3*: violet, *T4*: red line). The green dot on the trajectories indicates the moment of saccade initiation. **b** A linear discriminant analysis of the distribution of state vectors for all combination of saccade directions was performed. The percentage of correctly predicted choices based on this analysis is plotted as a function of the time bin during which the state vectors were defined. The red bar indicates those time periods, in which the percentage of correct predictions was significantly larger than chance, as determined through a permutation test

can decode the monkey's choices based on this approach, and at what point in time our predictive power is better than chance. The result of this analysis for the same set of 10 SEF neurons is shown in Fig. 6.4b. As we can see, the neuronal assembly does not allow us to predict the chosen saccade direction up until 60 ms before saccade onset. However, after this point the predictive power rapidly increases until it reaches a choice probability of $\sim 75\,\%$ just before saccade onset. This 'decision point' matches the estimate derived from the time-direction map for the moment at which the activity map differentiates (Fig. 6.3).

6.7 Perturbing Spiking Activity in the Brain

Showing that SEF activity is correlated with decisions in the gambling task is a good beginning towards establishing a link between neuronal activity and mental processes. However, the critical step is clearly the establishment of causality. This requires perturbation experiments, to show that changes in neuronal activity cause changes in behavior.

To this end, we used a cooling probe to temporarily inactivate the SEF in both hemispheres, while a monkey performed the gambling task [117]. It has been known

for some time that cooling suppresses the generation of action potentials. This inactivation is fast (with 1–5 min.), reversible, and causes no damage to the affected tissue. All these factors make this technique easier to use than pharmacological inactivation. At the same time, the size and extent of the affected area can be easily controlled by the shape of the cooling probe. Thus, large brain areas can be influenced simultaneously, a relative advantage over optogenetic techniques that are limited by the ability to spread light evenly. On the downside, cooling affects all neurons near the probe, and therefore does not allow the specific manipulation of functionally or anatomically defined neuronal sub-groups.

The behavioral effects of the inactivation are shown in Fig. 6.5. We plot the probability of choosing the less valuable target as a function of the difference in subjective values. Behavior under normal conditions is shown by the blue bars. As one would expect, the probability of choosing the less valuable option is largest when the difference is small and the discrimination of the value difference is hard. For larger value differences ($> 20\,\%$), the monkey typically picks the less valuable target only rarely. If SEF plays a causal role in value-based decision making, we would expect the monkey to show an increased rate of sub-optimal choices when SEF can no longer guide motor selection. This is indeed the case, as shown by the red bars. The effect is largest for intermediate value differences. This is probably due to the fact that there is a ceiling effect for very small value differences, while for very large value differences the decision is so easy that other brain regions beside the SEF are sufficient to pick the better option. Nevertheless, the overall effect of cooling on behavior is significant ($p < 0.01$), and the size of the effect is comparable to the effect of permanent lesions of the orbitofrontal cortex through ablation [118]. Importantly, the fact that inactivation of SEF has an immediate effect on value-based decisions establishes a causal link between SEF single unit activity and the monkey's choice based on subjective preferences (at least with regards to eye movements).

SEF is clearly part of a larger network that is involved in value-based decision making. An important future direction will therefore be the exploration of the other brain areas in the network. One group of areas, such as dorsolateral prefrontal cortex, orbitofrontal cortex and amygdala, provide inputs to the SEF. Another set of areas including frontal eye field, superior colliculus and dorsal striatum, in turn, receive inputs from the SEF. Understanding the entire circuit responsible for value-based decision making will require us to describe the types of signals and their temporal order within this network of brain areas.

6.8 Future Developments

Recently, very ambitious proposals for large-scale projects have been suggested within neuroscience. Many prominent and accomplished neuroscientists have suggested that we should try to reconstruct the neuronal network for large parts if not the entire, brain of a small mammal, a 'structural' connectome [30]. Other researchers have suggested that we should attempt to record every action potential from every

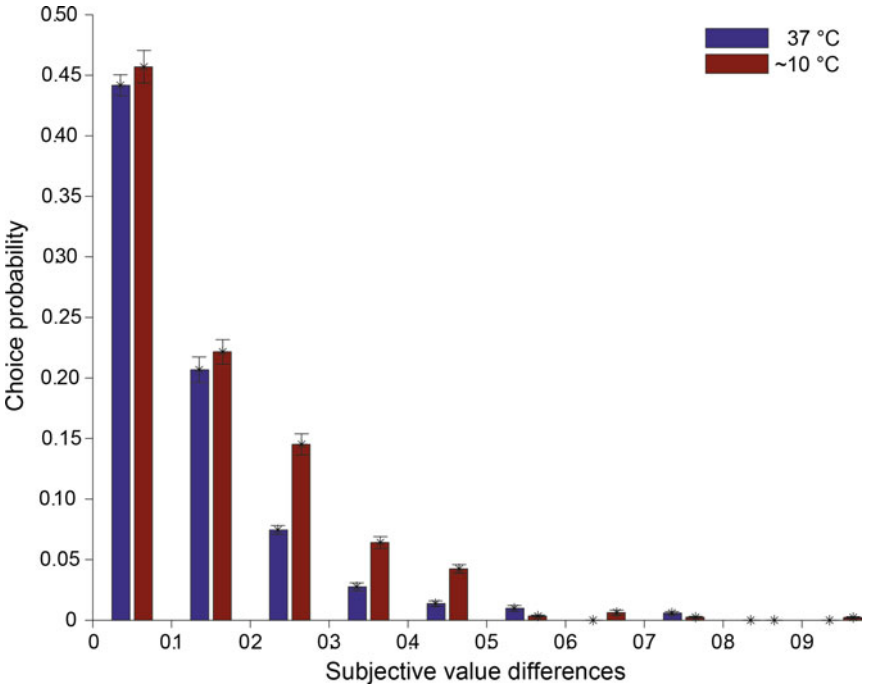


Fig. 6.5 Behavioral effects of bilateral inactivation of SEF through cooling. The probability of choosing the smaller of two options was plotted as a function of relative difference in subjective value of the options as determined behaviorally (see Fig. 6.1). The behavior under normal conditions is shown by the blue bars (normal temperature; 37°C). The behavior, when SEF is inactivated in both hemispheres, is shown by the red bars (~10°C)

neuron within a circuit, ultimately again within the entire brain, a 'functional' connectome [119]. Such ideas are very ambitious and attractive, since they promise to tackle directly one of the biggest problems in our current understanding of the brain. While we can observe the spiking activity of individual neurons, and can establish links between their activity and behavior or even mental states, we mostly do not know why the neurons show the activity pattern that we observe. The activity of neurons in the brain is ultimately an emergent property of the interactions between the different elements that make up the circuit that they belong to. Thus, only knowledge about the fine structure of this circuit and the distributed activity of the various elements that make up the circuit will provide us with a true mechanistic understanding of the brain.

However, there is reason to be cautious. It is easy to underestimate the true complexity of the brain [2, 120]. Moore proved that the number of steps necessary to learn about the internal structure of a computing machine is at least the same order of magnitude as the number of states of the machine [39]. It is easy to see the consequences of this relationship for a task such as fully characterizing the visual cortex of the mouse, which contains about 2 million neurons [120]. Any realistic hope

of progress relies on our ability to discover hierarchical structures in the network, which would allow us to simplify the level of complexity of the circuit that needs to be understood in order to explain the behavior of the entire network [120, 121]. Such insights ultimately require further conceptual breakthroughs, and the input of theorists, such as mathematical psychologists or computational neuroscientists.

In conclusion, we are living in exciting times for neuroscientists. Important technological breakthroughs have been made and there is the potential for the development of even more advanced methods for recording neural activity from hundreds, if not thousands, neurons simultaneously and to reconstruct nervous circuits in unprecedented detail [122, 123]. Without any doubt, these attempts at technical innovation will move neuroscience forward. This is true, not least, because work towards achieving these goals might lead to much-needed improvements in measurement technology, even if the ultimate goal should remain elusive. However, it is important not to ignore the main source of most of the real insights into the brain that have been acquired up to now, namely, the establishment of a functional and explanatory link between neural activity and mental phenomena using psychophysics and mathematical models. When the newly available techniques are combined with these established approaches, we will truly see great steps forward in our understanding of the brain.

Exercises

1. Given the fact that, in humans, for the foreseeable future we will have to rely on mass-activity measures of brain activity, the relationship between single-unit activity and mass-action recordings is of interest. What do you think is the relationship between individual neurons producing action potentials and fMRI or EEG?
2. New generations of neuroprobes will allow extreme miniaturization. Currently available probes allow the construction of devices with 456 electrodes. Within a few years we will likely have neuroprobes available that have up to 2000 electrodes [123]. Also currently already available are microchips that can be injected into the brain and that allow the recording of electrical fields, temperature, and the local emission of light, which would allow the spatially precise control of neural activity through optogenetics [124]. More futuristic approaches envision "nanometer-scale light-sensitive devices that could embed themselves in the membrane of neurons, power themselves from cellular energy, and wirelessly convey the activity of millions of neurons simultaneously" [123, 125]. If we assume for a moment that all these technical advances come to fruition, what are questions that could be answered using these new techniques?

Further Reading

1. Wise [32] provides a provocative review about the function of prefrontal cortex. It includes a discussion of the differences between the frontal cortex of primates and rodents.
2. In a seminal paper, Teller [1] lays out the internal logic that allows establishing a link between a set of physical events in the brain to psychological events or a functional concept. A more in depth discussion of these requirements is also provided by Parker & Newsome [3] in the sensory domain and by Schall [2] in the motor domain.
3. In a highly amusing and interesting book, Passingham [50] describes what is known about the specific characteristics of human brain anatomy and physiology, as opposed to the brains of other primates.
4. Glimcher [85] lays out, how psychology, economics, and neuroscience can be related in a new reductionist framework.
5. In this classic description, Gescheider [82] gives an overview over modern psychophysics. This topic is of utmost importance to anyone interested in connecting brain activity with behavior and mental states.
6. Modern approaches for the recording of multiple individual neurons are discussed by Kipke et al. [89] and Buzsaki [94]. An interesting new method for analyzing the resulting multi-neuron data is described in Yu, et al. [116]. These papers are of course just highlights out of a vast literature.
7. Denk et al. [30] and Alivisatos et al. [119] describe ambitious new proposals of obtaining structural and functional 'connectomes', that is, a complete description of a neuronal circuit.
8. Koch [120] discusses the difficulties for explaining the mammalian brain related to its astonishing complexity.

Acknowledgements We are grateful to K. Nielsen, D. Sasikumar and E. Emeric for comments on the manuscript. This work was supported by the National Eye Institute through grant R01-EY019039 to VS.

References

1. Teller DY (1984) Linking propositions. Vision Res 24(10):1233–1246
2. Schall JD (2004) On building a bridge between brain and behavior. Ann Rev Psychol 55:23–50
3. Parker AJ, Newsome WT (1998) Sense and the single neuron: probing the physiology of perception. Ann Rev Neurosci 21:227–277
4. Adrian ED (1928) The basis of sensation: the action of the sense organs. W. W. Norton, New York
5. Hartline HK, Milne LJ, Wagman IH (1947) Fluctuation of response of single visual sense cells. Fed Proc 6(1 Pt 2):124
6. Barlow HB (1995) The neuron doctrine in perception. In: Gazzaniga MS (ed) The cognitive neurosciences. MIT Press, Cambridge, pp 415–435

7. Mountcastle VB, Talbot WH, Darian-Smith I, Kornhuber HH (1967) Neural basis of the sense of flutter-vibration. Science 155(762):597–600
8. Rieke F, Warland DK, de Ruyter van Steveninck R, Bialek W (1997) Spikes: exploring the neural code. MIT Press, Cambridge
9. Koch C (1999) Biophysics of computation: information processing in single neurons. In: Computational Neuroscience Series, Stryker M (ed) Oxford University Press, Oxford
10. Silver RA (2010) Neuronal arithmetic. Nat Rev Neurosci 11(7):474–489
11. Softky WR, Koch C (1993) The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. J Neurosci 13(1):334–350
12. Shadlen MN, Newsome WT (1998) The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. J Neurosci 18(10):3870–3896
13. Singer W, Gray CM (1995) Visual feature integration and the temporal correlation hypothesis. Ann Rev Neurosci 18:555–586
14. London M, Roth A, Beeren L, Hausser M, Latham PE (2010) Sensitivity to perturbations in vivo implies high noise and suggests rate coding in cortex. Nature 466(7302):123–127
15. Werner G, Mountcastle VB (1965) Neural activity in mechanoreceptive cutaneous afferents: stimulus-response relations, weber functions, and information transmission. J Neurophysiol 28:359–397.
16. de Lafuente V Romo R (2005) Neuronal correlates of subjective sensory experience. Nature Neuroscience. [Research Support, Non-U.S. Gov't]. 8(12):1698–1703
17. Newsome WT, Britten KH, Salzman CD, Movshon JA (1990) Neuronal mechanisms of motion perception. Cold Spring Harb Symp Quant Biol 55:697–705
18. Shadlen MN, Newsome WT (2001) Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. J Neurophysiol 86(4):1916–1936
19. Vallbo AB, Johansson RS (1984) Properties of cutaneous mechanoreceptors in the human hand related to touch sensation. Hum Neurobiol 3(1):3–14
20. Houweling AR, Brecht M (2008) Behavioural report of single neuron stimulation in somatosensory cortex. Nature 451(7174):65–68
21. Li CY, Poo MM, Dan Y.(2009) Burst spiking of a single cortical neuron modifies global brain state. Science 324(5927):643–646
22. Crist RE, Lebedev MA (2007) Multielectrode recording in behaving monkeys. In: Nicolelis MAL (ed) Methods for neural ensemble recordings. CRC, Boca Raton
23. Evarts EV (1966) Pyramidal tract activity associated with a conditioned hand movement in the monkey. J Neurophysiol 29(6):1011–1027
24. Evarts EV (1968) Relation of pyramidal tract activity to force exerted during voluntary movement. J Neurophysiol 31(1):14–27
25. Mountcastle VB, Talbot WH, Sakata H, Hyvarinen J (1969) Cortical neuronal mechanisms in flutter-vibration studied in unanesthetized monkeys. Neuronal periodicity and frequency discrimination. J Neurophysiol 32(3):452–484
26. Wurtz RH (1968) Visual cortex neurons: response to stimuli during rapid eye movements. Science 162(858):1148–1150
27. Kepecs A, Uchida N, Zariwala HA, Mainen ZF (2008) Neural correlates, computation and behavioural impact of decision confidence. Nature 455(7210):227–231
28. Ogawa M, van der Meer MA, Esber GR, Cerri DH, Stalnaker TA, Schoenbaum G (2013) Risk-responsive orbitofrontal neurons track acquired salience. Neuron 77(2):251–258
29. Ohki K, Chung S, Kara P, Hubener M, Bonhoeffer T, Reid RC (2006) Highly ordered arrangement of single neurons in orientation pinwheels. Nature 442(7105):925–958
30. Denk W, Briggman KL, Helmstaedter M.(2012) Structural neurobiology: missing link to a mechanistic understanding of neural computation. Nat Rev Neurosci.13(5):351–358
31. Boyden ES, Zhang F, Bamberg E, Nagel G, Deisseroth K (2005) Millisecond-timescale, genetically targeted optical control of neural activity. Nat Neurosci 8(9):1263–1268
32. Wise SP (2008) Forward frontal fields: phylogeny and fundamental function. Trends Neurosci 31(12):599–608

33. Tanji J, Hoshi E (2008) Role of the lateral prefrontal cortex in executive behavioral control. Physiol Rev 88(1):37–57
34. Fuster JM (2008) The prefrontal cortex, 4th edn. Academic Press, Amsterdam
35. Uylings HB, Groenewegen HJ, Kolb B (2003) Do rats have a prefrontal cortex? Behav Brain Res 146(1–2):3–17
36. Sescousse G, Redoute J, Dreher JC (2010) The architecture of reward value coding in the human orbitofrontal cortex. J Neurosci 30(39):13095–13104
37. O'Neill M, Schultz W (2010) Coding of reward risk by orbitofrontal neurons is mostly distinct from coding of reward value. Neuron 68(4):789–800
38. Tobler PN, O'Doherty JP, Dolan RJ, Schultz W (2007) Reward value coding distinct from risk attitude-related uncertainty coding in human reward systems. J Neurophysiol 97(2):1621–1632
39. Moore EF (1956) Gedanken-experiments on sequential machines. In: Shannon CE, McCarthy J (eds) Automata studies. Princeton University Press, Princeton pp. 129–153
40. Ridley RM, Haystead TA, Baker HF (1981 Mar) An analysis of visual object reversal learning in the marmoset after amphetamine and haloperidol. Pharmacol Biochem Behav 14(3):345–351
41. Kruzich PJ, Grandy DK (2004) Dopamine D2 receptors mediate two-odor discrimination and reversal learning in C57BL/6 mice. BMC Neurosci 5:12
42. Mehta MA, Swainson R, Ogilvie AD, Sahakian BJ, Robbins TW (2001) Improved short-term spatial memory but impaired following the dopamine D-2 agonist bromocriptine reversal learning in human volunteers. Psychopharmacology 159(1):10–20.
43. Atasoy D, Betley JN, Su HH, Sternson SM (2012) Deconstruction of a neural circuit for hunger. Nature 488(7410):172–177
44. Plassmann H, O'Doherty J, Shiv B, Rangel A (2008) Marketing actions can modulate neural representations of experienced pleasantness. Proc Natl Acad Sci U S A 105(3):1050–1054
45. Cavanaugh J, Monosov IE, McAlonan K, Berman R, Smith MK, Cao V et al (2012) Optogenetic inactivation modifies monkey visuomotor behavior. Neuron 76(5):901–907
46. Nauhaus I, Nielsen KJ, Disney AA, Callaway EM (2012) Orthogonal micro-organization of orientation and spatial frequency in primate primary visual cortex. Nat Neurosci 15(12):1683–1690
47. Han X, Qian X, Bernstein JG, Zhou HH, Franzesi GT, Stern P et al (2009) Millisecond-timescale optical control of neural dynamics in the nonhuman primate brain. Neuron 62(2):191–198
48. Ozden I, Wang J, Lu Y, May T, Lee J, Goo W et al (2013) A coaxial optrode as multifunction write-read probe for optogenetic studies in non-human primates. J Neurosci Methods 219:142–154
49. Diester I, Kaufman MT, Mogri M, Pashaie R, Goo W, Yizhar O et al (2011) An optogenetic toolbox designed for primates. Nat Neurosci 14(3):387–397
50. Passingham R (2008) What is special about the human brain? Oxford University Press, Oxford
51. Suthana N, Fried I (2012 Aug) Percepts to recollections: insights from single neuron recordings in the human brain. Trends Cogn Sci 16(8):427–436
52. Tankus A, Fried I, Shoham S (2012) Structured neuronal encoding and decoding of human speech features. Nat Commun 3:1015
53. Quiroga RQ, Reddy L, Kreiman G, Koch C, Fried I (2005) Invariant visual representation by single neurons in the human brain. Nature 435(7045):1102–1107
54. Cerf M, Thiruvengadam N, Mormann F, Kraskov A, Quiroga RQ, Koch C et al (2010) On-line, voluntary control of human temporal lobe neurons. Nature 467(7319):1104–1108
55. Burbaud P, Clair AH, Langbour N, Fernandez-Vidal S, Goillandeau M, Michelet T et al (2013) Neuronal activity correlated with checking behaviour in the subthalamic nucleus of patients with obsessive-compulsive disorder. Brain 136(Pt 1):304–317
56. Gold JI, Shadlen MN (2007) The neural basis of decision making. Ann rev neurosci 30:535–574

57. Rangel A, Camerer C, Montague PR (2008) A framework for studying the neurobiology of value-based decision making. Nat Rev Neurosci 9(7):545–556
58. Balleine BW, Dickinson A (1998) Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. Neuropharmacology 37(4–5):407–419
59. Yin HH, Knowlton BJ (2006) The role of the basal ganglia in habit formation. Nat rev Neurosci 7(6):464–476
60. Daw ND, Niv Y, Dayan P (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci 8(12):1704–1711
61. Vickery TJ, Chun MM, Lee D (2011) Ubiquity and specificity of reinforcement signals throughout the human brain. Neuron 72(1):166–177
62. Padoa-Schioppa C (2011) Neurobiology of economic choice: a good-based model. Annu Rev Neurosci 34:333–359
63. Kable JW, Glimcher PW (2009) The neurobiology of decision: consensus and controversy. Neuron 63(6):733–745
64. Plassmann H, O'Doherty J, Rangel A (2007) Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. J Neurosci 27(37):9984–9988
65. Plassmann H, O'Doherty JP, Rangel A (2010) Appetitive and aversive goal values are encoded in the medial orbitofrontal cortex at the time of decision making. J Neurosci 30(32):10799–10808
66. Bermudez MA, Schultz W (2010) Reward magnitude coding in primate amygdala neurons. J Neurophysiol 104(6):3424–3432
67. Grabenhorst F, Hernadi I, Schultz W (2012) Prediction of economic choice by primate amygdala neurons. Proc Natl Acad Sci U S A 109(46):18950–18955
68. Lau B, Glimcher PW (2007) Action and outcome encoding in the primate caudate nucleus. J Neurosci 27(52):14502–14514
69. Samejima K, Ueda Y, Doya K, Kimura M (2005) Representation of action-specific reward values in the striatum. Science 310(5752):1337–1340
70. Kim S, Hwang J, Lee D (2008) Prefrontal coding of temporally discounted values during intertemporal choice. Neuron 59(1):161–172
71. Hernandez A, Nacher V, Luna R, Zainos A, Lemus L, Alvarez M et al (2010) Decoding a perceptual decision process across cortex. Neuron 66(2):300–314
72. Seo H, Barraclough DJ, Lee D (2007) Dynamic signals related to choices and outcomes in the dorsolateral prefrontal cortex. Cereb Cortex 17(suppl 1):i110–i117
73. So NY, Stuphorn V (2010) Supplementary eye field encodes option and action value for saccades with variable reward. J Neurophysiol 104(5):2634–2653
74. Cisek P (2012) Making decisions through a distributed consensus. Curr Opin Neurobiol 22(6):927–936
75. Cisek P, Kalaska JF (2010) Neural mechanisms for interacting with a world full of action choices. Ann Rev Neurosci 33:269–298
76. Platt ML, Glimcher PW (1999) Neural correlates of decision variables in parietal cortex. Nature 400(6741):233–238
77. Sugrue LP, Corrado GS, Newsome WT (2004) Matching behavior and the representation of value in the parietal cortex. Science 304(5678):1782–1787
78. Shadlen MN, Kiani R, Hanks TD, Churchland AK (2008) Neurobiology of decision making, an intentional framework. In: Engel C, Singer W (eds) Better than conscious? MIT Press, Cambridge, pp. 71–101
79. Kingdom FAA, Prins N (2010) Psychophysics: a practical introduction. Academic Press, Amsterdam
80. Maloney LT, Yang JN (2003) Maximum likelihood difference scaling. J Vis 3(8):573–585
81. Fechner GT (1876) Vorschule der Aesthetik. Breitkopf & Haerterl, Leibzig
82. Gescheider GA (1997) Psychophysics: the fundamentals. 3rd edn. Lawrence Erlbaum Assoc., Mahwah
83. Thurstone LL (1927) A law of comparative judgment. Psychol Rev 34:273–286

84. McFadden D (1974) Conditional logit analysis of qualitative choice behavior. In: Zarembka P (ed) Frontier in econometrics. Academic Press, New York, pp. 105–142
85. Glimcher P (2011) Foundations of neuroeconomic analysis. Oxford University Press, Oxford
86. Stevens SS (1951) Handbook of experimental psychology. Wiley, New York
87. Einevoll GT, Franke F, Hagen E, Pouzat C, Harris KD (2012) Towards reliable spike-train recordings from thousands of neurons with multielectrodes. Curr Opin Neurobiol 22(1):11–17
88. Hernandez A, Nacher V, Luna R, Alvarez M, Zainos A, Cordero S et al (2008) Procedure for recording the simultaneous activity of single neurons distributed across cortical areas during sensory discrimination. Proc Natl Acad Sci U S A 105(43):16785–16790
89. Kipke DR, Shain W, Buzsaki G, Fetz E, Henderson JM, Hetke JF et al (2008) Advanced neurotechnologies for chronic neural interfaces: new horizons and clinical opportunities. J Neurosci 28(46):11830–11838
90. Du J, Riedel-Kruse IH, Nawroth JC, Roukes ML, Laurent G, Masmanidis SC (2009) High-resolution three-dimensional extracellular recording of neuronal activity with microfabricated electrode arrays. J Neurophysiol 101(3):1671–1678
91. Kelly RC, Smith MA, Samonds JM, Kohn A, Bonds AB, Movshon JA et al (2007) Comparison of recordings from microelectrode arrays and single electrodes in the visual cortex. J Neurosci 27(2):261–264
92. Nicolelis MA, Dimitrov D, Carmena JM, Crist R, Lehew G, Kralik JD et al (2003) Chronic, multisite, multielectrode recordings in macaque monkeys. Proc Natl Acad Sci U S A 100(19):11041–11046
93. McNaughton BL, O'Keefe J, Barnes CA (1983) The stereotrode: a new technique for simultaneous isolation of several single units in the central nervous system from multiple unit records. J Neurosci Methods 8(4):391–397
94. Buzsaki G (2004) Large-scale recording of neuronal ensembles. Nat Neurosci 7(5):446–451
95. Miller EK, Wilson MA (2008) All my circuits: using multiple electrodes to understand functioning neural networks. Neuron 60(3):483–488
96. Fujisawa S, Amarasingham A, Harrison MT, Buzsaki G (2008) Behavior-dependent short-term assembly dynamics in the medial prefrontal cortex. Nat neurosci 11(7):823–833
97. Seymour JP, Kipke DR (2007) Neural probe design for reduced tissue encapsulation in CNS. Biomaterials 28(25):3594–3607
98. Salazar RF, Dotson NM, Bressler SL, Gray CM (2012) Content-specific fronto-parietal synchronization during visual working memory. Science 338(6110):1097–1100
99. Hoffman KL, McNaughton BL (2002) Coordinated reactivation of distributed memory traces in primate neocortex. Science 297(5589):2070–2073
100. Wilson MA, McNaughton BL (1994) Reactivation of hippocampal ensemble memories during sleep. Science 265(5172):676–679
101. Cisek P (2012) Making decisions through a distributed consensus. Curr Opin Neurobiol 22(6):927–936
102. Huerta MF, Kaas JH (1990) Supplementary eye field as defined by intracortical microstimulation: connections in macaques. J Comp Neurol 293:299–330
103. Ghashghaei HT, Hilgetag CC, Barbas H (2007 Feb 1) Sequence of information processing for emotions based on the anatomic dialogue between prefrontal cortex and amygdala. Neuroimage 34(3):905–923
104. Lau B, Glimcher PW (2008) Value representations in the primate striatum during matching behavior. Neuron 58(3):451–463
105. Shook BL, Schlag-Rey M, Schlag J (1991) Primate supplementary eye field. II. Comparative aspects of connections with the thalamus, corpus striatum, and related forebrain nuclei. J Comp Neurol 307(4):562–583
106. Coe B, Tomihara K, Matsuzawa M, Hikosaka O (2002) Visual and anticipatory bias in three cortical eye fields of the monkey during an adaptive decision-making task. J Neurosci 22(12):5081–5090
107. So NY, Stuphorn V (2010) Supplementary eye field encodes option and action value for saccades with variable reward. J Neurophysiol 104(5):2634–2653

108. So NY, Stuphorn V (2012) Supplementary eye field encodes reward prediction error. J Neurosci 32(9):2950–2963
109. Barlow HB, Levick WR, Yoon M (1971) Responses to single quanta of light in retinal ganglion cells of the cat. Vision Res Suppl 3:87–101
110. Thompson KG, Hanes DP, Bichot NP, Schall JD (1996) Perceptual and motor processing stages identified in the activity of macaque frontal eye field neurons during visual search. J Neurophysiol 76(6):4040–4055
111. Churchland MM, Yu BM, Sahani M, Shenoy KV (2007) Techniques for extracting single-trial activity patterns from large-scale neural recordings. Curr Opin Neurobiol 17(5):609–618
112. Duda RO, Hart PE, Stork DG (2000) Pattern classification, 2nd edn. Wiley, New York
113. Broome BM, Jayaraman V, Laurent G (2006) Encoding and decoding of overlapping odor sequences. Neuron 51(4):467–482
114. Briggman KL, Abarbanel HD, Kristan WB, Jr (2005) Optical imaging of neuronal populations during decision-making. Science 307(5711):896–901
115. Harvey CD, Coen P, Tank DW (2012) Choice-specific sequences in parietal cortex during a virtual-navigation decision task. Nature 484(7392):62–68
116. Yu BM, Cunningham JP, Santhanam G, Ryu SI, Shenoy KV, Sahani M (2009) Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. J Neurophysiol 102(1):614–635
117. Lomber SG, Payne BR, Horel JA (1999) The cryoloop: an adaptable reversible cooling deactivation method for behavioral or electrophysiological assessment of neural function. J Neurosci Methods 86(2):179–194
118. Noonan MP, Walton ME, Behrens TE, Sallet J, Buckley MJ, Rushworth MF (2010) Separate value comparison and learning mechanisms in macaque medial and lateral orbitofrontal cortex. P Natl Acad Sci USA 107(47):20547–20552
119. Alivisatos AP, Chun M, Church GM, Greenspan RJ, Roukes ML, Yuste R (2012)The brain activity map project and the challenge of functional connectomics. Neuron 74(6):970–974.
120. Koch C (2012) Systems biology. Modular biological complexity. Science 337(6094):531–532
121. Mountcastle VB (1997) The columnar organization of the neocortex. Brain 120(Pt 4):701–722
122. Kim TI, McCall JG, Jung YH, Huang X, Siuda ER, Li Y et al (2013) Injectable, cellular-scale optoelectronics with applications for wireless optogenetics. Science 340(6129):211–216
123. Abbott A (2013) Neuroscience: solving the brain. Nature 499(7458):272–274
124. Kim TI et al (2013) Injectable, cellular-scale optoelectronics with applications for wireless optogenetics. Science 340:211–216
125. Alivisatos AP et al (2013) Nanotools for neuroscience and brain activity mapping. ACS Nano 7:1850–1866

Chapter 7
Model-Based Cognitive Neuroscience: A Conceptual Introduction

Birte U. Forstmann and Eric-Jan Wagenmakers

Abstract This tutorial chapter shows how the separate fields of mathematical psychology and cognitive neuroscience can interact to their mutual benefit. Historically, the field of mathematical psychology is mostly concerned with formal theories of behavior, whereas cognitive neuroscience is mostly concerned with empirical measurements of brain activity. Despite these superficial differences in method, the ultimate goal of both disciplines is the same: to understand the workings of human cognition. In recognition of this common purpose, mathematical psychologists have recently started to apply their models in cognitive neuroscience, and cognitive neuroscientists have borrowed and extended key ideas that originated from mathematical psychology. This chapter consists of three main sections: the first describes the field of mathematical psychology, the second describes the field of cognitive neuroscience, and the third describes their recent combination: model-based cognitive neuroscience.

7.1 Introduction

The griffin is a creature with the body of a lion and the head and wings of an eagle. This mythical hybrid is thought to symbolize the rule over two empires, one on the earth (the lion part) and the other in the skies (the eagle part). The preceding six tutorial chapters may have given the impression that the field of model-based cognitive neuroscience is similar to a griffin in that it represents the union of two fundamentally incompatible disciplines. After all, the methods and concepts from

B. U. Forstmann (✉)
Cognitive Science Center Amsterdam, University of Amsterdam, Nieuwe Achtergracht 129, 1018 WS, Amsterdam, The Netherlands
e-mail: buforstmann@gmail.com

Department of Psychology, University of Amsterdam, Nieuwe Prinsengracht 129 1018 VZ, Amsterdam, The Netherlands

E.-J. Wagenmakers
University of Amsterdam, Department of Psychological Methods, Weesperplein 4, 1018 XA,Amsterdam, The Netherlands
e-mail: EJ.Wagenmakers@gmail.com

B. U. Forstmann, E.-J. Wagenmakers (eds.), *An Introduction to Model-Based Cognitive Neuroscience,* DOI 10.1007/978-1-4939-2236-9_7

Fig. 7.1 The griffin—part lion, part eagle—as depicted in Jonston (1660); copper engraving by Matthius Merian

the field of formal modeling, explained in Chaps. 1, 2, and 3, appear to have little in common with the methods and concepts from the field of cognitive neuroscience as discussed in Chaps. 4, 5, and 6. The goal of this tutorial chapter is to explain that this impression is mistaken—the griffin analogy is apt because it highlights the added possibilities and novel insights that can be obtained when formal models for behavior are combined with methods from cognitive neuroscience ([1; Fig. 7.1).

In this chapter we explain why it is natural to combine behavioral modeling with cognitive neuroscience; furthermore, we illustrate the benefits of the symbiotic relationship between the two disciplines by means of concrete examples. However, before we discuss our model-neuroscience griffin in detail, it is informative to first discuss its component disciplines separately.

7.2 Mathematical Psychology

Mathematical psychologists are concerned with the formal analysis of human behavior. Objects of study include perception, decision-making, learning, memory, attention, categorization, preference judgments, and emotion. Whenever researchers propose, extend, or test formal models of human behavior they are practising mathematical psychology. Thus, the field of mathematical psychology is relatively broad, and defined more by method than by topic or subject matter. To give you an impression of the work done by mathematical psychologists, Table 7.1 provides an overview of the articles published in the June 2012 issue of the *Journal of Mathematical Psychology*.

The inner core of card-carrying mathematical psychologists is comprised of only about a few hundred researchers, and consequently progress in the field can be agonizingly slow. In his 2008 editorial in the *Journal of Mathematical Psychology*, the society's president Jim Townsend wrote:

> It can prove a frustrating experience to compare psychology's pace of advance with progress in the 'hard' sciences. [...] steps in filling in data about a phenomenon not to mention testing of major theoretical issues and models, seem to occur with all the urgency of a glacier. One may wait years, before a modeler picks up the scent of an intriguing theoretical problem and carries it ahead. It is disheartening to contrast our situation with, say, that of microbiology. [9, p. 270]

Table 7.1 Articles published in the June 2012 issue of the *Journal of Mathematical Psychology*

Title	Reference
A tutorial on the Bayesian approach for analyzing structural equation models	[2]
Symmetry axiom of Haken-Kelso-Bunz coordination dynamics revisited in the context of cognitive activity	[3]
Quantum-like generalization of the Bayesian updating scheme for objective and subjective mental uncertainties	[4]
Torgerson's conjecture and Luce's magnitude production representation imply an empirically false property	[5]
A predictive approach to nonparametric inference for adaptive sequential sampling of psychophysical experiments	[6]
On a signal detection approach to m-alternative forced choice with bias, with maximum likelihood and Bayesian approaches to estimation	[7]
How to measure post-error slowing: A confound and a simple solution	[8]

One solution to this glacier-like progress is for mathematical psychologists to collaborate with researchers from other disciplines; when more researchers are interested in a particular phenomenon this greatly increases the speed with which new discoveries are made. This is in fact exactly what happened when cognitive neuroscientists became interested in quantitative models for speeded decision making (e.g., [10–12]; prior to this development, such models were proposed, adjusted, and tested only by a handful of mathematical psychologists—for example, from 1978 to 2001 Roger Ratcliff stood alone in his persistent efforts to promote the *drift diffusion model* as a comprehensive account of human performance in speeded two-choice tasks.

7.2.1 The Drift Diffusion Model

In the drift diffusion model (DDM), shown in Fig. 7.2, noisy information is accumulated over time until a decision threshold is reached and a response is initiated. The DDM provides a formal account of how people make speeded decisions between two choice alternatives. In other words, the model yields parameter estimates (e.g., for drift rate and boundary separation) that represent specific psychological processes (e.g., ability and response caution) in order to account for error rates as well as response time distributions for both correct choices and errors. Put differently, the DDM takes observed behavior—which may be difficult to interpret—and decomposes it into psychological processes that are easier to interpret. For instance, the boundary separation parameter in the DDM reflects the amount of information that a participant seeks to accumulate before being confident enough to respond. Higher levels of boundary separation reflect a more cautious response regime, one in which responding is slow but errors are few.

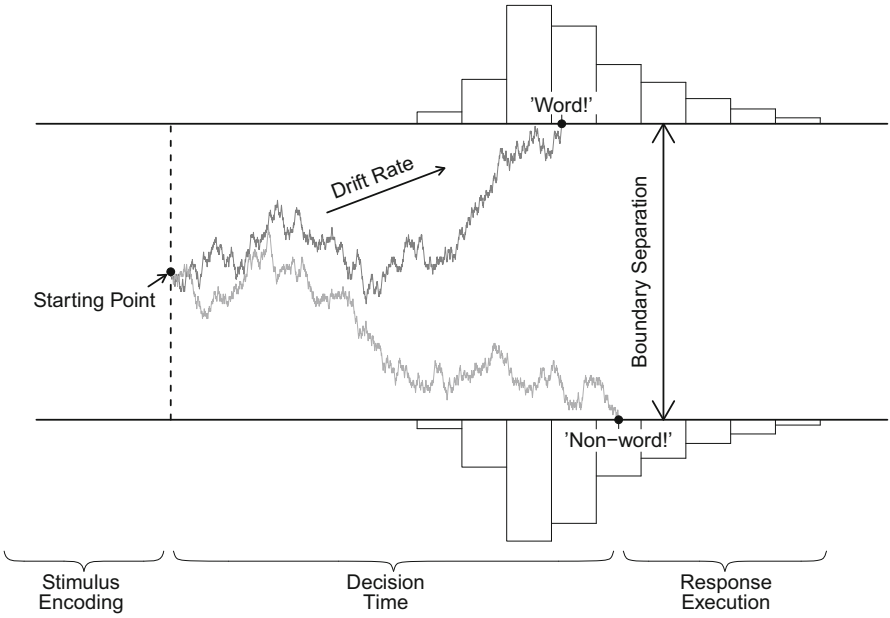


Fig. 7.2 A drift diffusion model for the lexical decision task. In this task, the participant is shown a letter string and has to decide quickly whether it is an existing word (e.g., *tiger*) or not (e.g., *drapa*). Noisy evidence is accumulated over time until a boundary is reached and the corresponding response is initiated. Drift rate quantifies decision difficulty and boundary separation quantifies response caution. Predicted response time equals the decision time plus the time required for non-decision processes such as stimulus encoding and response execution. (Figure as originally published in [13])

Throughout the years, Ratcliff repeatedly demonstrated how the DDM allows for deeper insight in the processes that underlie observed behavior (e.g., [14]). Consider, for instance, the finding that older adults respond more slowly than younger adults, a general empirical regularity that holds even in relatively simple tasks such as lexical decision. The once-dominant explanation of this age-related slowing holds that older adults have a reduced rate of information processing, perhaps as a result of neural degradation; hence, the age-related slowing was assumed to hold generally, across a wide range of different tasks and processes [15–17]. However, when the DDM was applied to the data from older adults, Ratcliff and colleagues discovered something surprising [18, 19]: in most speeded two-choice tasks, drift rates did *not* differ between the young and the old. That is, older adults were accumulating diagnostic information as efficiently as the young. Instead, the age-related slowdown was usually due to a combination of two factors: (1) an increase in non-decision time, that is, the time needed for encoding and response execution, and (2) an increase in response caution. These results suggest that the age-related slowing can be undone, at least in part, by encouraging the elderly to adopt a more risky response strategy (for a confirmation of this prediction see for instance [20]).

Currently, the DDM can be considered one of the most successful quantitative models in mathematical psychology: not only does it provide fits to empirical data that are consistently good, it has also driven theoretical progress in fields traditionally dominated by verbal or quasi-formal accounts. These intuitive accounts were often unable to withstand quantitative scrutiny (e.g., [21]).

The main weakness of the DDM is that it provides a decomposition of performance that is relatively abstract, that is, the DDM does not commit to any representational assumptions. This makes the model less interesting from a psychological point of view. The main weakness of the DDM, however, is also its main strength: because its account is relatively abstract it can be applied to a wide range of different tasks and paradigms.

For the first 25 years, the development and application of the DDM was guided by statistical and pragmatic considerations; Of particular relevance here is that the dynamics of decision-making in neural circuits is remarkably similar to that postulated by the DDM (e.g., [12]) in that neurons appear to accumulate noisy evidence until threshold. Thus, the DDM does not only capture behavioral data but holds the promise to capture underlying neural dynamics as well. This may not be accidental: the DDM describes performance of a decision-maker who is statistically optimal in the sense of minimizing mean response time for a fixed level of accuracy (e.g., [22]) and it is plausible that for simple perceptual tasks, evolution and individual learning has curtailed those neural dynamics that lead to suboptimal outcomes. Cognitive neuroscientists have not only applied the DDM to neural data, they have also proposed theoretical extensions to the model. For instance, high-profile extensions concern the generalization to more than two choice-alternatives [23, 24], collapsing bounds [25], urgency-gating [26], and drift rates that change during stimulus processing [27].

7.2.2 Ambivalence Towards Neuroscience

Although mathematical psychologists are increasingly interested in the neural underpinnings of cognition, the overall attitude towards the neurosciences is one of ambivalence or even open distrust.[1] Some of this ambivalence stems from the concern that brain measurements alone may not be theoretically meaningful. For instance, Coltheart claimed that "no functional neuroimaging research to date has yielded data that can be used to distinguish between competing psychological theories" [28, p. 323] (see the exercise at the end of this chapter).

To demonstrate the limitations of neuroscientific methods, Ulrich presented the following thought experiment [29]. Suppose you are intrigued by the ability of a computer program to provide analytic solutions to integrals. In the Maple program, for instance, you can enter the integral $\int x \sin x \, dx$ as `int(x*sin(x),x);` and Maple will immediately return the solution: `sin(x)-x*cos(x)`. How can we learn more about how Maple accomplishes these and other computations?

[1] At the 2009 annual meeting of the *Society for Mathematical Psychology*, one of the plenary speakers discussed some of his beginning exploits in cognitive neuroscience. Following his talk, the first question from the audience was whether he had now "joined the dark force".

Ulrich argues that neuroscientists may tackle this problem in different ways, as illustrated in Fig. 7.3: analogous to functional brain imaging, one might perform a heat scan on the laptop as it computes integrals, and compare this with a control condition where it is just waiting for input (Fig. 7.3, top left panel). Analogous to EEG measurements, one could attach surface electrodes to the laptop, have the laptop repeatedly perform integrals, and compute a stimulus-locked or a response-locked event-related potential (Fig. 7.3, top right panel). Analogous to single-cell recordings in monkeys, one might implant electrodes and register the activity of small components within the laptop (Fig. 7.3, lower left panel). Finally, analogous to neurosurgical methods, one might lesion the laptop, for instance by hitting it with a hammer. With luck, one might even discover a double dissociation, that is, lesioning one part of the laptop harms the computation of integrals but does not harm word processing, whereas lesioning another part of the laptop harms word processing but not the computation of integrals (Fig. 7.3, lower right panel).

Ulrich ([29, p. 29]) concludes that "(...) none of these fancy neuroscience techniques can directly unravel the hidden mechanisms of this symbolic math program" and hence, brain measurement techniques alone cannot replace formal theories of cognition. We suspect that most mathematical psychologists subscribe to the Ulrich laptop metaphor of neuroscience. The laptop metaphor is insightful and thought-provoking, but it should not be misinterpreted to mean that neuroscientific methods are by definition uninformative. For example, consider a race of aliens who discover a refrigerator and wish to learn how it works. They may first conduct behavioral experiments and conclude that parts of the refrigerator are cooler then others. They may study the speed of cooling in the different compartments, and its relation to a host of relevant factors (e.g., the extent to which the refrigerator door is left open, the temperature of various products just before they are put inside, and the volume occupied by the products). The aliens may develop theoretical concepts such as homeostasis, they may propose sets of axioms, and they may develop competing formal models about how the refrigerator does what it does. Unfortunately, the behavioral data are rather sparse and therefore they will fail to falsify many of the more complicated theories. It is evident that "neuroscientific" measures of studying the refrigerator (e.g., examining its underlying circuitry) will yield additional insights that can be used either to adjudicate between the competing theories or to develop new theories that are more appropriate.

We will leave it up to philosophers to decide whether the study of cognition is similar to a laptop or to a refrigerator. This decision may well depend on the cognitive phenomenon under study. For instance, perceptual illusions are perhaps best understood by taking into account the neural processes that subserve perception, whereas a different approach is warranted when one wants to understand loss-aversion in gambling.

Pragmatically, any approach is worthwhile as long as it yields theoretical progress, that is, a deeper understanding of human cognition. It is undeniably true that brain measurements, just as response times or error rates, constitute data that are potentially informative about the underlying cognitive process. The main difficulty, therefore, is to develop formal models that allow the brain measurements to make contact with putative cognitive processes.

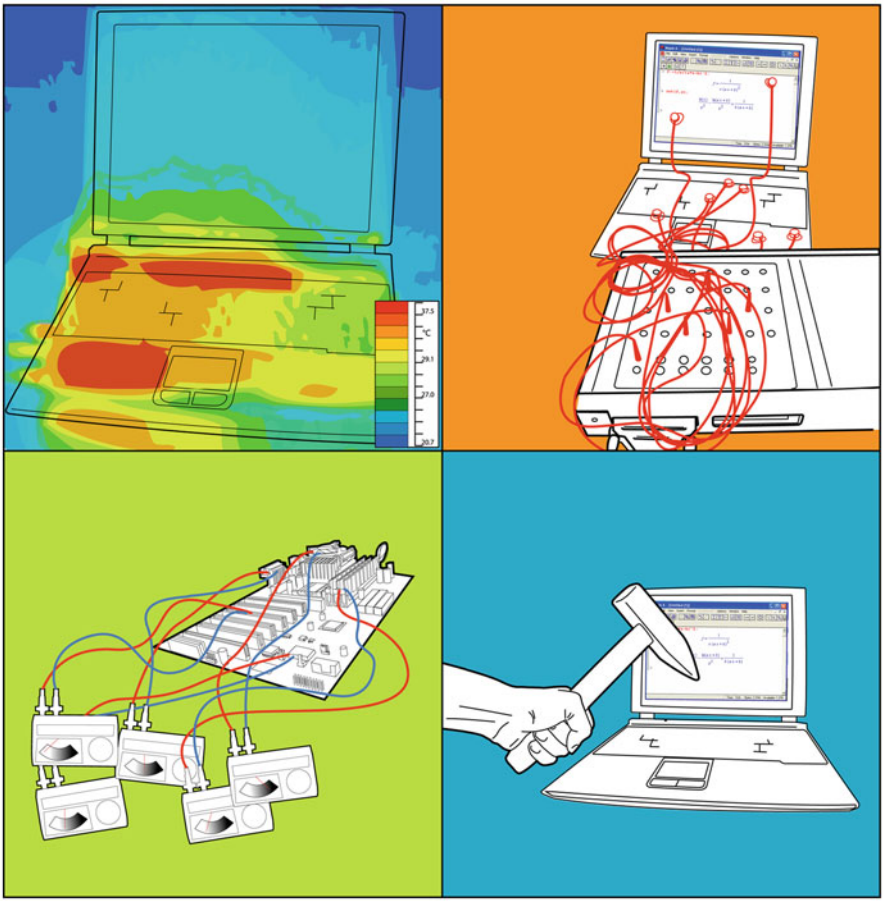


Fig. 7.3 Illustration of Ulrich's thought experiment. The operations of a computer program are studied with methods from neuroscience. *Top left panel*: heat radiation scan; *top right panel*: event-related potentials; *lower left panel*: single-unit recordings; *lower right panel*: experimental lesions. Figure reprinted with permission from [29]

7.3 Cognitive Neuroscience

The annual meetings of the *Society for Neuroscience* attract up to 40,000 participants, and plenary lectures are given by celebrities such as the Dalai Lama. Based on the attendance to their respective annual meeting, neuroscientists outnumber mathematical psychologist by a factor of 200 to 1. Cognitive neuroscientists use brain measurement techniques to study cognitive processes such as perception, attention, learning, emotion, decision-making, etc. Most of this work involves an empirical comparison between groups, treatments, or experimental conditions. For instance, Rouw and Scholte [30] compared a group of control participants with a group of

Table 7.2 First seven articles published in the June 2012 issue of the *Journal of Cognitive Neuroscience*

Title	Reference
Focal brain lesions to critical locations cause widespread disruption of the modular organization of the brain	[33]
Playing a first-person shooter video game induces neuroplastic change	[34]
Closing the gates to consciousness: Distractors activate a central inhibition process	[35]
TMS of the FEF interferes with spatial conflict	[36]
Local field potential activity associated with temporal expectations in the macaque lateral intraparietal area	[37]
Spatio-temporal brain dynamics mediating post-error behavioral adjustments	[38]
Hippocampal involvement in processing of indistinct visual motion stimuli	[39]

grapheme-color synesthetes, people who experience a specific color whenever they see a particular letter or number (e.g., "T is bright red"). Diffusion tensor imaging confirmed the hypothesis that the added sensations in synesthesia are associated with more coherent white matter tracts in various brain areas in frontal, parietal, and temporal cortex. In another example, Jepma and Nieuwenhuis [31] used a reinforcement learning task in which participants have to maximize rewards by making a series of choices with an uncertain outcome. The main result was that baseline pupil diameter was larger preceding exploratory choices (i.e., choices associated with a large uncertainty in outcome) than it was preceding exploitative choices (i.e., choices associated with a small uncertainty in outcome). Pupil diameter is an indirect marker for the activity of the locus coeruleus, a nucleus that modulates the norepinephrine system. Hence, the results are consistent with *adaptive gain theory*, according to which activity in the locus coeruleus regulates the balance between exploration and exploitation. A final example concerns the work by Ding and Gold [32], who showed that electrical microstimulation of the monkey caudate nucleus biases performance in a random-dot motion task.[2] This result suggests that the caudate has a causal role in perceptual decision making.

To give you a further impression of the work done by cognitive neuroscientists, Table 7.2 provides an overview of the articles published in the June 2012 issue of the *Journal of Cognitive Neuroscience*. Compared to the mathematical psychology approach, the cognitive neuroscience approach is geared towards understanding cognition on a relatively concrete level of implementation: what brain areas, neural processes, and circuits are involved in a particular cognitive process?

It is tempting to believe that the level of implementation is the level that is somehow appropriate for the study of cognition. This is suggested, for example, by the adage "the mind is what the brain does". However, Ulrich's laptop metaphor shows that such a conclusion is premature; clearly, the analytical integration that Maple

[2] In this popular perceptual task, the participant has to judge the apparent direction of a cloud of moving dots.

accomplishes is "what the laptop does", but it does not follow that we need or want to study the properties of the laptop in order to understand how Maple handles integrals analytically. Thus, even though "the mind is what the brain does", it is not automatically the case that when we measure the brain we learn a great deal about the mind. Readers who doubt this statement are advised to read the contributions that follow the article by Coltheart [28]; here, the discussants have to put in hard work to come up with just a single example of how functional neuroimaging has provided data to discriminate between competing psychological theories.

In order for cognitive neuroscience to have impact on psychological theory, it is important that the two are linked [40–42]. One way to accomplish such linking is by elaborating the psychological theory such that it becomes explicit about the brain processes involved [43]; another way is by using formal models to connect findings from neuroscience to the cognitive processes at hand. For instance, a mathematical psychologist may use the DDM to state that, when prompted to respond quickly, participants become less cautious, that is, they require less evidence before they are willing to make a decision. This description of cognition is relatively abstract and does not speak to how the brain implements the process. A neuroscientist may make this more concrete and suggest that the instruction to respond quickly leads to an increase of the baseline level of activation in the striatum, such that less input from cortex is needed to suppress the output nuclei of the basal ganglia, thereby releasing the brain from tonic inhibition and allowing an action to be executed [44, 45]. Thus, the DDM may provide an estimate of a latent cognitive process (e.g., response caution) which may then be compared against activation patterns in the brain. By using formal models that estimate psychological processes, this particular neuroscience approach furthers real theoretical progress and potentially bridges the divide between the implementational level and the algorithmic level [46].

7.4 Model-Based Cognitive Neuroscience: Symbiosis of Disciplines

The goal of model-based cognitive neuroscience is to bridge the gap between brain measurements and cognitive process with the help of formal models (e.g.,[10, 47–51]). This interdisciplinary approach is illustrated in Fig. 7.4. The figure shows that experimental psychology, mathematical psychology, and cognitive neuroscience all pursue a common goal: a better understanding of human cognition. It is often difficult, however, to learn about the relevant cognitive processes from the data directly – often, one first needs a mathematical model to provide quantitative estimates for the cognitive processes involved. Next, the estimates of the cognitive processes can be related to the brain measurements.

The "model-in-the-middle" [52] symbiosis of disciplines is useful in several ways. Rather than discuss the advantages abstractly, the next two sections provide concrete illustrations of the mutually beneficial relationship between mathematical models and brain measurements (see also [1]).

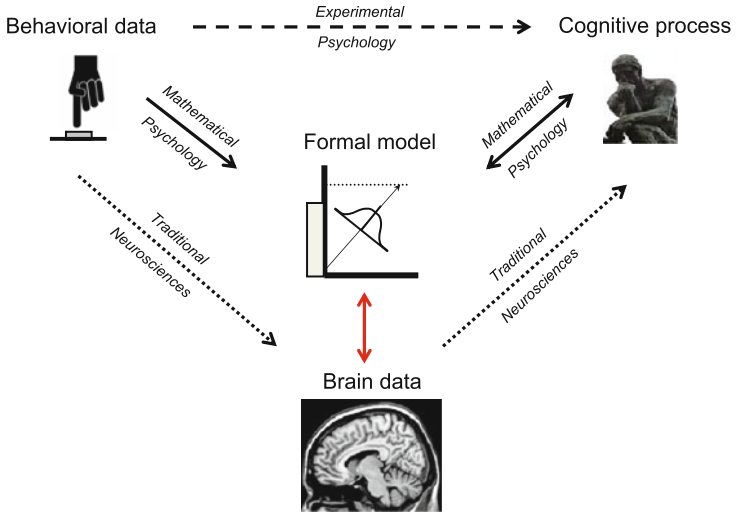


Fig. 7.4 The model-in-the-middle approach unites experimental psychology, mathematical psychology, and cognitive neuroscience, as the main goal of all three disciplines is to understand more deeply the processes and mechanisms that constitute human cognition. The *red* arrow indicates the reciprocal relation between measuring the brain and modeling behavioral data. (Figure reprinted with permission from [1])

7.4.1 Use of Mathematical Models for Cognitive Neuroscience

Mathematical models are useful in many ways. First, they decompose observed behavioral data into latent cognitive processes. Brain measurements can then be associated with particular cognitive processes instead of behavioral data. For example, Chap. 2 explained how the LBA model—just as the DDM model discussed in Sect. 7.2 and in Chap. 3—decomposes response time distributions and error rates into underlying concepts such as *response caution* and the *speed of information processing*. This decomposition can be used to demonstrate that a particular experimental manipulation had the desired effect. For instance, participants in a study by Forstmann and colleagues [44] performed a random-dot motion task under various cue-induced levels of speed-stress. That is, before each stimulus a cue indicated whether the stimulus needed to be classified accurately or quickly. Because the authors were interested in the neural basis of the speed-accuracy tradeoff, they hoped that the cue would selectively affect the LBA response caution parameter. And indeed, the model decomposition confirmed that this was the case. Note that without concrete model fitting, this conclusion had been premature, unwarranted, and potentially misleading—it is certainly possible that instructions to respond more quickly can, for some tasks, also induce a change in speed of processing, or a change in the time required for peripheral processes.

Another advantage that the model-based decomposition brings is that, even when a particular manipulation is not process-pure, one may associate brain measurements specifically with the parameter of interest. For instance, suppose that in the speed-accuracy experiment by Forstmann and colleagues [44], a cue to respond more quickly had also lowered the speed of information processing. This means that the critical fMRI contrasts are contaminated because they reflect a combination of two effects: the change in response caution associated with the speed-accuracy tradeoff, and the change in drift rate associated with the lower speed of processing. One method to address this complication is to correlate the contaminated brain measures (e.g., the average change in the BOLD response for each participant) with the process of interest (e.g., the individual estimates of the change in response caution), perhaps after partialling out the effects of the nuisance process. This method identifies those voxels that relate to the cognitive construct of response caution.

Finally, the model-based decomposition allows one to take into account individual differences. For instance, Forstmann and colleagues [44] found that speed cues activated the right anterior striatum and the right pre-supplementary motor area (pre-SMA). This result was corroborated by an analysis of individual differences: participants with a relatively large cue-induced decrease in response caution also showed a relatively large increase in activation in the right anterior striatum and right pre-SMA. Of course, such an analysis is only meaningful if there are substantial individual differences to begin with; if all participants respond to the cue in approximately the same way then the group-average result will be highly significant but the individual difference analysis may not be significant at all.

In another example of the importance of individual differences, Forstmann and colleagues [53] studied the neural basis of prior knowledge in perceptual decision-making. As before, participant performed a random-dot motion task; this time, the cue gave prior information about the likely direction of movement of the upcoming stimulus. The cue "L9", for example, indicated that the probability was 90 % that the upcoming stimulus would move to the left (see also [54]). The cue-induced bias was clearly visible in the behavioral data: responses were much faster and more often correct when the cue was reliable and informative. Surprisingly, however, the fMRI contrast did not reveal any significant results. After including an LBA response bias parameter as a covariate in the fMRI analysis, however, the results showed significant cue-related activation in regions that generally matched the theoretical predictions (e.g., putamen and orbitofrontal cortex). The reason for the discrepancy is that by adding the response bias parameter we can account for individual differences in people's reactions to the cue. Some participants exhibited a lot of bias, and others only a little. These individual differences in the latent cognitive process are usually not incorporated the fMRI analysis and hence add to the error term instead. By explicitly accounting for individual differences the error term is reduced and experimental power is increased.

Mathematical models are also useful because they can drive the search for brain areas involved in a particular cognitive function. In fMRI research, for instance, this means that a model's predictions are convolved with the hemodynamic response function. Next, the predicted blood oxygenation level dependent signal (BOLD)

response profiles are used to search for areas in the brain with similar activation profiles. The search can be exploratory or more confirmatory. A prominent example of the latter approach is the recent work that attempts to link the components of the ACT-R model to different brain regions (see the final chapter of this book as well as [55, 56]).

Another way in which mathematical models are useful is that they may help demonstrate that performance in superficially different tasks may be driven by the same cognitive process. For example, the cognitive neuroscience literature suggests that perceptual categorization and old-new recognition are subserved by separate neural systems. In contrast, mathematical models of these tasks suggests that similar processes are involved. Specifically, exemplar models posit that people categorize a test object by comparing it to exemplars stored in memory. Both categorization and recognition decisions are thought to be based on the summed similarity of the test object to the exemplars [57, 58]. These conflicting perspectives were recently reconciled by Nosofsky and colleagues [59], who argued that categorization and recognition differ not in terms of the underlying process, but in terms of the underlying criterion settings: in recognition, the criterion needs to be strict, since the test object needs to match one of the study items exactly; in categorization, the criterion can be more lax, as exact matches are not needed. In an fMRI experiment, Nosofsky and colleagues [59] induced participants to use different criterion settings; the resulting data were then fit by an exemplar model. Results confirmed that (1) the exemplar model provides a good account of both categorization and recognition, with only criterion settings free to vary; (2) the average task-related differences in brain activation can be explained by differences in evidence accumulation caused by systematically varying criterion settings; and (3) participants with high criterion settings show large BOLD differences between old and random stimuli in the frontal eye fields and the anterior insular cortex. Hence, Nosofsky and colleagues [59] concluded that there is little evidence that categorization and recognition are subserved by separate memory systems. The most important lesson to be learned from this work is that differences in brain activation do not necessarily indicate different underlying mechanisms or processes. Differences in brain activation can also come about through differences in stimulus surface features (which Nosofsky et al. controlled for) and differences in criterion settings. A mathematical model can estimate these criterion settings and allow a statistical assessment of their importance.

In addition to the above, mathematical models have general worth because they provide (1) a concrete implementation of a theoretical framework; (2) a coherent interpretive framework; and (3) a guide to experimental manipulations that are particularly informative. In sum, it is evident that for cognitive neuroscience, the use of mathematical models comes with considerable advantages. The reverse—the advantages of cognitive neuroscience for mathematical models—is the topic of the next section.

7.4.2 Use of Cognitive Neuroscience for Mathematical Models

Until recently, findings from cognitive neuroscience had little impact on model development in mathematical psychology. Exceptions that confirm the rule are parallel distributed processing models [60, 61] and neurocomputational models developed to take into account the details of neural processing [43, 62–64]. This state of affairs is changing, and for good reason: response times, error rates, and confidence judgements ultimately provide little information to tell apart mathematical models with incompatible assumptions and architectures [41]. For instance, Ditterich [23] showed that behavioral data are insufficient to discriminate between multiple-choice response time models that have evidence integrators with and without leakage, with and without feedforward and feedback inhibition, and with and without linear and non-linear mechanisms for combining information across choice alternatives. Neural data, however, can able to adjudicate between the hypothesized mechanisms, at least in potentia [23].

As a specific example, consider a generic response time model with N evidence accumulators, one for each choice alternative, that race to a threshold. The model can account for the speed-accuracy tradeoff by changing the distance from baseline to threshold. However, the model is mute on whether instructions to respond accurately increase threshold or decrease baseline; in fact, these mechanisms are mathematically equivalent. Nevertheless, the mechanisms are not conceptually equivalent, and neural data could discriminate between the two accounts. A similar example considers the change in processing that occurs as the number of incorrect choice alternatives increases. Such an increase in task difficulty requires a longer period of evidence accumulation in order to reach an acceptable level of accuracy in identifying the target alternative. The evidence accumulation process can be extended either by increasing thresholds or by decreasing the baseline. Using behavioral data, there is no way to tell these two mechanisms apart. However, Churchland and colleagues [65] used single-cell recordings to show that, confronted with the prospect of having to chose between four instead of two random-dot choice alternatives, monkeys had decreased firing rates in the lateral intraparietal area. These single cell recordings are consistent with a changing baseline account rather than a shifting threshold account.

Hence, the general promise is that data from cognitive neuroscience may provide additional constraints. An interesting illustration of this principle is provided by Purcell and colleagues [66], who studied how monkeys perform a visual search task in which they have to make an eye movement toward a single target presented among seven distractors. Several models were fit to the data, and initial constraint was gained by using the measured spike trains as input to the evidence accumulators. This creates an immediate challenge: the models must determine when the accumulators start to be driven by the stimulus, because the neural activity that precedes stimulus onset is uninformative and its accumulation can only harm performance. Hence, models with perfect integration failed, as they were overly impacted by early spiking activity that was unrelated to the stimulus. Models with leaky integration did not suffer from early spiking activity, but their predictions were inconsistent with another neural

constraint: the spiking activity from movement neurons. In the end, the only class of models that survived the test were gated integration models, models that block the influence of noise inputs until a certain threshold level of activation is reached.

In general, it is clear that neuroscience data hold tremendous potential for answering questions that mathematical psychologists can never address with behavioral data alone.

7.5 Open Challenges

The above examples have only scratched the surface of the work conducted within model-based cognitive neuroscience. Nevertheless, the field can greatly expand by considering a broader range of mathematical models, and a broader range of brain measures (e.g., structural MRI, event-related potentials in EEG, genetics, pharmacology, etc.). Also, model dynamics can be linked more closely to brain dynamics, either by constructing a single overarching statistical model, or by developing single trial estimates of cognitive processes. For example, van Maanen and colleagues [67] extended the LBA model to estimate drift rate and thresholds on a trial-by-trial basis. This allows a more direct comparison with neurophysiological data, which also vary on a trial-by-trial basis.

Another challenge is to balance the desire for parsimonious models (i.e., models with few parameters and clear mechanisms) against the reality of the brain's overwhelming complexity. The appropriate level of model complexity depends very much on the goals of the researchers. If the goal is to obtain estimates of latent cognitive processes, then the model needs to be relatively simple—the behavioral data simply do not provide sufficient support for models that are relatively complex. On the other hand, if the goal is to create a model that accounts for the detailed interactions between neurons or brain systems, the model needs to be more intricate.

A final challenge is that, despite the intuitive attractiveness of results from rats and monkeys, we should remain aware of the possibility that some of the results obtained with these species may not carry over to humans. This may be due to differences in anatomy, but other factors can contribute as well. For instance, recent work suggests that monkeys who perform a speeded choice task may experience an increased urgency to respond [26] that can expresses itself in response thresholds that decreases over time [25, 68]. Before concluding that response urgency or collapsing bounds are a universal signature of human decision making, however, we need to make sure that the pattern in monkeys is obtained in humans as well. This requires a careful modeling exercise in which benchmark data sets are fit with two versions of the same sequential sampling model: one that has constant thresholds and one that has collapsing bounds [69]. It is entirely possible, for instance, that collapsing bounds are used by monkeys because they want to maximize reward rate [68]; first-year psychology undergraduates, however, are usually not reinforced with squirts of orange juice and may approach the task with a different goal. The collapsing-bound hypothesis shows promise and is worth exploring, but its generality is pending investigation.

7.6 Concluding Comments

The examples in this chapter have shown how mathematical models can advance cognitive neuroscience, and how cognitive neuroscience can provide constraint for mathematical models. The increasing collaboration between these historically separate fields of study is an exciting new development that we believe will continue in the future.

Exercises

1. Section 7.2: in what sense is the DDM similar to signal-detection theory?
2. Section 7.2: Can you find a concrete example to refute Coltheard's claim that "no functional neuroimaging research to date has yielded data that can be used to distinguish between competing psychological theories"?
3. Section 7.3: Read the articles by Miller [70] and by Insel [71] on the impact of neuroscience on psychiatry and clinical psychology. Who do you agree with, and why?
4. Section 7.2: Read the Gold and Shadlen [12] article and prepare a 30-min presentation on it, critically summarizing and explaining its content.
5. Describe a mathematical model (not discussed in this chapter) that could find application in cognitive neuroscience.
6. Mention one pro and one con for each of the following brain measures: single-cell recordings, ERP, fMRI, and DWI.
7. Can you think of concrete research questions in cognitive neuroscience that could profit from a model-based approach?

Further Reading

1. Ratcliff and McKoon [74] offer an overview of the drift diffusion model and its relation to cognitive neuroscience.
2. http://neuroskeptic.blogspot.com/2012/02/mystery-joker-parodies-neuroscience.html tells a tale about neuroscience and Sigmund Freud.
3. Churchland and Ditterich [75] discuss recent developments in models for a choice between multiple alternatives.
4. We consider our work on bias [53] as one of our better efforts. Unfortunately, the reviewers did not agree, and one even commented "Flawed design, faulty logic, and limited scholarship engender no confidence or enthusiasm whatsoever".

References

1. Forstmann BU, Wagenmakers EJ, Eichele T, Brown S, Serences JT (2011) Reciprocal relations between cognitive neuroscience and formal cognitive models: Opposites attract? Trends Cognit Sci 15:272
2. Song XY, Lee SY (2012) A tutorial on the Bayesian approach for analyzing structural equation models. J Math Psychol 56:135
3. Frank TD, Silva PL, Turvey MT (2012) Symmetry axiom of Haken–Kelso–Bunz coordination dynamics revisited in the context of cognitive activity. J Math Psychol 56:149
4. Asano M, Basieva I, Khrennikov A, Ohya M, Tanaka Y (2012) Quantum-like generalization of the Bayesian updating scheme for objective and subjective mental uncertainties. J Math Psychol 56:166
5. Luce RD (2012) Torgerson's conjecture and Luce's magnitude production representation imply an empirically false property. J Math Psychol 56:176
6. Poppe S, Benner P, Elze T (2012) A predictive approach to nonparametric inference for adaptive sequential sampling of psychophysical experiments. J Math Psychol 56:179
7. DeCarlo LT (2012) On a signal detection approach to *m*–alternative forced choice with bias, with maximum likelihood and Bayesian approaches to estimation. J Math Psychol 56:196
8. Dutilh G van Ravenzwaaij D, Nieuwenhuis S, van der Maas HLJ, Forstmann BU, Wagenmakers EJ (2012) How to measure post–error slowing: A confound and a simple solution. J Math Psychol 56:208
9. Townsend JT (2008) Mathematical psychology: Prospects for the 21st century: A guest editorial. J Math Psychol 52:269
10. Gold JI, Shadlen MN (2001) Neural computations that underlie decisions about sensory stimuli. Trends Cognit Sci 5:10
11. Gold JI, Shadlen MN (2002) Banburismus and the brain: Decoding the relationship between sensory stimuli, decisions, and reward. Neuron 36:299
12. Gold JI, Shadlen MN (2007) The neural basis of decision making. Annu Rev Neurosci 30:535
13. van Ravenzwaaij D, Mulder M, Tuerlinckx F, Wagenmakers EJ (2012) Do the dynamics of prior information depend on task context? An analysis of optimal performance and an empirical test. Front Cognit Sci 3:132
14. Wagenmakers EJ (2009) Methodological and empirical developments for the Ratcliff diffusion model of response times and accuracy. Eur J Cognit Psychol 21:641
15. Brinley JF (1965) Cognitive sets, speed and accuracy of performance in the elderly. In: Welford AT, Birren JE (eds) Behavior, aging and the nervous system. Thomas, Springfield, pp 114–149
16. Cerella J (1985) Information processing rates in the elderly. Psychol Bull 98:67
17. Salthouse TA (1996) The processing–speed theory of adult age differences in cognition. Psychol Rev 103:403
18. Ratcliff R, Thapar A, Gomez P, McKoon G (2004) A diffusion model analysis of the effects of aging in the lexical–decision task. Psychol Aging 19:278
19. Ratcliff R, Thapar A, McKoon G (2007) Application of the diffusion model to two–choice tasks for adults 75–90 years old. Psychol Aging 22:56
20. Ratcliff R, Thapar A, McKoon G (2004) A diffusion model analysis of the effects of aging on recognition memory. J Mem Lang 50:408
21. Wagenmakers EJ, Ratcliff R, Gomez P, McKoon G (2008) A diffusion model account of criterion shifts in the lexical decision task. J Mem Lang 58:140
22. Bogacz R, Brown E, Moehlis J, Holmes P, Cohen JD (2006) The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced choice tasks. Psychol Rev 113:700
23. Ditterich J (2010) A comparison between mechanisms of multi–alternative perceptual decision making: Ability to explain human behavior, predictions for neurophysiology, and relationship with decision theory. Front Decis Neurosci 4:184
24. Usher M, McClelland JL (2001) On the time course of perceptual choice: The leaky competing accumulator model. Psychol Rev 108:550

25. Drugowitsch J, Moreno–Bote R, Churchland AK, Shadlen MN, Pouget A (2012) The cost of accumulating evidence in perceptual decision making. J Neurosci 32:3612
26. Cisek P, Puskas GA, El–Murr S (2009) Decisions in changing conditions: The urgency–gating model. J Neurosci 29:11560
27. Krajbich I, Armel C, Rangel A (2010) Visual fixations and comparison of value in simple choice. Nat Neurosci 13:1292
28. Coltheart M (2006) What has functional neuroimaging told us about the mind (so far)? Cortex 42:323
29. Ulrich R (2009) Uncovering unobservable cognitive mechanisms: The contribution of mathematical models. In: Rösler F, Ranganath C, Röder B, Kluwe RH (eds) Neuroimaging of human memory: linking cognitive processes to neural systems. Oxford University Press, Oxford, pp 25–41
30. Rouw R, Scholte HS (2007) Increased structural connectivity in grapheme-color synesthesia. Nat Neurosci 10:792
31. Jepma M, Nieuwenhuis S (2011) Pupil diameter predicts changes in the exploration-exploitation trade-off: Evidence for the adaptive gain theory. J Cognit Neurosci 23:1587
32. Ding L, Gold JI (2012) Separate, causal roles of the caudate in saccadic choice and execution in a perceptual decision task. Neuron 75:865
33. Gratton C, Nomura EM, Pérez F, D'Esposito M (2012) Focal brain lesions to critical locations cause widespread disruption of the modular organization of the brain. J Cognit Neurosci 24:1275
34. Wu S, Cheng CK, Feng J, D'Angelo L, Alain C, Spence I (2012) Playing a first–person shooter video game induces neuroplastic change. J Cognit Neurosci 24:1286
35. Niedeggen M, Michael L, Hesselmann G (2012) Closing the gates to consciousness: Distractors activate a central inhibition process. J Cognit Neurosci 24:1294
36. Bardi L, Kanai R, Mapelli D, Walsh V (2012) TMS of the FEF interferes with spatial conflict. J Cognit Neurosci 24: 1305
37. Premereur E, Vanduffel W, Janssen P (2012) Local field potential activity associated with temporal expectations in the macaque lateral intraparietal area. J Cognit Neurosci 24:1314
38. Manuel AL, Bernasconi F, Murray MM, Spierer L (2012) Spatio–temporal brain dynamics mediating post–error behavioral adjustments. J Cognit Neurosci 24:1331
39. Fraedrich EM, Flanagin VL, Duann JR, Brandt T, Glasauer S (2012) newblock Hippocampal involvement in processing of indistinct visual motion stimuli. JJ Cognit Neurosci 24:1344
40. Brown SD (2012) Common ground for behavioural and neuroimaging research. Aust J Psychol 64:4
41. Schall JD (2004) On building a bridge between brain and behavior. Annu Rev Psychol 55:23
42. Teller D (1984) Linking propositions. Vis Res 24:1233
43. Ashby FG, Helie S (2011) A tutorial on computational cognitive neuroscience: Modeling the neurodynamics of cognition. J Math Psychol 55:273
44. Forstmann BU, Dutilh G, Brown S, Neumann J, von Cramon DY, Ridderinkhof KR, Wagenmakers EJ (2008) Striatum and pre–SMA facilitate decision–making under time pressure. Proc Natl Acad Sci U S A 105:17538
45. Mink JW (1996) The basal ganglia: Focused selection and inhibition of competing motor programs. Prog Neurobiol 50:381
46. Marr D (1982) Vision: a computational investigation into the human representation and processing of visual information. Henry Holt and Company, San Francisco (H W. Freeman)
47. Dolan RJ (2008) Neuroimaging of cognition: Past, present, and future. Neuron 60:496
48. Friston KJ (2009) Modalities, modes, and models in functional neuroimaging. Science 326:399
49. Hanes DP, Schall JD (1996) Neural control of voluntary movement initiation. Science 274:427
50. Mars RB, Shea NJ, Kolling N, Rushworth MF. (2012) Model-based analyses: Promises, pitfalls, and example applications to the study of cognitive control. Q J Exp Psychol (Hove), 65(2): 252–67
51. O'Doherty JP, Hampton A, Kim H (2007) Model–based fMRI and its application to reward learning and decision making. Ann N Y Acad Sci 1104:35

52. Corrado G, Doya K (2007) Understanding neural coding through the model–based analysis of decision making. J Neurosci 27:8178
53. Forstmann BU, Brown S, Dutilh G, Neumann J, Wagenmakers EJ (2010) The neural substrate of prior information in perceptual decision making: A model-based analysis. Front Hum Neurosci 4:40
54. Mulder M, Wagenmakers EJ, Ratcliff R, Boekel W, Forstmann BU (2012) Bias in the brain: A diffusion model analysis of prior probability and potential payoff. J Neurosci 32:2335
55. Anderson JR, Fincham JM, Qin Y, Stocco A (2008) A central circuit of the mind. Trends Cognit Sci 12:136
56. Borst JP, Taatgen NA, Stocco A van Rijn H (2010) The neural correlates of problem states: Testing fMRI predictions of a computational model of multitasking. PLoS ONE 5
57. Nosofsky RM (1986) Attention, similarity, and the identification–categorization relationship. J Exp Psychol: Gener 115:39
58. Nosofsky RM, Palmeri TJ (1997) An exemplar–based random walk model of speeded classification. Psychol Rev 104:266
59. Nosofsky RM, Little DR, James TW (2012) Activation in the neural network responsible for categorization and recognition reflects parameter changes. Proc Natl Acad Sci U S A 109:333
60. Rumelhart DE, Hinton GE, McClelland JL (1986) A general framework for parallel distributed processing. In: Rumelhart DE, McClelland JL, the PDP Research Group (eds) Parallel distributed processing: explorations in the microstructure of cognition (Vol 1). MIT Press, Cambridge, pp. 45–76
61. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning internal representations by error propagation. In: Rumelhart DE, McClelland JL, the PDP Research Group (eds) Parallel distributed processing: explorations in the microstructure of cognition, Vol 1. MIT Press, Cambridge, pp 318–362
62. Ratcliff R, Frank MJ (2012) Reinforcement–based decision making in corticostriatal circuits: Mutual constraints by neurocomputational and diffusion models. Neural Comput 24:1186
63. Stocco A, Lebiere C, Anderson JR (2010) Conditional routing of information to the cortex: A model of the basal ganglia's role in cognitive coordination. Psychol Rev 117:541
64. Stocco A (2012) Acetylcholine–based entropy in response selection: A model of how striatal interneurons modulate exploration, exploitation, and response variability in decision–making. Front Neurosci 6:18
65. Churchland AK, Kiani R, Shadlen MN (2008) Decision–making with multiple alternatives. Nat Neurosci 11:693
66. Purcell BA, Heitz RP, Cohen JY, Schall JD, Logan GD, Palmeri TJ (2010) Neurally constrained modeling of perceptual decision making. Psychol Rev 117:1113
67. van Maanen L, Brown S, Eichele T, Wagenmakers EJ, Ho T, Serences J, Forstmann BU (2011) Neural correlates of trial–to–trial fluctuations in response caution. J Neurosci 31:17488
68. Deneve S (2012) Making decisions with unknown sensory reliability. Front Neurosci 6:75
69. Milosavljevic M, Malmaud J, Huth A, Koch C, Rangel A (2010) The Drift Diffusion Model can account for the accuracy and reaction time of value–based choices under high and low time pressure. Judgm Decis Making 5:437
70. Miller GA (2010) Mistreating psychology in the decades of the brain. Perspect Psychol Sci 5:716
71. Insel T (2010) Faulty circuits. Sci Am 302:44
72. Wagenmakers EJ, van der Maas HJL, Grasman PP (2007) Psychon Bull Rev 14:3
73. Logothetis NK (2003) In: Parker A (ed) The physiology of cognitive processes. Oxford University Press, New York
74. Ratcliff R, McKoon G (2008) The diffusion decision model: Theory and data for two–choice decision tasks. Neural Comput 20:873
75. Churchland AK, Ditterich J (in press) New advances in understanding decisions among multiple alternatives. Curr Opin Neurobiol

Part II
How Cognitive Models Inform the Cognitive Neurosciences

Chapter 8
Linking Across Levels of Computation in Model-Based Cognitive Neuroscience

Michael J. Frank

Abstract Computational approaches to cognitive neuroscience encompass multiple levels of analysis, from detailed biophysical models of neural activity to abstract algorithmic or normative models of cognition, with several levels in between. Despite often strong opinions on the 'right' level of modeling, there is no single panacea: attempts to link biological with higher level cognitive processes require a multitude of approaches. Here I argue that these disparate approaches should not be viewed as competitive, nor should they be accessible to only other researchers already endorsing the particular level of modeling. Rather, insights gained from one level of modeling should inform modeling endeavors at the level above and below it. One way to achieve this synergism is to link levels of modeling by quantitatively fitting the behavioral outputs of detailed mechanistic models with higher level descriptions. If the fits are reasonable (e.g., similar to those achieved when applying high level models to human behavior), one can then derive plausible links between mechanism and computation. Model-based cognitive neuroscience approaches can then be employed to manipulate or measure neural function motivated by the candidate mechanisms, and to test whether these are related to high level model parameters. I describe several examples of this approach in the domain of reward-based learning, cognitive control, and decision making and show how neural and algorithmic models have each informed or refined the other.

8.1 Introduction

Cognitive neuroscience is inherently interested in linking levels of analysis, from biological mechanism to cognitive and behavioral phenomena. But there are not just two levels, rather, a continuum of many. One can consider the implications of particular ion channel conductances and receptors, the morphological structure of individual neurons, the translation of mRNA, intracellular molecular signaling cascades involved in synaptic plasticity, and so forth. At the cognitive level, the

M. J. Frank (✉)
Cognitive, Linguistic & Psychological Sciences, Brown Institute for Brain Science, Providence, USA
e-mail: Michael_Frank@Brown.edu

B. U. Forstmann, E.-J. Wagenmakers (eds.), *An Introduction to Model-Based Cognitive Neuroscience,* DOI 10.1007/978-1-4939-2236-9_8

field typically discusses constructs such as working memory, executive control, reinforcement learning, episodic memory, to name a few. In between these levels reside architectures of neural systems, such as the frontal cortex, parietal cortex, hippocampus and basal ganglia, the interactions among all of these systems, and their modulations by neurotransmitters in response to relevant task events. Computational models greatly facilitate the linking of levels of analysis, because they force one to be explicit about their assumptions, to provide a unifying coherent framework, and to specify the computational objectives of any given cognitive problem which provide constraints on interpreting the underlying mechanisms.

Nevertheless, the question remains of which level of modeling to use. The field of computational neuroscience for example typically considers how low level mechanisms can give rise to higher level "behaviors", but where behavior here is defined in terms of the changes in membrane potentials of individual neurons or even compartments within neurons, or in terms of synchrony of neural firing across populations of cells. The field of computational cognitive science, on the other hand, considers how behavioral phenomena might be interpreted as optimizing some computational goal, like minimizing effort costs, maximizing expected future reward, or optimally trading off uncertainty about multiple sources of perceptual and cognitive information to make inferences about causal structure. In between, computational cognitive neuroscience considers how mechanisms within neural systems can solve tradeoffs, for example between pattern separation and pattern completion in hippocampal networks [56], between updating and maintenance of working memory in prefrontal cortex [11, 34], or between speed and accuracy in perceptual decision making as a function of connectivity between cortex and basal ganglia [9, 52]. Even with this limited number of examples however, models took multiple levels of description, from those using detailed spiking neurons to higher level computations capturing reaction time distributions, where latent estimated parameters are correlated with neural measures extracted from functional imaging. In general, theorists and experimentalists are happy to "live" at one level of analysis which intuitively has greatest aesthetic, even though there is large variation in the appreciation of what constitutes the "right" level. Although there is a rich literature in mathematical psychology on how to select the most parsimonious model that best accounts for data without overfitting, this issue primarily pertains to applications in which one quantitatively fits data, and not to the endeavor of constructing a generative model. For example, if given only error rates and reaction time distributions, a mathematical psychologist will select a minimalist model that can best account for these data with a few free parameters, but this model will not include the internal dynamics of neural activities. Some researchers also perform model selection together with functional imaging to select the best model that accounts for correlations between brain areas and psychological parameters (e.g., [47]), but even here the neural data are relatively sparse and the observations do not include access to internal circuitry dynamics, neurotransmitters, etc—even though everyone appreciates that those dynamics drive the observed measurements.

The reason everyone appreciates this claim is that the expert cognitive neuroscientist has amassed an informative prior on valid models based on a large body

of research that spans multiple methods, species, analysis tools etc. Nevertheless, it is simply not feasible to apply these informative priors into a quantitative model selection process given much more sparse data (e.g. given BOLD fMRI data and behavioral responses one would not be advised to try to identify the latent parameters of a detailed neuronal model that includes receptor affinities, membrane potential time constants, etc).

In this chapter, I advocate an alternative, multi-level modeling strategy to address this issue. This strategy involves critical integration of the two levels to derive predictions for experiments and to perform quantitative fits to data. In one prong, theoreticians can create detailed neural models of interacting brain areas, neurotransmitters, etc, constrained by a variety of observations. These models attempt to specify interactions among multiple neural mechanisms and can show how the network dynamics recapitulate those observed in electrophysiological data, and how perturbation of those dynamics (by altering neurotransmitter levels or receptor affinities) can lead to observable changes in behavior that qualitatively match those reported in the literature. In a second prong, the modeler can construct a higher level computational model motivated by the mathematical psychological literature which summarizes the basic cognitive mechanism. Examples of this level include simple reinforcement learning models from the machine learning literature (e.g., Q learning; [72]), or sequential sampling models of decision making such as the drift diffusion model (see [59], for a review). These models should be constructed such that the parameters are identifiable, meaning that if one generates fake data from the model they should be able to reliably recover the generative parameters and to differentiate between changes that would be due to alterations in one parameter separately from other parameters. Ideally, the experimental task design will be informed by this exercise prior to conducting the experiment, so that the task can provide conditions that would be more diagnostic of differences in underlying parameters. The identifiability of a model is a property of not only of the model itself but also the different task conditions. To see this, consider a model in which task difficulty of one sort or another is thought to selectively impact a given model parameter. One might include two or more different difficulty levels, but the degree to which these influence behavior, and thus lead to observable differences that can be captured by the relevant model parameter, often interact with the other task and model parameters.

Given an identifiable model and appropriate task, the next step is to link the levels of modeling to each other. Here, one generates data from the detailed neural model exposed to the task such that it produces data at the same level as one would obtain from a given experiment—error rates and RT distributions for example, or perhaps also some summary statistic of neural activity in a given simulated brain area in one condition vs. another as one might obtain with fMRI or EEG. Then, these outputs are treated just as one does when fitting human (or other animal) data: assuming they were generated by the higher level model (or several candidate higher level models). Model selection and parameter optimization proceeds exactly as it would when fitting these models to actual human data. The purpose of this exercise is to determine which of the higher level models best summarizes the effective computations of the detailed model. Further, one can perform systematic parameter manipulations in the neural

model to determine whether these have observable selective effects on the higher level parameter estimates. If the fit is reasonable (for example if the measures of model fit are similar to those obtained by fitting the higher level model to human data), this process can lead to a higher level computational description of neural circuit function not directly afforded by the detailed neural dynamic simulations. (This approach is complementary to that taken by Bogacz and colleagues, who have used a single level of computation but made observations about how distinct aspects of the computational process can be mapped onto distinct nuclei within the cortico-basal ganglia network [7]. In Bogacz's work the computations afforded by this circuitry is identical to that of the optimal Bayesian model of decision making. In our models described below, these computations emerge from nonlinear neural dynamics and the mapping is approximate, not exact, and hence allow for potential refinements in the higher level description that best characterizes human cognition and behavior.)

When successful, this exercise can also motivate experiments in which the same biological manipulation is performed on actual participants (e.g. a medication manipulation, brain stimulation, pseudoexperimental manipulation via genetics). Indeed, by linking across levels of computation one derives precise, falsifiable predictions about which of the higher level observable and identifiable parameters will be affected, and in which direction, by the manipulation. It can also provide informative constraints on interpreting how component processes are altered as a function of mental illness [53]. Of course, one may derive these predictions intuitively based on their own understanding of neural mechanisms, but there are various instances in which explicit simulations with more detailed models can lend insight into interactions that may not have been envisioned otherwise.

8.1.1 Applications to Reinforcement Learning, Decision Making and Cognitive Control

The above "recipe" for linking levels of computation is relatively abstract. The rest of this chapter focuses on concrete examples from my lab. We study the neurocomputational mechanisms of cortico-basal ganglia functions—including action selection, reinforcement learning and cognitive control. For theory development, we leverage a combination of two distinct levels of computation. First, our lab and others have simulated interactions within and between corticostriatal circuits via dynamical neural systems models which specify the roles of particular neural mechanisms [26, 27, 30, 34, 43, 58, 73, 75]. These models are motivated by anatomical, physiological, and functional constraints. The mechanisms included in the detailed neural models were originally based on data from animal models, but integrated together into a systems-level functional model (i.e., one that has objectives and links to behavior). As such they have reciprocally inspired rodent researchers to test, and validate, key model predictions using genetic engineering methods in rodents by manipulating activity in separable corticostriatal pathways [44, 50, 66]. Second, we adopt and refine higher level mathematical models to analyze the functional properties of the

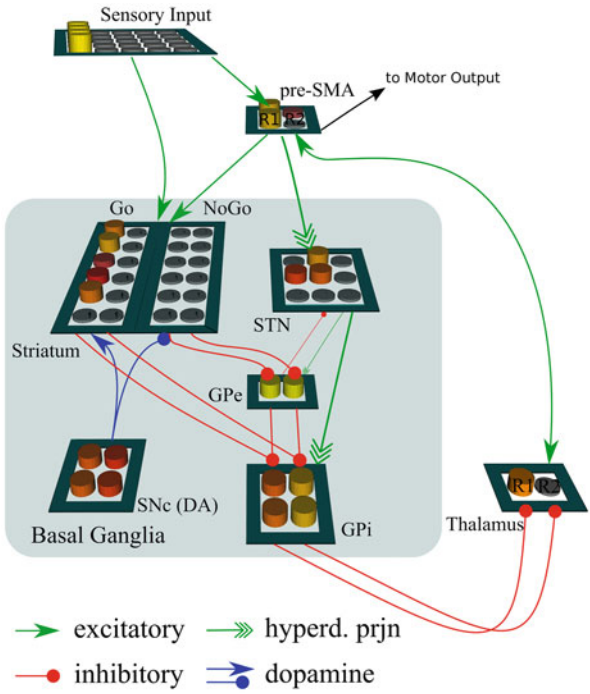


Fig. 8.1 Neural network model of a single cortico-basal ganglia circuit [2]. Sensory and motor (pre-SMA) cortices project to the basal ganglia. Two opposing "Go" and "NoGo" (direct and indirect) pathways regulate action facilitation and suppression based on reward evidence for and against each decision option. Dopamine (DA) modulates activity levels and plasticity in these populations, influencing both choice and learning. The 'hyperdirect' pathway from cortex to STN acts to provide a temporary Global NoGo signal inhibiting the selection of all actions, particularly under conditions of decision conflict (co-activation of competing pre-SMA units). *GPi/e* Globus Pallidus internal/ external segment

neurocognitive systems, affording a principled computational interpretation and allowing for tractable quantitative fits to brain-behavior relationships [5, 13, 21, 22, 58]. Examples of the two levels of description are shown in Fig. 8.1 (neural systems) and Figs. 8.2, 8.3 and 8.4 (abstractions).

For empirical support, we design computerized tasks sensitive to the hypothesized neural computations that probe *reinforcement learning, cognitive control, and reward-based decision making under uncertainty*. We provide quantitative estimates of individual performance parameters using mathematical models, yielding objective assessments of the degree to which subjects rely on specific computations when learning and making decisions [13, 14, 28, 30, 33, 38, 61]. We assess how these parameters vary with markers of neural activity (EEG, fMRI), and how they are altered as a function of illness, brain stimulation, pharmacology, and genetics [13, 14, 33, 41, 53, 57].

This approach has contributed to a coherent depiction of frontostriatal function and testable predictions. As one example, we have identified mechanisms underlying two distinct forms of impulsivity. The first stems from a deficit in learning from "negative reward prediction errors" (when decision outcomes are worse than expected, dependent on dips in dopamine and their resultant effects on activity and plasticity in a subpopulation of striatal neurons expressing D2 dopamine receptors). Deficiencies in the mechanism lead to a failure to properly consider negative outcomes of prospective decisions, and hence lead to a bias to focus primarily on gains.

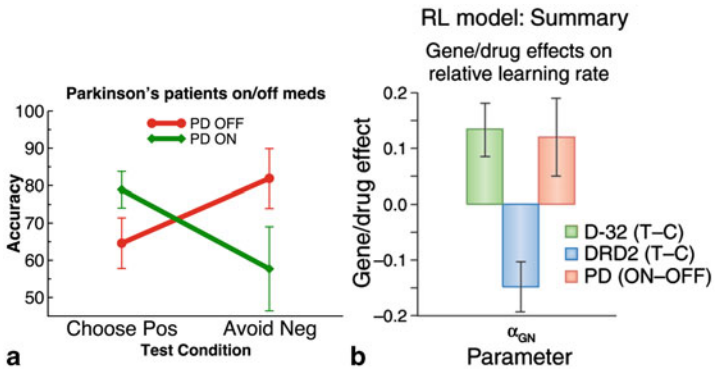


Fig. 8.2 Effects of **a** dopamine medication manipulations in Parkinson's disease [35] and **b** genotypes related to striatal D1 and D2 pathways in healthy participants on choice accuracy [22, 38]. "Choose Pos" assesses the ability to choose the probabilistically most rewarded (positive) action based on previous Go learning; "Avoid Neg" assesses the ability to avoid the probabilistically most punished (negative) action based on previous NoGo learning. **c** Quantitative fits with a reinforcement learning (RL) model capture these choice dissociations by assigning asymmetric Go vs. NoGo learning rates that vary as a function of PD, medications, and genotype [19, 31]. Unlike studies linking candidate genes to complex disease phenotypes, where findings often fail to replicate [55], this linking of neurogenetic markers of corticostriatal function to specific computational processes has been replicated across multiple experiments

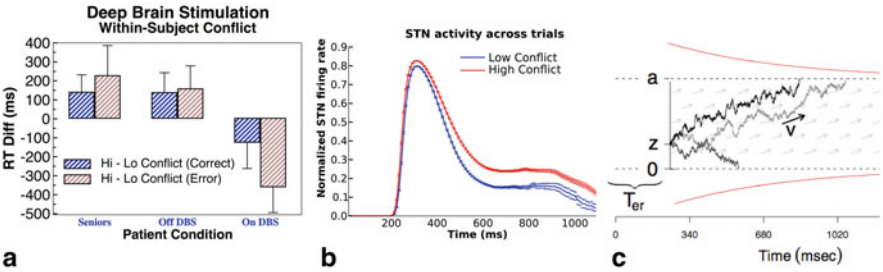


Fig. 8.3 **a** Decision conflict-induced response time slowing in PD patients and controls ("Seniors"). Hi Conflict: alternative actions have similar reward probabilities (high entropy). Lo Conflict: alternative actions have qualitatively different reward probabilities. DBS reverses conflict-induced slowing, leading to impulsive choice (large effect for suboptimal "error" choices) **b** STN firing rate in the neural model surges during action selection, to a greater extent during high conflict trials, delaying responding (not shown). **c** Two-choice decision making is captured by the drift diffusion model. Evidence accumulates for one option over the other from a starting point "z" at a particular average rate "v"; choices are made when this evidence crosses the decision threshold ("a"). Noisy accumulation leads to variability in response times (example RT distributions shown). Dynamics of STN function are captured in this framework by an increased decision threshold when conflict is detected (*red curve*), followed by a collapse to a static asymptotic value, similar to STN activity in (b); see [58]. Quantitative fits of the BG model with the DDM show that STN strength parametrically modulates decision threshold, as supported by experimental manipulations of STN and model-based fMRI studies showing STN activity varying with decision threshold estimated with the DDM [41]

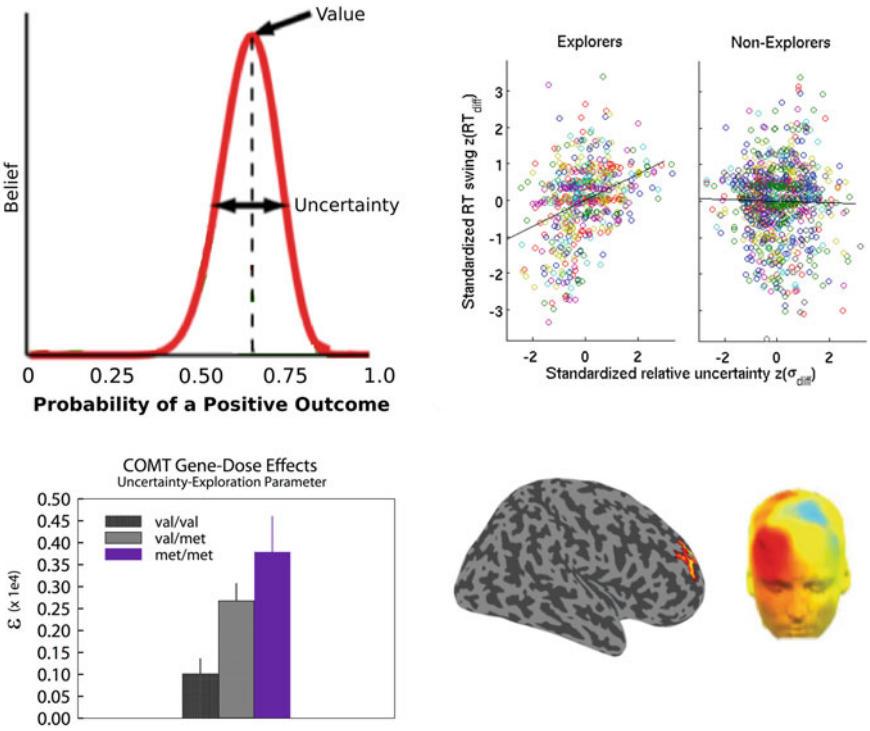


Fig. 8.4 The exploration-exploitation tradeoff. *Top*: probability density functions representing the degree of belief in values of two alternative actions. One action has a higher value estimate, but with uncertainty, quantified by Bayesian updating of belief distributions as a function of experience. Exploration is predicted to occur when alternative actions have relatively higher uncertainty. Approximately half of participants ("Explorers") employ this exploration strategy, with choice adjustments proportional to relative uncertainty. *Bottom*: The degree of uncertainty-driven exploration estimated by model parameter ε varies as a function of a genetic variant in COMT, associated with prefrontal dopamine [40]. fMRI data show that activity in RLPFC varies parametrically with relative uncertainty in Explorers [5, 14]. EEG topographical map shows theta-band activity correlates with relative uncertainty, maximal over rostrolateral electrodes

The second form of impulsivity stems from a failure to adaptively pause the decision process given conflicting evidence about the values of alternative actions (dependent on communication between frontal cortex and the subthalamic nucleus (STN), which temporarily provides a soft brake on motor output). Deficiencies in this mechanism lead to rash choices, i.e., a failure to consider the value of all the decision options but instead to quickly accept or reject an option based on its value alone. Notably these two forms of impulsivity are differentially impacted by distinct forms of treatment in Parkinson's disease –medications that produce elevations in dopamine and deep brain stimulation of the STN—supporting the posited mechanisms by which these treatments alter neural circuitry [13, 14, 38, 79]. Further insights come from linking these precise mechanisms to their high level computations. To do so, we next present

a more detailed overview of the neural model, followed by its linking to abstract computations.

8.2 Cortico-Striatal Interactions During Choice and Learning

The basal ganglia (BG) are a collection of subcortical structures that are anatomically, neurochemically, and functionally linked [1, 28, 43, 54]. Through a network of interconnected loops with the frontal cortex, they modulate motor, cognitive, and affective functions. The defining characteristic of this interaction is a "gating function": the frontal cortex first generates candidate options based on their prior history of execution in the sensory context and the BG facilitates selection [43] of one of these candidates given their relative learned reward values [26, 28].

Choice Two main projection pathways from the striatum go through different striatal nuclei on the way to thalamus and up to cortex. Activity in the *direct* "Go" pathway provides evidence in favor of facilitation of the candidate cortical action, by disinhibiting thalamocortical activity for the action with highest reward value. Conversely, activity in the *indirect* "*NoGo*" pathway indicates that the action is maladaptive and hence should not be gated. Thus for any given choice, direct pathway neurons convey the positive evidence in favor of that action based on learned reward history, whereas indirect pathway neurons signal the negative evidence (likelihood of leading to a negative outcome). The action most likely to be gated is a function of the relative difference in Go/NoGo activity for each action [26]. Dopamine influences the cost/benefit tradeoff by modulating the balance of activity in these pathways via differential effects on D1 and D2 receptors, thereby modulating choice incentive (whether choices are determined by positive or negative potential outcomes). Similar principles apply to the selection of cognitive actions (notably, working memory gating) in BG-prefrontal cortical circuits [26, 28, 34, 41]. Finally, conflict between competing choices, represented in mediofrontal/premotor cortices, activates the subthalamic nucleus (STN) via the *hyperdirect* pathway. In turn, STN activity delays action selection, making it more difficult for striatal Go signals to facilitate a choice and buying more time to settle on the optimal choice [27]; in abstract formulations, this is equivalent to a temporary elevation in the *decision threshold* [58] (see below; Fig. 8.3).

Learning The models simulate phasic changes in dopamine levels that occur during positive and negative *reward prediction errors* (difference between expected and obtained reward), and their effects on plasticity in the two striatal pathways. Phasic bursts of dopamine cell firing during positive prediction errors act as teaching signals that drive Go learning of rewarding behaviors via D1 receptor stimulation [26, 35, 63]. Conversely, negative prediction errors lead to pauses in dopamine firing [63], supporting NoGo learning to avoid unrewarding choices via D2 receptor disinhibition. An imbalance in learning or choice in these pathways can lead to a host of aberrant neurological and psychiatric symptoms [53].

8.3 Higher Level Descriptions

Thus far we have considered basic mechanisms in dynamical models of corticostriatal circuits and their resultant effects on behavior. However, these models are complex (cf. Fig. 8.1—this is the 'base' model and other variants build from there to include multiple circuits and their interactions). Moreover, because they simulate internal neural dynamics consistent with electrophysiological data, they require many more parameters than is necessary to relate to behavior alone. Higher level computational descriptions allow us to abstract away from the detailed implementation. In particular, the tendency to incrementally learn from reward prediction errors and to select among multiple candidate options has been fruitfully modeled using reinforcement learning (RL) models inherited from the computer science literature. These models summarize valuation of alternative actions, reflecting the contributions of the striatal units in the neural models, in terms of simple "Q values" which are incremented and decremented as a function of reward prediction errors. A simple choice function is used to compare Q values across all options and to stochastically select that with the highest predicted value: this function summarizes the effective computations of the gating circuitry that facilitates cortical actions (where noise in both cortex and striatum results in stochastic choice function). An asymmetry in learning from positive or negative prediction errors (reflecting high or low dopamine levels and their effects on activity/plasticity) can be captured by using separate learning rates (Frank et al. 2007) [22]. However, a better depiction of the neural model allows for not only differential modulation of learning, but also differential modulation of choice incentive during action selection [19]. This model uses separate Q values, QG and QN, to represent the Go and NoGo pathway respectively, each with their own learning rate, and where the 'activity', i.e. the current QG or QN value, further influences learning. The choice probability is then a function of the relative difference in QG and QN values for each decision option, with a gain factor that can differentially weigh influences of QG vs QN. This gain factor can be varied to simulate dopamine effects on choice incentive by boosting the extent to which decisions are made based on learned QG or QN values, even after learning has taken place.

Because these models are simple and minimal, they can be used to quantitatively fit behavioral data, and to determine whether the best fitting parameters vary as a function of biological manipulations. But because there is a clear mapping from these models to the neural versions, there are strong *a priori* reasons to manipulate particular neural mechanisms and to test whether the resulting estimates of computational parameters are altered as predicted or not.

Indeed, evidence validating these multi-level model mechanisms has mounted over the last decade across species. Monkey recordings combined with Q learning model fits indicate that separate populations of striatal cells code for positive and negative Q values associated with action facilitation and suppression [23, 51, 61, 71]. In mice, targeted manipulations confirm selective roles of direct and indirect pathways in the facilitation and suppression of behavior [49], which are necessary and sufficient to induce reward and punishment learning, respectively [44, 50]. Phasic

stimulation or inhibition of dopamine neurons induces reward/approach and aversive/avoidance learning, respectively [67, 68]. Synaptic plasticity studies reveal dual mechanisms for potentiation and depression in the two pathways as a function of D1 and D2 receptors [64], as in the models. In humans, striatal dopamine manipulation influences the degree to which individuals learn more from positive or negative outcomes (Fig. 8.2), with DA elevations enhancing reward learning but impairing punishment learning, and vice-versa for DA depletion [6, 33, 35, 57, 69] (Frank et al. 2007). Quantitative fits using RL models reveal that these can be accounted for by differential effects of dopamine manipulations on learning rates from positive and negative prediction errors. Moreover, in the absence of any acute manipulation, individual differences in these fit learning rate parameters are associated with genetic polymorphisms that differentially impact the efficacy of striatal D1 and D2 pathways [22, 31, 32, 40, 41] (Fig. 8.2).

One of the advantages of high level models, besides being simpler and more naturally used for quantitative behavioral fits, is that they can also include relevant processes that are out of scope in the neural versions. For example, when humans perform a "reinforcement learning task", they are not only incrementally learning probabilistic stimulus-action-outcome associations and choosing between them, but they also engage in other cognitive strategies involving hypothesis testing and working memory. Fitting their behavior with a RL model alone—no matter how well this model summarizes the corticostriatal learning process and its contribution to behavior—is then misleading, because it will capture variance that is really due to working memory capacity by absorbing this into the learning rate parameters of the RL process. Collins and Frank [17] showed clear evidence of such effects by manipulating the number of stimuli in the set to be learned (and hence working memory load). They found that when using RL models alone and without factoring in working memory, one needed to include a separate learning rate for each set size to capture the data, and that a gene related to prefrontal but not striatal function was predictive of this learning rate. However, when an augmented model which included a capacity-limited working memory process was used, the overall fits to the data were improved, and the RL process could be captured by a single learning rate that applies across all set sizes. Further, this learning rate in this best fit model varied with striatal genetic function, whereas the prefrontal gene was now related to working memory capacity.

On the other hand, algorithmic RL models that only predict choice probability miss out on the dynamics of choice, reflected in RT distributions, which emerge naturally from the neural model because it is a *process model*. First, firing rate noise throughout the network produces variance in the

speed with which an action is gated. Second, the action value of the candidate option impacts not only the likelihood of selecting that option relative to its competitors, but also the speed with which this option is selected. Finally, as mentioned above, when multiple candidate options have similar frequencies of execution based on their choice history—that is, when there is conflict or choice entropy—this elicits hyperdirect pathway activity from mediofrontal cortex to the STN, which provides a

temporary brake on the striatal gating process, thereby slowing down response time and increasing the likelihood in settling on the optimal response [27].

High level descriptions of process models have been extensively used to simulate dynamics of simple decision making in cognitive psychology for over 3 decades. In particular, the drift diffusion model (DDM) belongs to a class of sequential sampling models in which noisy evidence is accumulated in favor of one of two options, and a choice is executed once this evidence cross a critical *decision threshold*. The slope at which evidence accumulates is called the drift rate and reflects the ease of the decision. These models capture not only choice proportions and mean RT, but the entire shape of the RT distribution for correct and erroneous responses.

Notably, when fitting the behavioral outputs of the neural model with the DDM, we found that parametric manipulations of both corticostriatal and STN output projection strengths were related to estimated decision threshold, with corticostriatal strength decreasing threshold (see [24]) and STN strength increasing threshold [58, 74].

Studies with Parkinson's patients on and off STN deep brain stimulation provide an opportunity to test the impact of interference of the STN pathway, which can also lead to clinical impulsivity. Indeed, this procedure provides a selective disruption of conflict-induced slowing, without impacting learning [19, 39] (Fig. 8.3). We have extended this finding in three critical ways. First, EEG revealed that in healthy participants and patients off DBS, the amount of medial prefrontal (mPFC) theta-band activity during high conflict trials was predictive on a trial-to-trial basis of the amount of conflict-induced RT slowing. STN-DBS reversed this relationship, presumably by interfering with hyperdirect pathway function, without altering mPFC theta itself. Second, we developed a toolbox for hierarchical Bayesian parameter estimation allowing us to estimate the impact of trial-to-trial variations in neural activities on decision parameters [77]. We found that mPFC theta was predictive of decision threshold adjustments (and not other decision parameters), and, moreover, that DBS reversed this mPFC-threshold relationship [13, 14]. Third, electrophysiological recordings within STN revealed decision conflict-related activity in a similar time and frequency range as mPFC in both humans and monkeys [4, 13, 14, 41, 45, 46, 79]. These findings thus provide support for a computational account of hyperdirect pathway function, and a potential explanation for the observed impulsivity that can sometimes result from DBS.

Thus far we have considered the ability of existing abstract formulations to summarize the computations of more detailed neural models, providing a link between levels. It is also possible however, that aspects of the neural models, if valid, should alter the way we think about the abstract formulation. In the above example, we claimed that the STN was involved in regulating decision threshold. Consider its internal dynamics however (Fig. 8.3b). STN activity is not static throughout a trial, but rather exhibits an initial increase in activity, which then subsides with time during the action selection process. Moreover the initial STN surge is larger and more prolonged when there is higher decision conflict. This model dynamic is supported by electrophysiological evidence in both monkeys and humans [46, 79], and implies that STN effects on preventing BG gating should be transient and decrease with time, implying a collapsing rather than fixed decision threshold. Functionally this

collapsing threshold ensures that a decision is eventually made, preventing decision paralysis (this collapsing threshold is optimal when there are response deadlines; [42]). Indeed, quantitative fits using the DDM to capture RT distributions of the BG model showed that a collapsing threshold provided a good account of the model's behavior, notably, with the temporal dynamics of the best fitting exponentially collapsing threshold matching reasonably well to the dynamics of STN activity—despite the fact that the DDM fits had no access to this activity but only to RT distributions [58]. This study also found that when fitting human behavioral data in the same reward conflict decision-making task, fits were improved when assuming a higher and collapsing threshold in conflict trials, compared to the fixed threshold model.

This last result supports the assertion that neural mechanism constraints can be included to refine higher level descriptions. However, we must also admit that we do not have well constrained neural mechanistic models for all cognitive processes. The next example I turn to is the exploration-exploitation tradeoff in reinforcement learning, a process studied in machine learning for many years but only recently considered in the cognitive neurosciences.

8.4 Beyond Basic Mechanisms: Uncertainty Driven Exploration and Hierarchical Learning

Often individuals need to explore alternative courses of action to maximize potential gains. *But how does one know when to explore rather than exploit learned values?* Basic RL models usually assume a degree of random exploration, but a more efficient strategy is to keep track of the uncertainty about value estimates, and to guide exploration toward the action with higher uncertainty [20]. We have reported evidence for just such a mechanism, whereby trial-by-trial behavioral adjustments are quantitatively related to a Bayesian model estimate of relative outcome uncertainty. In this case, there is no existing neural model for how this relative uncertainty measure is encoded or updated as a function of reward experiences. Nevertheless, individual differences in the employment of this uncertainty-driven exploration strategy are predicted by genetic variations in the COMT (Catechol-O-methyltransferase) gene, which is related to prefrontal cortical dopamine function [40]. Further, a recent model-based fMRI study [5] revealed that the rostrolateral prefrontal cortex (RLPFC) parametrically tracks the relative uncertainty between outcome values, preferentially so in "Explorers" (defined based on behavioral fits alone). In EEG, relative uncertainty is reflected by variations in theta power over RLPFC (in contrast to the mPFC indices of conflict noted above), again preferentially in Explorers [13]. These converging data across modeling, behavior, EEG, genetics and fMRI indicate a potential prefrontal strategy for exploring and over-riding reward-based action selection in the BG. Notably, patients with schizophrenia, specifically those with anhedonia, exhibit profound reductions in uncertainty-driven exploration [65]. Thus this measure has potential relevance for understanding motivational alterations in clinical populations,

and motivates the development of mechanistic models of how relative uncertainty estimates are computed and updated in populations of prefrontal neurons.

As the field matures, it becomes less clear which level of modeling motivated the other—and this is a good thing, as mutual constraints become available. Collins and Frank [18] confronted the situation in which a learner has to decide whether, when entering a new context, the rules dictating links between states, actions and outcomes ("task-sets") should be re-used from those experienced in previous contexts, or whether instead a new task-set should be created and learned.

They developed a high level "context-task-set" (C-TS) computational model based on non-parametric Bayesian methods (Dirichlet process mixtures), describing how the learner can cluster contexts around task-set rules, generalizable to novel situations [18]. This model was motivated by analogous clustering models in category learning (e.g., [1, 62]), but applied to hierarchical cognitive control, and as such was similarly motivated by the hierarchical structure of prefrontal cortical basal ganglia networks and modeling implementations thereof [10, 29, 48, 60]. They also constructed a refined hierarchical PFC-BG network which confronted the same tasks, and showed that its functionality is well mimicked by the C-TS model. Quantitative model fitting linking these levels showed that particular neural mechanisms were associated with specific C-TS model parameters. For example, the prior tendency to re-use vs. create new structure in C-TS, captured by Dirichlet alpha parameter, was directly related to the sparseness of the connectivity matrix from contextual input to PFC (Fig. 8.5). Thus in this case, there existed well established and validated models of interactions between PFC and BG during learning, working memory, and action selection (including some hierarchical implementations), but the computations afforded by the novel C-TS model further inspired refinement and elaboration of the network. In turn, this exercise reciprocally allowed us to derive more specific predictions about mechanisms leading to differential response times and error patterns (which were confirmed behaviorally), and to marry the reinforcement learning models described previously with the cognitive control mechanisms involving decision threshold regulation.

One novel finding from this modeling work was that in such environments, the STN mechanism, previously linked only to decision making and impulsivity, plays a key role in learning. In particular, the simulations showed that early in the trial, when there is uncertainty about the identity of the PFC task-set, this conflict between alternative PFC states activated the STN, preventing the motor loop from responding. This process ensures that the PFC state is resolved prior to motor action selection, and as such, when the outcome arrives, stimulus-action learning is conditionalized by the selected PFC state. As the STN contribution is reduced, there is increasing interference in learning across task-sets, hence learning is less efficient. This novel theory specifying the role of the STN in conditionalizing learning by PFC state needs to be tested empirically (e.g. with DBS or fMRI), but each component is grounded by prior empirical and theoretical work, yet it would not likely have emerged without this multi-level modeling endeavor. In related work, Frank and Badre [29] considered hierarchical learning tasks with multidimensional stimuli with two levels of modeling. A Bayesian mixture of experts model summarized how participants may learn hierarchical structure of the type, "if the color is red, then the response is determined

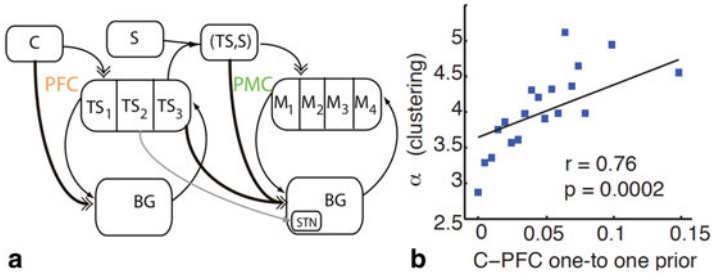


Fig. 8.5 *Left*: Schematic of hierarchical corticostriatal network model for creating task-sets (TS) which are gated into prefrontal cortex depending on the context C. The lower motor loop selects motor actions M depending on the selected TS in PFC and the current sensory state S. The same TS can be reused across contexts, supporting clustering and generalization of behaviors, or if needed, a new TS can be gated, preventing interference in learned state-action mappings between different contexts/TS. *Right*: Parametric manipulation of the sparseness of the connectivity matrix from Context to PFC (enforcing a prior tendency to encode distinct C's as distinct PFC TS) is well fit by an increased α Dirichlet process clustering parameter in the C-TS model which creates and re-uses TS according to non-parametric Bayesian methods. (Adapted from Collins and Frank [18])

by the shape, whereas if the color is blue, the response is determined by the orientation." Quantitative modeling showed that estimated attention to the hierarchical expert was linked to speeded learning in hierarchical conditions, and when fit to a PFC-BG network, was related to a measure of gating policy abstraction, learned via RL, in the hierarchical connections from PFC to striatum. Badre and Frank (2012) then used model-based fMRI to show that in participants, estimated attention to hierarchical structure was linked to PFC-BG activity within a particular rostrocaudal level of the network consistent with the "second-order" rule level of the task.

8.5 Concluding Comments

The examples described above demonstrate mutual, reciprocal constraints between models of neural circuitry and physiology to models of computational function. This exercise leads to multiple testable predictions using model-based cognitive neuroscience methods. Ultimately, models are judged based on their predictive power, and as such, they can inspire informative experiments valuable even to those who question the validity or assumptions of either of the levels of modeling employed.

Exercises

1. Give examples of implementational neural models and higher level algorithmic models in any domain. What sorts of data do these models attempt to capture?
2. Think of some examples in which an abstract model exists but would benefit from a mechanistic elaboration for making cognitive neuroscience predictions.

3. Can you think of potential advantages of combining the models? How about some pitfalls?
4. Describe how dopamine may contribute both to learning and to choice incentive (the degree to which decisions are made based on positive vs negative consequences).
5. Reinforcement learning models and sequential sampling models of decision making have been largely separate literatures in mathematical psychology yet each of these classes of models have been fit to the basal ganglia neural model described in this chapter. Read Bogacz and Larsen [8] for a complementary approach to linking these formulations within an algorithmic framework.
6. Conversely, read Wong and Wang [78] for a complementary example of a single neural model capturing dynamics of decision making and working memory.

Further Reading

1. Collins and Frank [18] present two levels of modeling describing the interactions between cognitive control and learning needed to construct task-set rules generalizable to novel situations. This endeavor reaps the benefits of both RL models and the temporal dynamics of decision making, and how each affects the other. It also shows theoretically how a non-parametric Bayesian approach to task-set clustering can be implemented in hierarchical PFC-BG circuitry.
2. Wang [70] reviews neural models of decision making and their relation to normative theory.
3. Brittain et al. [12] present evidence for STN involvement in deferred choice under response conflict in a non-reward based task, complementing findings described in this chapter.

References

1. Alexander GE, Crutcher MD (1990) Functional architecture of basal ganglia circuits: neural substrates of parallel processing. Trends Neurosci 13(7):266–271
2. Anderson JR (1991) The adaptive nature of human categorization. Psychol Rev 98(3):409–429
3. Aron AR, Behrens TE, Smith S, Frank MJ, Poldrack R (2007) Triangulating a cognitive control network using diffusion-weighted magnetic resonance imaging (MRI) and functional MRI. J Neurosci 27(14):3743–3752
4. Badre D, Frank MJ (2012) Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 2: Evidence from fMRI. Cerebral Cortex 22:527–536
5. Badre D, Doll BB, Long NM, Frank MJ (2012) Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration. Neuron 73:595–607
6. Bódi N, Kéri S, Nagy H, Moustafa A, Myers CE, Daw N, Dibó G, Takáts A, Bereczki D, Gluck MA (2009). Reward-learning and the novelty-seeking personality: a between- and within-subjects study of the effects of dopamine agonists on young Parkinson's patients. Brain 132:2385–2395

7. Bogacz R, Gurney K (2007) The basal ganglia and cortex implement optimal decision making between alternative actions. Neural Comput 19(2):442–477
8. Bogacz R, Larsen T (2011) Integration of reinforcement learning and optimal decision-making theories of the basal ganglia. Neural Comput 23(4):817–851
9. Bogacz R, Wagenmaker EJ, Forstmann BU, Nieuwenhuis S (2010) The neural basis of the speed-accuracy tradeoff. Trends Neurosci 33(1):10–16
10. Botvinick MM, Niv Y, Barto AC (2009) Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. Cognition 113(3):262–280
11. Braver TS, Cohen JD (2000) On the control of control: the role of dopamine in regulating pre- frontal function and working memory. In: Monsell S, Driver J (eds) Control of cognitive processes: attention and performance XVIII. MIT Press, Cambridge, pp 713–737
12. Brittain JS, Watkins KE, Joundi RA, Ray NJ, Holland P, Green AL, Aziz TJ, Jenkinson N (2012) A role for the subthalamic nucleus in response inhibition during conflict. J Neurosci 32(39):13396–13401
13. Cavanagh JF, Frank MJ, Klein TJ, Allen JJB (2010) Frontal theta links prediction error to behavioral adaptation in reinforcement learning. Neuroimage 49(4):3198–3209
14. Cavanagh JF, Figueroa CM, Cohen MX, Frank MJ (2011a) Frontal theta reflects uncertainty and unexpectedness during exploration and exploitation. Cereb Cortex 22(11):2575–2586
15. Cavanagh JF, Wiecki TV, Cohen MX, Figueroa CM, Samanta J, Sherman SJ, Frank MJ (2011b) Subthalamic nucleus stimulation reverses mediofrontal influence over decision threshold. Nat Neurosci 14(11):1462–1467
16. Collins AGE, Frank MJ (2012) How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. Eur J Neurosci 35(7):1024–1035
17. Collins AGE, Frank MJ (2013) Cognitive control over learning: creating, clustering and generalizing task-set structure. Psychol Rev 120(1):190–229
18. Collins AGE, Frank MJ (2013) Opponent Actor Learning (OpAL): modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. Psychol Rev 121:337–366
19. Coulthard EJ, Bogacz R, Javed S, Mooney LK, Murphy G, Keeley S, Whone AL (2012). Distinct roles of dopamine and subthalamic nucleus in learning and probabilistic decision making. Brain 135:3721–3734
20. Dayan P, Sejnowksi T (1996) Exploration bonuses and dual control. Mach Learn 25:5–22
21. Doll BB, Jacobs WJ, Sanfey AG, Frank MJ (2009) Instructional control of reinforcement learning: a behavioral and neurocomputational investigation. Brain Res 1299:74–94
22. Doll BB, Hutchison KE, Frank MJ (2011) Dopaminergic genes predict individual differences in susceptibility to confirmation bias. J Neurosci 31(16):6188–6198
23. Ford KA, Everling S (2009) Neural activity in primate caudate nucleus associated with pro- and antisaccades. J Neurophysiol 102(4):2334–2341
24. Forstmann BU, Dutilh G, Brown S, Neumann J, von Cramon DY, Ridderinkhof KR, Wagenmakers EJ (2008a) Striatum and pre-SMA facilitate decision-making under time pressure. Proc Natl Acad Sci USA 105(45):17538–17542
25. Forstmann BU, Jahfari S, Scholte HS, Wolfensteller U, van den Wildenberg WP, Ridderinkhof KR (2008b) Function and structure of the right inferior frontal cortex predict individual differences in response inhibition: a model-based approach. J Neurosci 28(39):9790–9796
26. Frank MJ (2005) Dynamic dopamine modulation in the basal ganglia: a neurocomputational account of cognitive deficits in medicated and nonmedicated Parkinsonism. J Cogn Neurosci 17(1):51–72
27. Frank MJ (2006) Hold your horses: a dynamic computational role for the subthalamic nucleus in decision making. Neural Netw 19(8):1120–1136
28. Frank MJ (2011) Computational models of motivated action selection in corticostriatal circuits. Curr Opin Neurobiol 2:381–386
29. Frank MJ, Badre D (2012) Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: computational analysis. Cereb Cortex 22(3):509–526

30. Frank MJ, Claus ED (2006) Anatomy of a decision: striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. Psychol Rev 113(2):300–326
31. Frank MJ, Fossella JA (2011) Neurogenetics and pharmacology of learning, motivation, and cognition. Neuropsychopharmacology 36:133–152
32. Frank MJ, Hutchison K (2009) Genetic contributions to avoidance-based decisions: striatal D2 receptor polymorphisms. Neuroscience 164(1):131–140
33. Frank MJ, O'Reilly RC (2006) A mechanistic account of striatal dopamine function in human cognition: psychopharmacological studies with cabergoline and haloperidol. Behav Neurosci 120(3):497–517
34. Frank MJ, Loughry B, O'Reilly RC (2001) Interactions between frontal cortex and basal ganglia in working memory: a computational model. Cogn Affect Behav Neurosci 1(2):137–160
35. Frank MJ, Seeberger LC, O'Reilly RC (2004) By carrot or by stick: cognitive reinforcement learning in parkinsonism. Science 306(5703):1940–1943
36. Frank MJ, Santamaria A, O'Reilly R, Willcutt E (2007a) Testing computational models of dopamine and noradrenaline dysfunction in attention deficit/hyperactivity disorder. Neuropsychopharmacology 32(7):1583–1599
37. Frank MJ, D'Lauro C, Curran T (2007b) Cross-task individual differences in error processing: neural, electrophysiological, and genetic components. Cogn Affect Behav Neurosci 7(4):297–308
38. Frank MJ, Moustafa AA, Haughey H, Curran T, Hutchison K (2007c) Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. Proc Natl Acad Sci 104(41):16311–16316
39. Frank MJ, Samanta J, Moustafa AA, Sherman SJ (2007d) Hold your horses: impulsivity, deep brain stimulation and medication in Parkinsonism. Science 318:1309–1312
40. Frank MJ, Doll BB, Oas-Terpstra J, Moreno F (2009). Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. Nat Neurosci 12(8):1062–1068
41. Frank MJ, Gagne C, Nyhus E, Masters S, Wiecki TV, Cavanagh JF, Badre D (2015) fMRI and EEG Predictors of dynamic decision parameters during human reinforcement learning. J Neurosci 35:484-494
42. Frazier P, Yu AJ (2008) Sequential hypothesis testing under stochastic deadlines. Adv Neural Inf Process Syst 20:465–472. (MIT Press, Cambridge)
43. Gurney K, Prescott TJ, Redgrave P (2001) A computational model of action selection in the basal ganglia. II. Analysis and simulation of behaviour. Biol Cybern 84(6):411–423
44. Hikida T, Kimura K, Wada N, Funabiki K, Nakanishi S (2010) Distinct roles of synaptic transmission in direct and indirect striatal pathways to reward and aversive behavior. Neuron 66(6):896–907
45. Isoda M, Hikosaka O (2007) Switching from automatic to controlled action by monkey medial frontal cortex. Nat Neurosci 10(2):240–248
46. Isoda M, Hikosaka O (2008) Role for subthalamic nucleus neurons in switching from automatic to controlled eye movement. J Neurosci 28(28):7209–7218
47. Jahfari S, Verbruggen F, Frank MJ, Waldorp LJ, Colzato L, Ridderinkhof KR, Forstmann BU (2012) How preparation changes the need for top-down control of the basal ganglia when inhibiting premature actions. J Neurosci 32(32):10870–10878
48. Koechlin E, Summerfield C (2007) An information theoretical approach to the prefrontal executive function. Trends Cogn Sci 11(6):229–235
49. Kravitz AV, Freeze BS, Parker PRL, Kay K, Thwin MT, Deisseroth K, Kreitzer AC (2010) Regulation of parkinsonian motor behaviours by optogenetic control of basal ganglia circuitry. Nature 466(7306):622–626
50. Kravitz AV, Tye LD, Kreitzer AC (2012) Distinct roles for direct and indirect pathway striatal neurons in reinforcement. Nat Neurosci 15:816–818
51. Lau B, Glimcher PW (2008) Value representations in the primate striatum during matching behavior. Neuron 58(3):451–463

52. Lo CC, Wang XJ (2006) Cortico-basal ganglia circuit mechanism for a decision threshold in reaction time tasks. Nat Neurosci 9(7):956–963
53. Maia TV, Frank MJ (2011) From reinforcement learning models to psychiatric and neurological disorders. Nat Neurosci 2:154–162
54. Mink JW (1996) The basal ganglia: focused selection and inhibition of competing motor programs. Prog Neurobiol 50(4):381–425
55. Munafò MR, Stothart G, Flint J (2009) Bias in genetic association studies and impact factor. Mol Psychiatry 14:119–120
56. O'Reilly RC, McClelland JL (1994) Hippocampal conjunctive encoding, storage, and recall: avoiding a trade-off. Hippocampus 4(6):661–682
57. Pessiglione M, Seymour B, Flandin G, Dolan RJ, Frith CD (2006) Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. Nature 442(7106):1042–1045
58. Ratcliff R, Frank MJ (2012) Reinforcement-based decision making in corticostriatal circuits: mutual constraints by neurocomputational and diffusion models. Neural Comput 24:1186–1229
59. Ratcliff R, McKoon G (2008) The diffusion decision model: theory and data for two-choice decision tasks. Neural Comput 20(4):873–922
60. Reynolds JR, O'Reilly RC (2009) Developing PFC representations using reinforcement learning. Cognition 113(3):281–292
61. Samejima K, Ueda Y, Doya K, Kimura M (2005) Representation of action-specific reward values in the striatum. Science 310(5752):1337–1340
62. Sanborn AN, Griffiths TL, Navarro DJ (2010) Rational approximations to rational models: alternative algorithms for category learning. Psychol Rev 117(4):1144–1167
63. Schultz W (2002) Getting formal with dopamine and reward. Neuron 36(2):241–263
64. Shen W, Flajolet M, Greengard P, Surmeier DJ (2008) Dichotomous dopaminergic control of striatal synaptic plasticity. Science 321(5890):848–851
65. Strauss GP, Frank MJ, Waltz JA, Kasanova Z, Herbener ES, Gold JM (2011) Deficits in positive reinforcement learning and uncertainty-driven exploration are associated with distinct aspects of negative symptoms in schizophrenia. Biol Psychiatry 69:424–431
66. Tai LH, Lee AM, Benavidez N, Bonci A, Wilbrecht L (2012) Transient stimulation of distinct subpopulations of striatal neurons mimics changes in action value. Nat Neurosci 15:1281–1289
67. Tan KR, Yvon C, Turiault M, Mirabekov JJ, Doehner J, Labouèbe G, Deisseroth K, Tye KM, Lüscher C (2012) GABA neurons of the VTA drive conditioned place aversion. Neuron 73:1173–1183
68. Tsai HC, Zhang F, Adamantidis A, Stuber GD, Bonci A, de Lecea L, Deisseroth K (2009) Phasic firing in dopaminergic neurons is sufficient for behavioral conditioning. Science 324:1080–1084
69. Voon V, Pessiglione M, Brezing C, Gallea C, Fernandez HH, Dolan RJ, Hallett M (2010) Mechanisms underlying dopamine-mediated reward bias in compulsive behaviors. Neuron 65(1):135–142
70. Wang XJ (2012) Neural dynamics and circuit mechanisms of decision-making. Curr Opin Neurobiol 22:1039–1046
71. Watanabe M, Munoz DP (2009) Neural correlates of conflict resolution between automatic and volitional actions by basal ganglia. Eur J Neurosci 30(11):2165–2176
72. Watkins CJCH, Dayan P (1992) Q-Learning. Mach Learn 8:279–292
73. Wiecki TV, Frank MJ (2010) Neurocomputational models of motor and cognitive deficits in Parkinson's disease. Prog Brain Res 183:275–297
74. Wiecki TV, Frank MJ (in press). A computational model of inhibitory control in frontal cortex and basal ganglia. Psychological Review.
75. Wiecki TV, Riedinger K, Meyerhofer A, Schmidt W, Frank MJ (2009) A neurocomputational account of catalepsy sensitization induced by D2 receptor blockade in rats: context dependency, extinction, and renewal. Psychopharmacology (Berl) 204:265–277
76. Wiecki TV, Sofer I, Frank MJ (2012). Hierarchical Bayesian parameter estimation of Drift Diffusion Models (Version 0.4RC1) [software]. http://ski.clps.brown.edu/hddm_docs/.

77. Wiecki TV, Sofer I, Frank MJ (2013) HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. Fron Neuroinformatics 7:1–10
78. Wong KF, Wang XJ (2006) A recurrent network mechanism of time integration in perceptual decisions. J Neurosci 26(4):1314–1328
79. Zaghloul K, Weidemann CT, Lega BC, Jaggi JL, Baltuch GH, Kahana MJ (2012) Neuronal activity in the human subthalamic nucleus encodes decision conflict during action selection. J Neurosci 32(7):2453–2460

Chapter 9
Bayesian Models in Cognitive Neuroscience: A Tutorial

Jill X. O'Reilly and Rogier B. Mars

Abstract This chapter provides an introduction to Bayesian models and their application in cognitive neuroscience. The central feature of Bayesian models, as opposed to other classes of models, is that Bayesian models represent the beliefs of an observer as probability distributions, allowing them to integrate information while taking its uncertainty into account. In the chapter, we will consider how the probabilistic nature of Bayesian models makes them particularly useful in cognitive neuroscience. We will consider two types of tasks in which we believe a Bayesian approach is useful: optimal integration of evidence from different sources, and the development of beliefs about the environment given limited information (such as during learning). We will develop some detailed examples of Bayesian models to give the reader a taste of how the models are constructed and what insights they may be able to offer about participants' behavior and brain activity.

9.1 Introduction

In the second half of the eighteenth century, the French mathematician Pierre-Simon Laplace was confronting a dilemma. He wanted to use observations of the location of planets to test the predictions made recently by Isaac Newton about the motion of heavenly bodies and the stability of the solar system. However, the data Laplace was confronted with was cobbled together from sources all over the word and some of it was centuries old. Couple that with the imprecision of the instruments of the time and Laplace had what we now call noisy data on his hands.

J. X. O'Reilly (✉) · R. B. Mars
Centre for Functional MRI of the Brain (FMRIB), Nuffield Department of Clinical Neurosciences, John Radcliffe Hospital, University of Oxford, Oxford OX3 9DU, UK
e-mail: joreilly@fmrib.ox.ac.uk

Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, 6500 HB, Nijmegen, The Netherlands

R. B. Mars
Department of Experimental Psychology, University of Oxford, South Parks Road, Oxford OX1 3UD, UK
e-mail: rogier.mars@psy.ox.ac.uk

B. U. Forstmann, E.-J. Wagenmakers (eds.), *An Introduction to Model-Based Cognitive Neuroscience,* DOI 10.1007/978-1-4939-2236-9_9

Laplace decided he needed a method that would allow him to use the large amounts of data obtained by astronomers, some of which might be unreliable, to determine the real state of the universe they were observing. In other words, he needed a way to move back from observed events (the astronomers' observations) to the most probable cause (the position of a planet). In doing so he created a way of thinking fundamentally different from the established approaches at the time. Because Laplace unwittingly hit upon elements of an earlier work by the English Reverend Thomas Bayes [1], we now know this way of thinking as 'Bayesian'.

The Bayesian way of thinking is so different from conventional statistics that it was 'non grata' in most university departments for a long time. Only since the mid-twentieth century has this begun to change. Bayesian methods were starting to be applied pragmatically to solve a host of real-world problems. Moreover, the invention of computers enabled people to perform the often labor intensive computations required in Bayesian statistics automatically. Slowly, Bayesians dared to come out of the closet (see McGrayne [14] for a history of Bayesian thinking). Bayesian thinking has now been applied to almost every field imaginable, including code-breaking, weather prediction, improving the safety of coal mines, and—most relevant for the purpose of this book—the modeling of human behavior and human brain function.

This chapter is about the use of Bayesian models in cognitive neuroscience and psychology, with particular reference to the modeling of beliefs and behavior. The reason for using formal models in this context is to gain insight into internal representations of the environment and of experimental tasks that are held by participants, and to use them to predict behavior and brain activity. We will therefore begin by explaining how the representation of the world contained in a Bayesian model (or brain) differs from non-Bayesian representations, and go on to consider how these features can be used in the context of cognitive neuroscience and psychology research.

We will first discuss three key features of Bayesian system: (1) Bayesian systems represent beliefs as probability distributions, (2) Bayesian systems weight different sources of information according to their associated uncertainty, and (3) Bayesian systems interpret new observations in the light of prior knowledge.

After considering how a Bayesian model's worldview differs from that of a non-Bayesian model, we will briefly review some evidence from the psychology and neuroscience literature that human and animal observers behave in a Bayesian manner. In particular we will focus on two classes of problems in which Bayesian models behave differently from non-Bayesian ones: integration of sensory evidence from different sources, and learning.

In the final section of the chapter, we will look in more detail at how Bayesian models can be constructed and what insights can be gained from them. We will consider Bayesian approaches to two problems: inferring a spatial distribution from a few observations, and inferring the probability of targets or rewards appearing in one of two locations in a gambling task. By constructing Bayesian 'computer participants' for each of these tasks, we will gain insights into factors that might predict the performance of human or animal participants on the same tasks.

9.2 The defining features of a Bayesian model

Bayesian statistics is a framework for making inferences about the underlying state of the world, based on observations and prior beliefs. The Bayesian approach, to try and infer *causes* from their observed *effects,* differs philosophically from other approaches to data analysis. Other approaches, often referred to as 'frequentist' approaches, focus on obtaining summary statistics for the observed data (such as the mean or expected value of an observation) without reference to the underlying causes that generated the data.

9.2.1 Bayesian Systems Represent Beliefs as Probability Distributions

A Bayesian approach to understanding data is to consider a range of possible causes for the observed data, and assign probabilities to each of them. A subtle but crucial consequence of this approach is that, although the true state of the environment takes a single value, the observer's idea of the environment can be represented as a distribution over many possible states of the environment. In other words, even though the observer knows there is only one true cause of his observations, he can still assign a graded probability to several possible causes. The observer's model represents these possibilities as a *probability density function* (pdf). This single feature, the representation of beliefs as probability density functions, gives rise to much of the behavior that differentiates Bayesian models from non-Bayesian ones.

Let's illustrate the use of probability density functions with an example. Consider the following scenario: Isaac Newton is foraging for apples in his garden when he sees an apple fall from a tree into long grass (Fig. 9.1a). Where should he go to retrieve the apple? If he saw the apple fall into the undergrowth, then the most likely place to look for the apple is near where it fell. We might, therefore, represent his belief about the location of the apple (his *internal model* of the state of the environment) as a single value, the location at which the apple entered the undergrowth (Fig. 9.1b; for simplicity, let's assume we can represent the location in a one-dimensional space). However, because the apple is now out of sight, Isaac can't be certain exactly where it is (it may have rolled along the ground in an unknown direction). This uncertainty can be incorporated into his internal model, if instead of using a single value the apple's position is represented as a probability distribution. Then we can make statements like 'there is a 95 % chance that there will be an apple between locations A and B' (Fig. 9.1c). Note that as well as the most likely location of the apple (the model of the distribution) this representation captures uncertainty (the width or variance of the distribution).

Note that the Bayesian use of probability density functions to represent degree of belief about a single state of the world is rather distinct from the typical use of probability density functions to represent the frequency of observations. In our apple

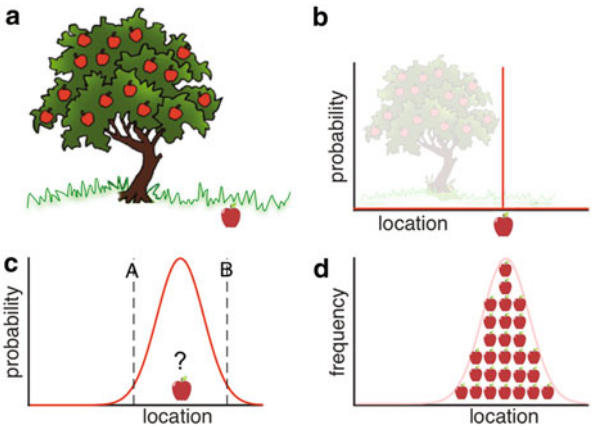


Fig. 9.1 **a** An apple falls from a tree into undergrowth. Where does the apple end up? **b** The single most likely location for the apple. **c** A Bayesian probabilistic representation: beliefs about the location of a single apple are represented as a probability distribution. **d** The equivalent frequentist concept: if I repeated the experiment many times, how often would an apple end up in each location?

example, Isaac Newton knows a *single* apple fell from the tree, and represents the location of that *single* apple as a probability density function, although in fact there is only one apple and it has only one true location. A more typical (frequentist) construction of a probability density function would be to represent the *frequency* with which apples were observed in different locations (Fig. 9.1d). Whilst for a hundred apples, the frequentist and Bayesian pdfs may look the same, for a single apple, the frequentist view is that the apple is either in a position, or it is not.

9.2.2 Bayesian Systems Integrate Information Using Uncertainty

When a belief about the state of the world (for example, about the location of an apple) is represented as a probability density function, the variance of that pdf, in other words the width of the pdf, represents the degree of uncertainty about the state of the world. One key feature of the Bayesian approach is that Bayesian systems take this uncertainty into account and use it to weight different sources of information according to their relative precisions.

Imagine that Isaac Newton has both *seen* an apple fall from the tree into long grass, and *heard* it hit the ground. His belief about the location of the apple based on each of these sources of information (vision and hearing) can be represented as a single probability density function. How should Isaac's brain use these two sources of information to get the best estimate of the apple's location? One solution is to use only the more reliable, or preferred sense. But this wastes the information from the other sense. A better solution is to combine the estimates of location based on vision and hearing.

How should the two sensory modalities be combined? Perhaps Isaac could take a point midway between the most likely location given what he saw, and the most likely location given what he heard? The Bayesian solution to this problem is to apply *precision weighting* the two sources of information, that is to give more weight to

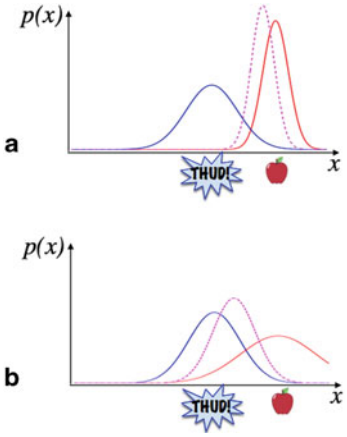


Fig. 9.2 Multisensory integration. The red line represents the probability distribution based on Isaac's visual information, the blue line represents the probability distribution based on hearing. The dotted line represents the combined distribution

the observation with the lowest variance. If, for example, vision gives a more precise estimate of where the apple fell, then visual evidence should be given more weight. On the other hand, if vision is unreliable (e.g. at night), auditory evidence should be given more weight.

Let's look at this graphically. In Fig. 9.2a, we can see that the pdf of Isaac's visual information (in red) is much less wide than his pdf based on hearing (in blue). Or, to put it more precisely, the variance of the vision pdf is smaller than that of the hearing pdf. Thus, optimally combining these two sources of information will result in a pdf closer to the vision pdf. However, in Fig. 9.2b, the vision is much less reliable, indicated by a greater variance in the red pdf. The combined pdf now is much closer to the hearing one.

Precision weighting is only possible if the observations (by vision and hearing) are represented as probability density functions; if each observation was represented in terms of a single most likely location, we could still combine the predictions by taking a point between the locations given vision and hearing, but there would be no way to take into account the relative reliability or precision of the two sources of information. However, given that observations are represented as probability density functions, precision weighting arises naturally from simply multiplying together the two probability distributions[1]. Then the probability of the apple's location given both visual and auditory information is highest where the two distributions overlap, and the mode (peak) of the combined distribution lies closer to the mode of the distribution with the lowest variance.

[1] In fact, the probability of each location given hearing and vision can only be obtained by multiplication if the variance in the two probability density functions is independent. In this case, we are talking about uncertainty that arises from noise in the sensory systems, which we can safely assume is independent between vision and hearing.

9.2.3 Bayesian Systems Interpret New Information in the Light of Prior Knowledge

Isaac Newton probably had some previous experience with apples falling from trees. Therefore, it would seem sensible if he used this prior knowledge to inform his model of where the apple might lie. For example, he might have some expectations about how far the apple might roll, the slope of the land, etc. Even if Isaac didn't see an apple fall, he would still have a prior belief that apples should be found under the apple tree—not, for example, under the lamppost. Isaac knows the apple should not fall far from the tree.

In the same way the location of the apple, given where Isaac saw it fall, can be represented as a probability density function, so can his prior beliefs. In Bayesian thinking these prior beliefs are called the *priors*. Furthermore, current observations can be combined with the prior, just as probability density functions based on vision and hearing were combined in the previous section. Combining the current observations with the prior gives a *posterior* distribution that takes both into account.

The ability to combine current observations with a prior, or to combine parallel sources of information like vision or hearing, is embodied in the central theorem of Bayesian statistics, called *Bayes' theorem*:

$$p(\textit{apple location}|\textit{observation}) \propto p(\textit{observation}|\textit{apple location}) \times p(\textit{apple location}) \tag{9.1}$$

... where p(apple location) is defined as the probability that a given hypothetical state of the environment (such as a given location for a planet or an apple) was true, based on all sources of information other than the observation currently being considered. The term 'other sources of information' can equally well include other sensory modalities or prior knowledge.

In Bayesian terminology, the left hand side of Eq. 9.1, p(apple location | observation), is called the *posterior*; the expression p(observation | apple location) is called the *likelihood*; and p(apple location) is called the *prior*. Bayes' theorem thus says that our belief about the true state of the environment after our observations is proportional to our prior beliefs weighted by the current evidence.

9.2.4 Priors and Learning

Because Bayes' theorem tells us how we should combine new observations with prior beliefs, it provides particularly useful insights about how the observer's beliefs should evolve in situations where information about the environment is obtained sequentially. For example, we can model how Isaac's beliefs evolve while he observes a number of falling apples. After each observation, he updates his prior to a new posterior. This posterior then serves as the new prior for the next apple.

In experimental paradigms in which participants learn by trial and error, we cannot assume the observer has complete knowledge of the state of the environment. These

paradigms are a key target for model-based cognitive neuroscience, since if we want to model a participant's behavior or brain activity, it is arguably more appropriate to base our predictions on a model of what the participant might *believe* the state of the environment to be, rather than basing our predictions about brain activity on the true state of the environment, which the participant could not in fact know, unless he/she/it was clairvoyant.

Of course, not all learning models are Bayesian—for example, temporal-difference learning models such as the Rescorla-Wagner algorithm are also popular. Non-Bayesian algorithms can do a good job of explaining behavior in many experiments. In a later section of the tutorial we will investigate the differences between Bayesian and non-Bayesian learning algorithms in more detail, in order to highlight cases where Bayesian models can give us enhanced insights into the participant's thought processes as compared to non-Bayesian learning algorithms.

9.3 Are Bayesian models valid for modeling behavior?

In the previous section we've seen how Bayesian thinking can be used to model the beliefs of an observer and can track how these beliefs should evolve when combining different sources of information or during learning based on repeated observations. Mathematically, it can be shown that the Bayesian approach is the best approach to combine information under uncertainty with the greatest precision [5]. However, for these models to be useful in cognitive neuroscience we need to know if people combine information in similar ways. Fortunately, it turns out they often do. People and animals can show behavior close to the optimum predicted by Bayesian theory. In this section, we will provide some examples of how human behavior can be described by Bayesian models. We will limit ourselves to illustrating how human behavior shows some of the Bayesian characteristics we described above. More in-depth reviews of how the Bayesian approach can inform our understanding of behavior and brain function are provided by O'Reilly et al. [17], Chater and Oaksford [3], and Körding and Wolpert [11].

At the most fundamental level, one can see the human brain as a device whose job it is, at least partly, to infer the state of the world. However, we know that the nervous system is noisy. Thus we need to deal with information under uncertainty. One way that psychologists have suggested we deal with this is the use of 'top-down information'. In the terms of Bayesian theory this means people have a prior that influences their information processing. The effect of such a prior has been demonstrated in vision by the existence of a variety of well-known visual illusions. For instance, in the famous Müller-Leyer illusion people see two line segments of equal length that have short lines on their ends, either pointing in the direction of the line or away from it. Most people report the second line to be longer than the first. Gregory [8] suggested this is because people have priors about perspective that they have learned from the buildings in the environment, in which the former configuration corresponds to an object which is closer and the latter with an object

far away. Interestingly, this predicts that people who have grown up in a different environment might not have this illusion. This indeed seems to be the case for some African tribes [22].

In our daily life, we often have to reconcile different sources of information. As shown above, Bayesian thinking implies that different sources of information should be combined using precision weighting. As one illustration of whether humans combine information in this way, Jacobs [9] asked participants to match the height of an ellipse to the depths of a simulated cylinder defined by texture and motion cues. The participants were either given motion information, texture information, or both about the depth of the cylinder. Bayesian models predicted how participants combined the sources of information. Similarly, Ernst and Banks [7] asked participants to combine visual and haptic information to judge the height of a raised bar. Participants were given conflicting information with the experimenter manipulating the precision of the information available by introducing noise in the visual stimulus. They reported that participants took the reliability of the visual information into account when combining the visual and haptic information, in a way that was predicted by Bayes' theorem.

Once we are satisfied that humans are able to behave in a Bayes-optimal fashion in general it becomes interesting to see in which situations their optimality breaks down. O'Reilly et al. [17] discusses some instances in which a deviation from Bayesian predictions informs us about the limits of our cognitive system.

The usefulness of Bayes' theorem for modeling behavior and its particular characteristics are perhaps best illustrated during learning. Therefore we will spend the remainder of this chapter looking at the behavior of Bayesian systems during the learning of environmental contingencies.

9.4 Learning

As experimenters in cognitive neuroscience, we create the experimental environment in which our participants produce behavior. Therefore, we know the true parameters of the environment (in the previous example, this would be equivalent to knowing where Newton's apple actually is). However, the participant does not know these true values; s/he must infer them from observations.

Since we are interested in the behavior and brain activity of the participant, it is advantageous to have an estimate of what the participant knows or believes about the state of the environment, as this might differ from the true state. This is particularly true when data about the environment are presented sequentially, as in many psychological tasks. For example, in the gambling tasks such as the one-armed, two-armed and multi-armed bandit tasks, participants, from humans [2, 21] to the humble bumble bee [20], learn the probability of rewards associated with certain actions by trial and error; similarly in uncued attentional tasks such as the uncued Posner task [19], participants learn over many trials that targets are more likely to appear in certain locations than others. In these sequential tasks, a number of trials

must be experienced before the participant's estimates of the probabilities associated with each action approach the true values; in environments that change, continuous learning may be required.

9.4.1 'Today's Posterior Is Tomorrow's Prior'

As we briefly suggested above, learning from a sequence of observations can be modeled using iterative application of Bayes' rule. For example, let's say we observe a number of apples falling to the ground at locations x_1, $x_2 \ldots x_i$, and we want to infer from this the most likely location of fallen apples. Let's make the assumption that the distribution of apples is Gaussian around the tree trunk, with unknown mean μ (the location of the tree trunk) and variance σ^2. Then we can say that the variable x, the location of any given apple, follows a Gaussian distribution $x \sim \mathcal{N}(\mu, \sigma^2)$. Our aim is to infer the values of μ and σ^2 from the observed values of x.

Let's assume we have no a-priori knowledge about where the apple might fall. In other words, we start with a prior distribution such that all possible values of the parameters μ, σ^2 are considered equally likely. This is called a *uniform or flat prior*.

Then, on trial 1, we observe a data point, say $x_1 = 67$. Remember that Bayes' rule (Eq. 9.1) tells us we can update our prior (which is flat) by our likelihood to give our posterior. In our current situation, we can determine the likelihood, since we know for each possible pair of parameters μ, σ^2 the probability that a value of 67 would have been observed. In this case this is the probability density of a Gaussian $\mathcal{N}(\mu, \sigma^2)$ for a region about the value 67 with unit width. Thus, we can work out the probability of each possible pair of parameters μ, σ^2 given this one observation, and plot a probability density function over 'parameter space'—the range of possible values of μ and σ^2 (Fig. 9.3a). This probability density distribution based on the current observation is sometimes called the 'likelihood function'.

To obtain the posterior probability for each pair of values μ, σ^2 (the left hand side of Bayes' rule), we also need to take into account the prior probability that the values μ, σ^2 are correct by multiplying the likelihood function with the prior probability distribution. On trial one, we had a uniform prior, so the posterior is equal to the likelihood distribution. On trial two, we use the posterior resulting from trial one as be basis for our new prior. Again we observe a data point and update our prior to a new posterior that functions as the prior on the new trial. Etcetera. In general, what happens during learning is that the prior at trial i is derived from the posterior at trial i-1. Hence we can write Bayes' rule on trial i as follows:

$$p(x \sim N(\mu, \sigma^2) \,|x_{1:i}\,) \propto p(x_i \,|x \sim N(\mu, \sigma^2)) \times p(\, x \sim N(\mu, \sigma^2) \,|x_{1:i-1}). \tag{9.2}$$

Thus, the posterior distribution on trial i is proportional to the likelihood of the observed data, x_i, times the prior distribution at trial i, which was derived from the posterior at trial i-1 and captures all that is known about how previous data $x_{1:i-1}$ predict the current data point x_i.

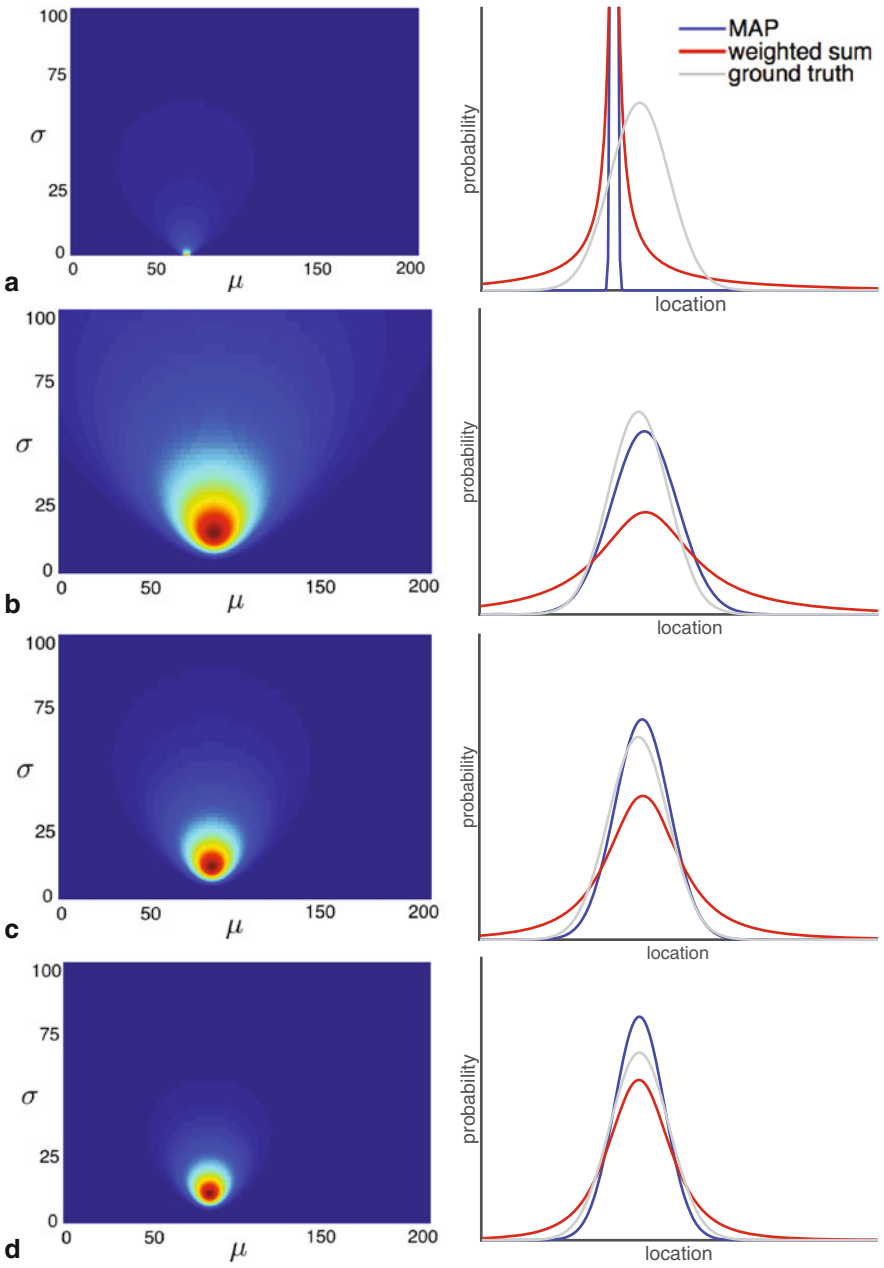


Fig. 9.3 Four trials of learning the mean and variance of a Gaussian distribution. The left hand panels show a "state space"–a space of candidate values for the mean (μ) and standard deviation (σ) of the Gaussian distribution from which the data are drawn. Colour indicates the posterior probability that each candidate pair of values of μ and σ matches the true values. The right hand panels show the corresponding probability distributions over location—the true distribution (*light grey line*), the maximum a posteriori distribution of apple locations (MAP, *blue lines*), and the weighted combination of all possible Gaussian distributions over locations (*red lines*)

9.4.2 In the Mind of Our Bayesian Participant

We can now look into the 'mind' or our Bayesian computer participant to see what it knows and believes about the environment on a trial-to-trial basis. After the first observation, the posterior distribution (our model's estimate of the true values of μ, σ^2) has a peak at $\mu = 67$ and a very low value for σ^2, since all observed apples (all one of them) fell near to $x = 67$.

We can represent the posterior over μ, σ^2 graphically on a grid (Fig. 9.3a, left panel) that represents all the possible combinations of mean and variance and their associated probabilities, which are denoted by colour. The space of all possible values for μ, σ^2 is called the *state space* or *parameter space*.

The next data point is $x_2 = 100$. This point is far from the previous observation. Therefore, the model's estimates shift. The estimated μ moves towards a point between 67 and 100, and the estimated σ^2 increases to create a distribution that encompasses *both* data points. As you can see the best fit Gaussian is now a much wider distribution, with a mean to somewhere in between 67 and 100 and a variance such that both data points are encompassed (Fig. 9.3b, left panel). However, the model is also relatively uncertain about the values of μ, σ^2 as can be seen from the wide spread of probability density across parameter space. As the model obtains more and more data the posterior distribution converges on a mean and standard deviation, and uncertainty decreases (Fig. 9.3c–9.3d, left panels).

We can translate our model's estimates of the values μ, σ^2 into a probability density function over location (i.e., plot probability as a function of the possible positions, x, at which apples could fall). We do this in the right-hand panels of Fig. 9.3.

To plot probability density over location, we need to decide how to summarize the distribution over parameter space, i.e. over μ, σ^2, which in fact represents our degree of belief in a range of different Gaussian distributions with different values of μ, σ^2 How should the distribution in parameter space be translated into a distribution over x? One option is to take the peak (or mode) of the distribution over μ, σ^2—the values at the deepest red spot in the left-hand panels of Fig. 9.3. This gives the most likely Gaussian distribution (the maximum a posteriori (MAP) estimate). The resulting distributions are shown in blue in the right hand panels of Fig. 9.3. However, this measure ignores uncertainty about that distribution, throwing away a lot of information. Another option is to take a weighted sum of all possible Gaussian distributions over space—as given by[2]:

$$p(x) = \sum_j \sum_k p(x \,|x \sim N(\mu_j, {\sigma^2}_k)) \times p(x \sim N(\mu_j, {\sigma^2}_k) \,|x_{1:i}). \qquad (9.3)$$

[2] In all the examples and exercises given here, we obtain an approximate solution by evaluating p(x) for discrete values of (μ, σ^2). In the continuous case, Eq. 9.3 would become: $p(x) = \int d\mu \int d\sigma^2 [p(x|x \sim N(\mu, \sigma^2)) \times p(x \sim N(\mu, \sigma^2)|x_{1i})]$

This gives a distribution over x that takes into account variance due to uncertainty over μ, σ^2. The resulting distributions are shown in red in the right hand panels of Fig. 9.3.

9.4.3 What Is it Useful For?

Using a Bayesian learner that iteratively integrates observed data with what is known from previous observations allows us to follow the dynamics of the different model parameters on a trial-by-trial basis. Using this trial-by-trial information, we can make predictions about behaviour—i.e. where Isaac should forage for apples on each trial. For example, we might hypothesize that he will search in an area centered on the estimated value of μ (i.e., he will search around where he thinks the tree is) and that the size of the area he searches in should be proportional to σ^2.

Furthermore, because the Bayesian model represents Isaac's beliefs about μ and σ^2 as probability distributions, our Bayesian model gives us an estimate of how uncertain he should be about those values (the spread of probability density in parameter space). Because we have this insight into uncertainty, which is the defining feature of Bayesian models, we can make and test predictions about behavior based on uncertainty about the environment (estimation uncertainty [10, 18])—for example, we might expect Isaac to express more exploratory behavior when uncertainty about μ and σ^2 is high [4, 6]. Moreover, we might expect that certain parameters of our Bayesian model might be reflected in neural activity. Although one has to be careful with interpretation [16], it is possible to link the values of the model's parameters to brain activity.

9.4.4 Another Example of a Bayesian Learner: One-armed Bandit

In the previous example, we showed how a Bayesian computer participant could be used to model what a human participant knows or believes about the parameters of a Gaussian distribution from which spatial samples (the location of apples) were drawn. In fact, this is one example in which the beliefs of the participant (at least, an optimal Bayesian participant) rapidly approach the true state of the environment as can be seen in Fig. 9.3.

There are many tasks, including some in common use in cognitive neuroscience, where an internal model based on sampling of the environment is a much weaker approximation of the true state of the environment. One such example is given by tasks in which the environment changes frequently (so the observer must constantly update his model of the environment) [2]. Another case is presented by tasks in which the parameters of the environment are learned slowly. These include probabilistic tasks—for example if we observe a binary variable (say, reward vs. no reward), we need several trials to estimate the underlying probability p(reward): to tell the

difference between a reward probability of 0.8 and 0.9 would require at least ten trials for example.

The more the participant's internal model of the environment differs from the true state of the environment, the more useful it is for the experimenter to have a model of what the participant knows/believes about the state of the world rather than assuming the true parameters of the environment are known.

We will now consider a Bayesian computer participant in a one-armed bandit task. This is a task in which learning naturally requires a larger number of trials and hence participants' model of the environment is likely to differ from the true state of the environment. We will see that in this task the Bayesian computer participant can give us rich insights into what participants might think/believe on each trial of the task.

In the one-armed bandit task, participants must choose between two actions, A and B (say, press a button with the left or right hand), only one of which would lead to delivery of a reward. The probability that action A is rewarded is set at some value q; the probability that action B would be rewarded is then $(1–q)$; formally we can say that the probability that action A is rewarded follows a Bernoulli distribution (a single-trial binomial distribution) with probability parameter q. From time to time during the task, the value of q changes to a new value; participants do not know when these changes will occur or how frequently. Hence the participant's task is to infer both the current value of q, and the probability v of a change in q, from the observed data. The details of this model are not central to our point here, which is to illustrate that a Bayesian model can give rich insights into the internal thought processes of the participant. However, for the interested reader we describe the model used to generate the figures in Appendix A.

Figure 9.4 (left hand panel) illustrates the task data and model fit. Values of q were generated randomly with a jump probability (true value of v) of $^1/_{15}$—the true value of q on each trial is indicated by the white line. The side that was actually rewarded on each trial is indicated by the presence of a dot on the left or right of the plot, respectively. Remember that as p(Left rewarded) = 1–p(Right rewarded), the player knows which side was rewarded on all trials, even when he chose the unrewarded side.

The red line and shaded areas represent the model's maximum likelihood estimate of the state of the environment (the value of q). The shading represents the probability density distribution over q on each trail, according to the model.

Inspecting the model's estimates of the environment across trials, we can see a number of interesting features. Firstly, we notice that the maximum likelihood estimate is close to the true value of q most of the time. However, when there are changes in the underlying environment, the model takes a few trials to 'catch up'. Secondly, we can see that the model's uncertainty about the state of the world generally decreases over time (the shaded area gets narrower over time), but uncertainty increases when there is a change in the environment, or when a change is suspected.

In the right hand panels of Fig. 9.4 (labeled a, b, and c), we take a closer look at the probability density distributions across parameter space for three sets of trials around change points or suspected change points. The data points in question are labeled a, b, and c in the left hand panel. For each data point we show the distribution of

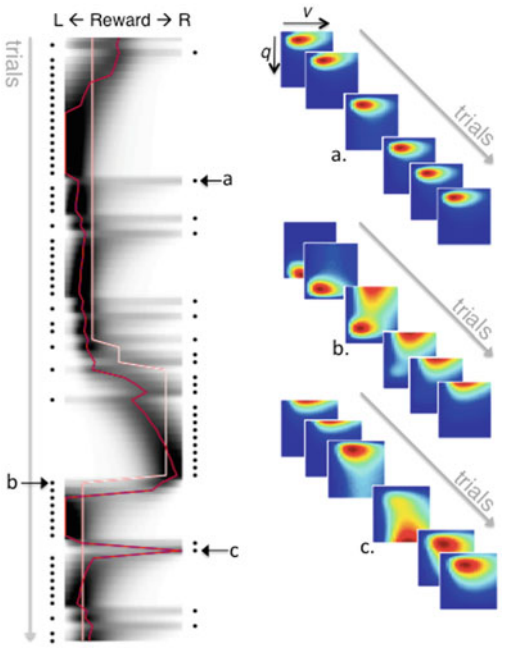


Fig. 9.4 Learning in an unstable environment. The left hand panel shows the rewarded sides on each trial (*black dots*); the true probability of reward, i.e., the true value of q (*white line*); the model's estimate of q (*red line*); and the uncertainty of the model's estimation (*shade*). The left hand panels show the model's beliefs of each possibility in state space. **a** No update—A point at which a single right side reward trial is observed. **b** Successful update. **c** False/temporary update

probability density across parameter space on that trial and surrounding trials (right hand panel). These plots are analogous to the parameter space plots in the left hand panel of Fig. 9.3, but instead of plotting the distribution of probability density across values of μ and σ^2 we are now plotting probability density across values of q and v.

Just before time point a, the model 'thinks' that q, the probability of the left side being rewarded, is near to 100 %, as it has just experienced a long run of left-rewarded trials. At point a, a right-rewarded trial is observed (the actual trial labeled a in the left hand panel is the same one labeled a in the right hand panel). The probability that associated with values of q and v other than those which were favored before point a increases. However, subsequent trials continue to be left-rewarded, and the model reverts to its previous state of believing the probability of left-rewarded trials to be very high.

In contrast, time point b represents a successful update. Prior to b, there was a long run of right rewarded trials, followed by an actual change in q (white line) and a series of left-rewarded trials starts. In this case, the model updates its estimate of q over a series of trials. On trial b itself, the model is clearly entertaining both the hypothesis that q has changed, and the hypothesis that q remains the same. Note that the 'change' hypothesis is associated with a higher value of v (the peak is further to the right), compared to the 'no change' hypothesis, as we would expect since v is the probability of change, which is inferred based on the number of change points that were observed.

Finally, point c represents a point at which the model is erroneously updated when there was in fact no change in q. Just before point c, the model 'thinks' q is almost 100 %, i.e. only left-rewarded trials can occur. It then observes two right-rewarded trials, leading it to think q has changed to favour right-rewarded trials. However, these two trials are followed by more left-rewarded trials, leading the model to revert to its former hypothesis (favouring the left) but with a more moderate probability value, so q is now nearer to 80 % than 100 % (indeed, the maximum likelihood estimate of q is now nearer to the true value of q, as seen from the white and red lines on the left-hand panel).

We have briefly described some snapshots of 'interesting behavior' of the Bayesian learning algorithm, in order to illustrate how constructing such a model could allow us to 'peek inside' the mind of a model participant to see how its beliefs about the state of the evolve. We have seen, for example, that learning models can capture lags when even an optimal participant could not yet have adjusted to a change in the environment. We have seen that when a model is fit to the actual data observed by a participant, it can indicate when the participant could mis-estimate the parameters of the environment (such as at point c). We have also seen that Bayesian models can give us insights into internal features of learning such as uncertainty, which may themselves predict neural and/or behavioral data. Hopefully this brief illustration will convince the reader that explicitly modeling the contents of the participant's mind, as with a Bayesian learning model, can generate and refine our predictions about what activity we might find in their brain, beyond what could be achieved by simply relating brain activity to stimuli or responses.

9.5 Conclusion

In this chapter, we have discussed the use of Bayesian models in cognitive neuroscience. We have illustrated of the main characteristics of Bayesian models, including the representation of beliefs as probability distributions, the use of priors, and sequential updating of information. These models can be highly predictive of the actual behavior displayed by humans and animals during a variety of tasks. We have looked closely at two learning tasks, one in a stable and one in an unstable environment, and charted how the beliefs of a Baysian model change over trials. The parameters of such a model can then be used to interrogate behavior and brain function.

Appendix A: One-Armed Bandit Model

We can write down the *generative* model, by which the rewarded action (A or B) is selected as follows:

p(A rewarded on trial i)$\sim$ **Bernoulli**(q_i)

$$q_i = \begin{cases} q_{i-1} & if\ J = 0 \\ rand(0,1) & if\ J = 1 \end{cases}$$

... where J is a binary variable which determines whether there was a jump in the value of q between trial i-1 and trial i; J itself is determined by

$$J\text{-}Bernoulli(v)$$

... where v is the probability of a jump, e.g. if a jump occurs on average every 15 trials, v = $^1/_{15}$.

Then we can construct a Bayesian computer participant which infers the values of q and v on trial i as follows:

$$p(q, v\,|x_{1:i}) = p(x_i\,|q_i\ , v)p(q_i, v)$$

where the prior at trial i, $p(q_i, v)$, is given by

$$p(q_i, v) = p(q_i|q_{i-1}, v)\ p(q_{i-1}, v|x_{1:i-1})$$

and the transition function $p(q_i\,|q_{i-1}, v)$ is given by

$$p(q_i\,|q_{i-1}, v) = (1 - v)\ q_{i-1} + v\left(\frac{1}{Uniform\ (0,1)}\right)$$

Exercises

Exercise 1. Look at Fig. 9.5. How do you interpret the shadow on the surface shapes? Most people see the left hand side bumps as convex and the right hand bumps as concaves. Can you explain why that might be, using your Bayesian perspective? Hint: think of the use of priors.

Exercise 2. In Fig. 9.4 we saw some interesting behavior by a Bayesian learner. For instance, at point c the model very quickly changed its belief of an environment where left was rewarded into one where right was rewarded. One important goal of model-based cognitive neuroscience is to link this type of changes probability distributions to observed neural phenomena. Can you come up with some phenomena that can be linked with changes in the model's parameters?

Exercise 3. In this final exercise we will ask you to construct a simple Bayesian model. The solutions include example Matlab code, although they are platform independent. Consider the following set of observations of apple positions x, which Isaac made in his garden:

1. *Find* the mean, $E(x)$, and variance, $E(x^2)$–$E(x)^2$, of this set of observations using the formulae

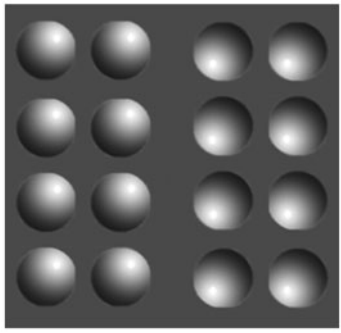


Fig. 9.5 Convex or concave?

i	x_i
1	63
2	121
3	148
4	114
5	131
6	121
7	90
8	108
9	76
10	126

$$E(x) = \frac{1}{n}\sum_i x_i$$

$$E(x^2) = \frac{1}{n}\sum_i {x_i}^2$$

2. If I tell you that these samples were drawn from a normal distribution, $x \sim N(\mu, \sigma^2)$ how could you use Bayes' theorum to find the mean and variance of x? Or more precisely, how could you use Bayes' theorem to estimate the parameters, μ and σ^2, of the normal distribution from which the samples are drawn?

Hint: remember from the text that we can write

$$p(x \sim N(\mu, \sigma^2) \,|x_1 \ldots x_n\,) \propto p(x_1 \ldots x_n \,|x \sim N(\mu, \sigma^2)) p(x \sim N(\mu, \sigma^2))$$

... where the likelihood function, $p(x_i \,|x \sim N(\mu, \sigma))$, is given by the standard probability density function for a normal distribution:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

. . . and you can assume:

1. The prior probability $p(x \sim N(\mu, \sigma^2))$ is equal for all possible values of μ and σ^2, and
2. The observations are independent samples such that $p(x_i \cap x_j) = p(x_i)p(x_j)$ for all pairs of samples $\{x_i, x_j\}$.

Now use MATLAB to work out the posterior probability for a range of pairs of parameter values μ and σ^2, and find the pair with the highest joint posterior probability. This gives a maximum likelihood estimate for μ and σ^2.

3. Can you adapt this model to process each data point sequentially, so that the posterior after observation *i* becomes the prior for observation $i + 1$?

Hint: remember from the text that (assuming the underlying values of μ and σ^2 cannot change between observations), we can write:

$$p(x \text{-} N(\mu, \sigma^2) \,|x_1 \ldots x_i) \propto p(x_i \,|x \text{-} N(\mu, \sigma^2)\,) p(x \text{-} N(\mu, \sigma^2) \,|x_1 \ldots x_{i-1})$$

. . . where the prior at trial *i*, $p(x \text{-} N(\mu, \sigma^2) \,|x_1 \ldots x_{i-1})$ is the posterior from trial *i-1*.

4. If you have done parts 2 and 3 correctly, the final estimates of $\{\mu, \sigma^2\}$ should be the same whether you process the data points sequentially, or all at once. Why is this?

Further Reading

1. McGrayne [14] provides an historical overview of the development of Bayes' theorem, its applications, and its gradual acceptance in the scientific community;
2. Daniel Wolpert's TED talk (available at http://www.ted.com/talks/daniel_wolpert_the_real_reason_for_brains.html) provides a nice introduction in to consequences of noise in neural systems and the Bayesian way of dealing with it;
3. O'Reilly [15] discusses Bayesian approaches to dealing with changes in the environment and how different types of uncertainty are incorporated into Bayesian models and dealt with in the brain.
4. Nate Silver's book *The signal and the noise* [23] contains some nice example about how humans make predictions and establish beliefs. Silver advocates a Bayesian approach to dealing with uncertainty. It served him very well in the 2012 USA presidential elections, when he correctly predicted for each of the 50 states whether they would be carried by Obama or Romney.
5. David MacKay's book *Information theory, inference, and learning algorithms* [12] is a much more advanced treatment of many of the principle of Bayesian thinking. It is available for free at http://www.inference.phy.cam.ac.uk/itprnn/book.html.

References

1. Bayes T (1763) An essay towards solving a problem in the doctrine of chances. Phil Trans 53:370–418
2. Behrens TE, Woolrich MW, Walton ME, Rushworth MF (2007) Learning the value of information in an uncertain world. Nat Neurosci 10:1214–1221
3. Chater N, Oaksford M (eds) (2008) The probabilistic mind: Prospects for Bayesian cognitive science. Oxford University Press, Oxford
4. Courville AC, Daw ND, Touretzky DS (2006) Bayesian theories of conditioning in a changing world. Trends Cogn Sci 10:294–300
5. Cox RT (1946) Probability, frequency and reasonable expectation. Am J Phys 14:1–13
6. Dayan P, Kakade S, Montague PR (2000) Learning and selective attention. Nat Neurosci 3:1218–1223
7. Ernst MO, Banks MS (2002) Humans integrate visual and haptic information in a statistically optimal fashion. Nature 415:429–433
8. Gregory R (1966) Eye and brain. Princeton University Press, Princeton
9. Jacobs RA (1999) Optimal integration of texture and motion cues to depth. Vis Res 39:3621–3629
10. Knight FH (1921) Risk, uncertainty and profit. Hart, Schaffner and Marx, Boston
11. Körding KP, Wolpert DM (2006) Bayesian decision theory in sensorimotor control. Trends Cogn Sci 10:319–326
12. MacKay DJC (2003) Information theory, inference, and learning algorithms. Cambridge University Press, Cambridge
13. Mars RB et al (2008) Trial-by-trial fluctuations in the event-related electroencephalogram reflect dynamic changes in the degree of surprise. J Neurosci 28:12539-12545
14. McGrayne SB (2011) The theory that would not die: How Bayes' rule cracked the enigma code, hunted down Russian submarines, and emerged triumphant from two centuries of controversy. Yale University Press, New Haven
15. O'Reilly JX (2013) Making predictions in a changing world-inference, uncertainty, and learning. Front Neurosci 7:105
16. O'Reilly JX, Mars RB (2011) Computational neuroimaging: Localising Greek letters? Trends Cogn Sci 15:450
17. O'Reilly JX, Jbabdi S, Behrens TE (2012) How can a Bayesian approach inform neuroscience? Eur J Neurosci 35:1169–1179
18. Payzan-LeNestour E, Bossaerts P (2011) Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. PLoS Comp Biol 7:e1001048
19. Posner MI, Snyder CRR, Davidson BJ (1980) Attention and the detection of signals. J Exp Psychol Gen 109:160–174
20. Real LA (1991) Animal choice behavior and the evolution of cognitive architecture. Science 253:980–986
21. Robbins H (1952) Some aspects of the sequential design of experiments. Bull Amer Math Soc 58:527–535
22. Segall MH, Campbell DT, Herskovits MJ (1963) Cultural differences in the perception of geometric illusions. Science 139:769–771
23. Silver N (2012) The signal and the noise: Why most predictions fail but some don't. Penguin, New York

Chapter 10
Constraining Cognitive Abstractions Through Bayesian Modeling

Brandon M. Turner

Abstract There are many ways to combine neural and behavioral measures to study cognition. Some ways are theoretical, and other ways are statistical. The predominant statistical approach treats both sources of data as independent and the relationship between the two measures is inferred by way of a (post hoc) regression analysis. In this chapter, we review an alternative approach that allows for flexible modeling of both measures simultaneously. We then explore and elaborate on several of the most important benefits of this modeling approach, and close with a model comparison of the Linear Ballistic Accumulator model and a drift diffusion model on neural and behavioral data.

10.1 Introduction

As this book will demonstrate, there are many ways in which the neurosciences can inspire mathematical models of cognition, and vice versa. Perhaps the most theoretically-oriented way is to develop models on the basis of what is observed at the neurophysiological level. This approach could be called "bottom-up" because at its very core, the models are designed to embody neurological principles [1–9].

If a model is not originally designed as a bottom-up model, it is possible to adjust the assumptions of the model so that it resembles a bottom-up structure. An example of this is the latest instantiation of the ACT-R model (see [10] and Chap. 18 of this text). Anderson and colleagues [11] developed a model of the process of equation solving that also predicted patterns of brain activity, as measured by the blood oxygen level dependent (BOLD) response. Their model assumed that observers maintain a set of modules that become active when required by subtask demands, and become inactive when no longer in use. In the model, BOLD responses are produced in the brain areas corresponding to active modules by convolving a module activation function (i.e., a function carrying binary active or inactive information) with a hemodynamic response function. The correspondence between modules and

B. M. Turner (✉)
Psychology Department, The Ohio State University, Lazenby Hall,
Room 200C, Columbus, OH 43227, USA
e-mail: turner.826@gmail.com

B. U. Forstmann, E.-J. Wagenmakers (eds.), *An Introduction to Model-Based Cognitive Neuroscience,* DOI 10.1007/978-1-4939-2236-9_10

brain areas was based on prior work, where each module was mapped to region(s) of the brain where the greatest brain activity occurred [12]. After estimating a few free parameters, the model was shown to provide a reasonable fit to both the neural and behavioral data.

While the bottom-up approach is useful for a number of reasons, it is often more desirable to take a "top-down" approach because for many cognitive models, it is not straightforward to map model mechanisms (e.g., module activation) to particular brain regions [12]. Top-down approaches generally proceed by (1) fitting a cognitive model to behavioral data, (2) fitting or examining patterns in the neural data, and (3) regressing the parameter estimates of the cognitive model to the neural signature of interest [13–21]. This approach has been highly successful in relating response caution adjustments across different speed emphasis instructions (e.g., subjects instructed to respond quickly or accurately) to the pre-supplementary motor area (pre-SMA) and the (anterior) striatum [14–16].

Both the top-down and bottom-up approaches are reciprocal in the sense that we derive an understanding about a cognitive process on the basis of formal cognitive models and cognitive neuroscience [22]. However, neither approach is *statistically* reciprocal. In the bottom-up approach, by "statistically reciprocal", we mean that the model is not usually fit to both neural and behavioral data simultaneously. By contrast, in the top-down approach, we mean that the information contained in the neural data do not inform the estimation of the behavioral model parameters. Instead, the relationship between the neural and behavioral data is inferred after two independent analyses.

In this chapter, we discuss an alternative top-down approach that enforces mutual constraint across both sources of data. The chapter elaborates on the work of Turner et al. [23], where the authors conjoin "submodels" of singular (i.e., neural or behavioral, but not both) measures by way of a hierarchical Bayesian framework. Bayesian modeling has become popular in many neural [24–29] and behavioral [30–39] modeling applications for a number of theoretical and practical reasons (see Chap. 9 of this text). We itemize and discuss several advantages of our approach below. We present these advantages first at a conceptual level, then if useful to articulate a point, we present relevant mathematical details. The less technically-oriented reader may safely skip these areas. We encourage the more technically-oriented reader to consult Turner et al. [23] for additional details.

10.2 Joint Modeling

The "joint modeling framework" is a statistical framework for relating a subject's neural data to their behavioral data and vice versa. As we discussed in the introduction, it is often difficult to specify a single model for describing both aspects of the data simultaneously without considerable additional theoretical overhead and increased computational complexity. A simpler approach is to instead focus on each facet of the data individually.

Let's begin with the behavioral data, which we denote B. We must first specify a model for how these data may have been produced. Over many years, cognitive modelers have developed an entire suite of behavioral models for a variety of tasks. Behavioral models assume a system of mathematical or statistical mechanisms governed by a set of parameters that are latent. Behavioral models are interesting because they derive explicit assumptions about how the mind works based on an overarching cognitive theory. However, the behavioral models are not meant to be taken literally; instead, the models are tools by which inferences can be made about how an underlying cognitive process is affected under different contexts or experimental manipulations. In this sense, behavioral models are simply instantiations of an abstract cognitive process, whose parameters serve as proxies.

The behavioral model we choose can be anything, such as a signal detection theory model for data from a perceptual discrimination experiment [31–40], the bind cue decide model of episodic memory [41], or the drift diffusion model for choice response times (see [42] and Chap. 15 of this text). As we will discuss later in this chapter, we can even choose a different model for the same data, so that we can compare the models on the basis of generalizability, and degree of model fit. We will write the probability distribution for the data under our assumed behavioral model as $p(B \mid \theta)$, where θ denotes the set of parameters in the model.

We can now turn to the neural data, which we denote N, and the neural model, whose probability density function we write as $p(N \mid \delta)$, where δ is the set of neural model parameters. As with the behavioral model, we are not constrained to any particular neural model, so for example, we could choose the generalized linear model [25, 26, 43], the topographic latent source analysis model [24], a wavelet process [44], or a simple hemodynamic response function [45]. However, neural models tend to differ from behavioral models because they generally do not make explicit connections to a cognitive theory. As such, neural models are statistical in nature, and so any joint model created within our framework embodies the principles assumed by the behavioral model alone.[1]

There are two main factors in selecting the neural model. First, the model should reduce the dimensionality of the neural data. Generally speaking, there will be a large amount of neural data (e.g., from an EEG experiment), and so reducing the data's dimensionality will reduce the computational complexity associated with fitting the joint model to data. Second, the model should provide information about the neural signal in a generative form. For example, the neural model could describe the location, shape, and degree of activation of the neural source(s) of interest by way of a mixture of normal distributions [23]. As another example, the neural model could describe how factors (i.e., covariates) in an experimental design are associated with a neural signal through a general linear model [25, 43].

[1] Note that the neural model could also make explicit theoretical assumptions. However, we feel that a theoretical model on the behavioral side with a statistical model on the neural side will be of greatest interest to the readers of this book.

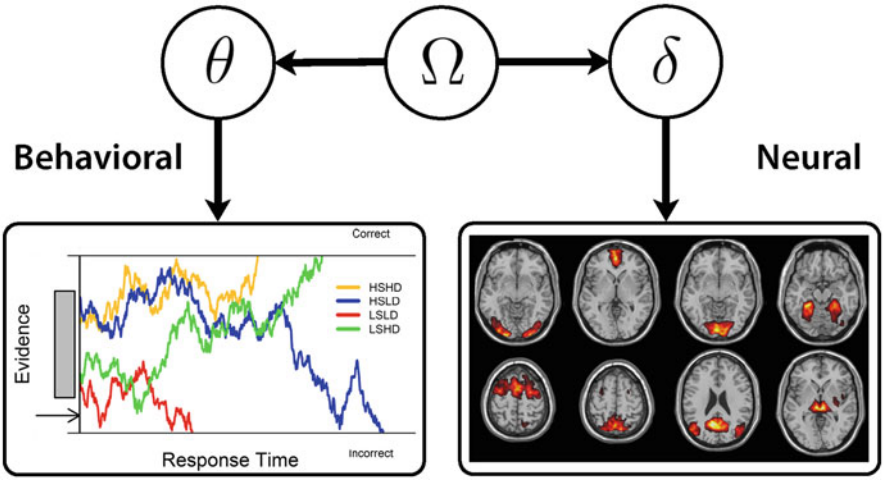


Fig. 10.1 Graphical diagram for the joint modeling approach. The *left* side shows a chosen behavioral submodel (i.e., the drift diffusion model) and the *right* side shows a chosen neural submodel (i.e., a method for single trial analysis of fMRI data [47]). The two submodels describe their respective data with a set of parameters θ (for the behavioral submodel) and δ (for the neural submodel), which are then connected through a hyper distribution with parameters Ω

Having selected our behavioral and neural models, which we will refer to as "submodels" henceforth, we now must make the connection between the submodels' parameters explicit. At this point, one could specify a discriminative (i.e., a causal) relationship between the parameters; however, here we are only interested in generative approaches where both sets of parameters are assumed to be manifestations of the same underlying construct. To make a fully generative model explicit, one need only specify a structural form for the joint distribution of the parameters of the submodels. As discussed in Turner et al. [23], a convenient choice is to assume the parameters are distributed according to a multivariate normal distribution. Figure 10.1 shows a graphical diagram for a joint model used in a recent study [46]. The right side of the diagram illustratively shows the neural data as a set of functional magnetic resonance imaging (fMRI) scans with the submodel parameters δ, whereas the left side of the diagram shows the behavioral model as a drift diffusion process with submodel parameters θ. In the middle of the diagram, the hyperparameters Ω connect the two submodels through an assumption about the joint distribution of θ and δ (see Eq. 10.1).

As illustrated in Fig. 10.1, the joint modeling approach is a convenient way to generate expanded models that capture multiple facets of the same underlying construct. However, to physically implement the approach, the technical details of the model must be carefully considered. In the following subsection, we discuss the technical aspects of our approach in greater detail. The uninterested reader may skip the next section and incur little conceptual loss.

10.2.1 Technical Details

We let $\mathbf{M}(\Omega)$ with parameters Ω denote the linking distribution of θ and δ, such that

$$(\delta, \theta) \sim \mathbf{M}(\Omega), \tag{10.1}$$

and denote the corresponding density function as $p(\delta, \theta \mid \Omega)$. The type of distribution we assume for $\mathbf{M}$ will depend on the properties of the behavioral and neural model parameters. Currently, the only linking distribution we have investigated is the multivariate normal distribution, because it provides a convenient distribution with infinite support and clear parameter interpretations. Assuming multivariate normality places some restriction on the types of submodel parameters we can use. Specifically, the assumption of multivariate normality should be reasonable, meaning that a multivariate normal distribution should adequately describe the variation between submodel parameters. Furthermore, the parameters should have infinite support (i.e., they should not be bounded). When a submodel parameter is bounded, transformations can be used, such as the log or logit, to produce infinite support for the parameter space. In this chapter, we will use the multivariate normal assumption.

Under the multivariate normality assumption, $\Omega = \{\phi, \Sigma\}$ consists of a set of hyper mean parameters ϕ and hyper dispersion parameters Σ, and Eq. 10.1 becomes

$$(\delta, \theta) \sim \mathcal{N}_p(\phi, \Sigma), \tag{10.2}$$

where $\mathcal{N}_p(a, b)$ denotes the multivariate normal distribution of dimension p with mean vector a and variance-covariance matrix b.

As we will discuss below, the properties of the hyperparameters will depend on how the lower-level parameters θ and δ are used. For example, θ and δ could represent subject-specific parameters meaning that Ω would describe the distribution of the model parameters between subjects in the group. By contrast, θ and δ could also represent trial-specific parameters meaning that Ω would be a set of condition- or subject-specific parameters. Regardless of the characterization of the model parameters, the hyper mean vector ϕ can be divided into the set of mean parameters for the neural submodel (δ_μ) and the behavioral submodel (θ_μ), such that $\phi = \{\delta_\mu, \theta_\mu\}$. Similarly, the variance-covariance matrix Σ can be partitioned as

$$\Sigma = \left[\begin{array}{c|c} \delta_\sigma^2 & \rho\delta_\sigma\theta_\sigma \\ \hline (\rho\delta_\sigma\theta_\sigma)^\intercal & \theta_\sigma^2 \end{array}\right]$$

to reflect that it consists of matrices that characterize various dispersions of the model parameters. Note that the variance-covariance matrix $\rho\delta_\sigma\theta_\sigma$ uses the parameter matrix ρ to model the correlation between submodel parameters. Specifying the model in this way allows us to directly infer the degree to which behavioral submodel parameters are related to neural submodel parameters. To reduce the number of model parameters, we can also constrain elements of this variance-covariance matrix to be

equal to zero. Such constraints are particularly useful when the intention of one's research is confirmatory rather than exploratory.

As for the variance-covariance matrices δ_σ^2 and θ_σ^2, there are two modeling approaches that we advocate. First, we can specify that these matrices be diagonal matrices so that only the variances of the corresponding submodel parameters are captured [23]. Doing so assumes that the submodel parameters are independent, which may not necessarily be true. However, making this assumption limits the number of parameters in the model while still allowing for relationships between the submodel parameters to be observed through the joint posterior distribution. Second, one can specify that these matrices be square (i.e., elements are symmetric about the diagonal), and build in relationships between submodel parameters explicitly [46]. Surprisingly, adding these extra off-diagonal elements in the matrix can facilitate the estimation process because the conditional distributions of ϕ and Σ become analytic, and so Gibbs sampling can be used to efficiently generate proposals for the hyperparameters [48, 49].

Once the model has been fully defined, the final step is to specify prior distributions for ϕ and Σ. There are a number of priors to choose from, and the choice will depend on how Σ is specified. When Σ is unconstrained so that it is symmetric and all of its elements are free to vary, we recommend a conjugate (dependent) prior on $\Omega = (\phi, \Sigma)$, such that

$$p(\Omega) = p\,(\phi, \Sigma) = p(\phi \mid \Sigma)p(\Sigma).$$

Here, we use a multivariate normal prior for $p(\phi \mid \Sigma)$ and an inverse Wishart prior on $p(\Sigma)$. An application of these priors appears below. Having fully specified the model, the joint posterior distribution of the model parameters is

$$p(\theta, \delta, \Omega \mid B, N) \propto p(B \mid \theta)p(N \mid \delta)p(\theta, \delta \mid \Omega)p(\Omega). \tag{10.3}$$

To estimate the model parameters, we require an algorithm to sample from the joint posterior distribution in Eq. 10.3. The type of algorithm we use will depend on how difficult it is to evaluate Eq. 10.3. For relatively simple applications, we can use software programs such as WinBUGS [50] or JAGS [51] to carry out the estimation procedure. For more complex problems, we may require algorithms that scale to high dimensions and/or tune to the shape of the joint posterior distribution (e.g., [52–56]). Finally, when the likelihood functions in Eq. 10.3 are difficult to evaluate, we may require algorithms that approximate [57–60] or efficiently evaluate [61] them.

Conjoining submodels in this way allows for a statistically reciprocal relationship between the behavioral and neural submodel parameters. Specifically, when there is a nonzero correlation between any of the submodel parameters, modeling both aspects of the data simultaneously provides greater parameter constraint than modeling each subset of the data independently. There are many other benefits of our joint modeling approach, many of which are discussed in Turner et al. [23]. In what follows, we will expand on and clarify the discussion of five of these benefits.

10.3 Prediction of Neural or Behavioral Measures

One of the major benefits of using a Bayesian approach lies in the flexible adjustment for, and prediction of, missing data. In the joint modeling framework, we can, for example, generate predictions for a particular subject's behavioral data given only that subject's neural data. The framework allows for predictions in the opposite direction as well, so that given only behavioral data, we can make predictions about a subject's neural data. The way this occurs in the model is through the hyper parameters Ω, and their relationship to the parameters θ and δ. Importantly, the hyper parameters Ω enforce that while the parameters θ and δ are conditionally independent, they are not marginally independent. Hence, the distribution of θ depends on the particular value of δ, and vice versa. From this type of dependency, the model can generalize the relationship between θ and δ to flexibly generate predictions about missing data.

Suppose we fit a joint model to a set of data consisting of a number of subjects. Assume that for a (small) subset of the subjects, we only have information about one data source (e.g., only neural data). The model would learn the relationship between the submodel parameters from the subjects who were fully observed (i.e., the subjects who had both data sources), and this information would be stored in the hyper parameters Ω. When given, say, only neural data for a particular subject, we could only form an estimate of the neural submodel parameters δ. However, given the estimated δ, the model can generalize the information stored in Ω to produce a prediction about the behavioral submodel parameters θ, which can then be used to generate predictions about the behavioral data B. This predictive process can be easily reversed to generate predictions about the neural data having only observed behavioral data.

Turner et al. [23] performed a simulation study to demonstrate a joint model's ability to flexibly predict either behavioral or neural data. In their study, Turner et al. formed a joint model by combining a finite mixture model as the neural submodel, and a signal detection theory model as the behavioral submodel. They simulated data for 24 subjects; however, when fitting the model, they withheld the neural data for two of the subjects, and they withheld the behavioral data for two other subjects. The remaining 20 subjects were fully observed, and no data were withheld. The objective of their study was to demonstrate how well the joint model could predict the withheld data for the four subjects, given only partial information about those subjects. They showed that the data predicted by the model were entirely consistent with the data that were withheld.

Figure 10.2 illustrates how the model predicts withheld data given partial information. The left side of the figure shows the neural model at the parameter level (upper), and the voxel level (i.e., the neural data space; lower). The right side of the figure similarly shows the behavioral submodel parameters (upper) and the behavioral data space (lower). The middle of the figure shows the relevant subset of the hyper parameters. The figure illustrates two predictions of the model: the orange path designates a prediction of behavioral data, whereas the blue path designates a prediction of neural data. On the orange path, the model is provided with neural data in the form of a pattern of brain activity in voxel space (green square). The neural data informs the estimates of the neural model parameters (green histogram),

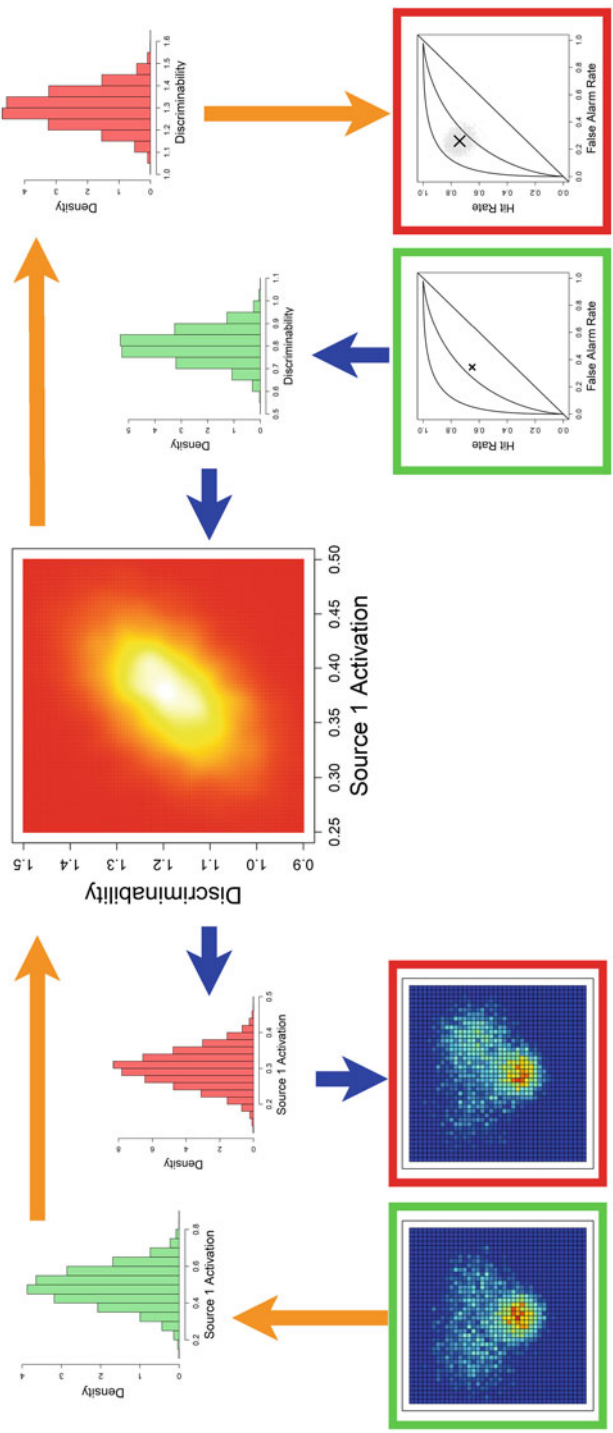


Fig. 10.2 Path diagram for generating predictions from a joint model. The paths begin and end at the lowest levels of the hierarchy. The *left* side of the figure shows the neural model at the parameter level (*upper*) and the voxel level (*lower*). Similarly, the *right* side shows the behavioral model at the parameter (*upper*) and behavioral data space (*lower*) levels. The *middle* of the figure represents the joint parameters Ω. The *orange* path designates a prediction made for behavioral data, whereas the *blue* path designates a prediction made for neural data. Data inputs are surrounded by *green* boxes, and model predictions are surrounded by *red* boxes

and the estimate is passed upward to the hyper parameters. At the hyper parameter level, the model learns the relationship between the neural and behavioral submodel parameters from the 20 subjects who were fully observed. The model then generalizes an estimate for the behavioral submodel parameters (red histogram) based on (1) the relationship between the submodel parameters, and (2) the particular neural submodel parameter estimate. Finally, the behavioral submodel parameter estimate is used to make a prediction about the hit and false alarm rate for this subject (red square). The gray cloud represents the distribution of predictions (i.e., the posterior predictive distribution), and the black "X" marks the withheld behavioral data. The blue path works in the same way, but in the opposite direction: the model is provided behavioral data (green square) from which an estimate of the behavioral submodel parameter is inferred (green histogram). The model then generalizes the parameter estimate to form an estimate of the neural model parameters (red histogram), which ultimately produces a prediction for the neural data (red square).

Although the joint modeling framework is able to flexibly predict missing data, our approach is a generative one and is likely to produce inferior predictions to discriminative approaches [62]. The two approaches are different in the way they treat the "inputs" or observed data. Discriminative models condition on one variable (e.g., the neural data) to make a prediction about another variable, and as a consequence, they necessarily have fewer random variables and greater certainty when generating predictions. One disadvantage of the discriminative approach is that one must first train the model to learn the relationship between neural and behavioral data by having a (large) set of both data types. Then, given only neural data, the discriminative model can make predictions about the behavioral data [63, 64]. However, to make a prediction for neural data (i.e., in the opposite direction), the discriminative model would need to be retrained on the full data set. Another disadvantage is that discriminative techniques are designed to make predictions about discrete variables (e.g., correct or incorrect), and are more difficult to use for predictions of continuous variables (e.g., response times). Finally, and most importantly from our perspective, discriminative models are statistical in nature, meaning that they make no connection to an explicit cognitive theory.

10.4 Additional Constraint on Cognitive Theory

Another important reason for using the joint modeling approach is the additional constraint provided by supplementary sources of information. From a cognitive modeling perspective, it is preferable to capture as many aspects of the underlying cognitive processes as possible. Because data from the neurosciences can be viewed as additional manifestations of the same cognitive process that the behavioral data provide, it follows that augmenting cognitive models with neuroimaging data is a way to advance our understanding of the mechanisms underlying cognition.

While augmenting our behavioral data with neuroimaging data is generally useful, in the joint modeling context there are certain conditions where the additional (neural)

information is not beneficial for all parameters. In particular, if the link between neural and behavioral submodel parameters is not well established (i.e., a near-zero correlation), neither submodel will benefit from the supplementary data, and no additional model constraint will be enforced. To demonstrate this idea, first suppose the following partition:

$$\left[\begin{array}{c}\theta\\\hline\delta\end{array}\right] \sim \mathcal{N}_p\left(\phi = \left[\begin{array}{c}\phi_1\\\hline\phi_2\end{array}\right], \Sigma = \left[\begin{array}{c|c}\Sigma_{1,1} & \Sigma_{1,2}\\\hline\Sigma_{2,1} & \Sigma_{2,2}\end{array}\right]\right)$$

Here, ϕ_1 is a $(p_1 \times 1)$ matrix, and ϕ_2 is a $(p_2 \times 1)$ matrix, where $p_1 + p_2 = p$. Partitioning the parameter space in this way allows us to identify the mean and variance components of the submodel parameter sets. Given the properties of the multivariate normal distribution, there are three important facts that we observe [65, 66].

1. The marginal distributions of θ and δ are multivariate normal. Specifically, $\theta \sim \mathcal{N}_{p_1}(\phi_1, \Sigma_{1,1})$ and $\delta \sim \mathcal{N}_{p_2}(\phi_2, \Sigma_{2,2})$.
2. The conditional distributions of θ and δ are multivariate normal. Specifically, the conditional distribution of θ given that $\delta = \delta^*$ is

$$\theta \mid \delta = \delta^* \sim \mathcal{N}_{p_1}\left(\phi_{1|2}, \Sigma_{1,1|2}\right),$$

   where

$$\begin{aligned}\phi_{1|2} &= \phi_1 + \Sigma_{1,2}\Sigma_{2,2}^{-1}(\delta^* - \phi_2)\\ \Sigma_{1,1|2} &= \Sigma_{1,1} - \Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1}.\end{aligned} \tag{10.4}$$

3. A zero covariance matrix $\Sigma_{1,2}$ implies that θ and δ are independent. Thus, θ and δ are independently distributed if and only if $\Sigma_{1,2} = \mathbf{0}$, where $\mathbf{0}$ indicates a $(p_1 \times p_2)$ matrix of zeros.

Together, Facts 1 and 2 provide some interesting (and perhaps surprising) information about exactly how neural measures constrain behavioral submodel parameters. Equation 10.4 shows that the mean of the conditional distribution of θ depends on the value of δ^*, whereas the variance of the conditional distribution of θ does not. However, both the mean and variance depend on the covariance matrix $\Sigma_{1,2}$. To illustrate the difference between the conditional and marginal distributions of θ, we can examine the absolute difference in the parameters of the two distributions. For illustrative purposes, suppose $p_1 = p_2 = 1$, $\phi_1 = \phi_2 = 0$, and $\Sigma_{1,1} = \Sigma_{2,2} = 1$. We can then examine the differences between the parameters of the conditional distribution of $\theta \mid \delta = \delta^*$ and the marginal distribution of θ by evaluating the differences $\phi_{1|2} - \phi_1$ and $\Sigma_{1,1|2} - \Sigma_{1,1}$ across a range of values for δ^* and $\Sigma_{1,2} = \Sigma_{1,1}^{1/2}\Sigma_{2,2}^{1/2}\rho = \rho$. Figure 10.3 shows these differences across $\delta^* \in [-3, 3]$ and $\rho \in [-1, 1]$ for $\phi_{1|2} - \phi_1$ (left panel) and $\Sigma_{1,1|2} - \Sigma_{1,1}$ (right panel). The left panel shows that the difference between the conditional and marginal distributions is small near the point $(\rho, \delta^*) = (0, 0)$, but that the difference increases in magnitude as one moves away

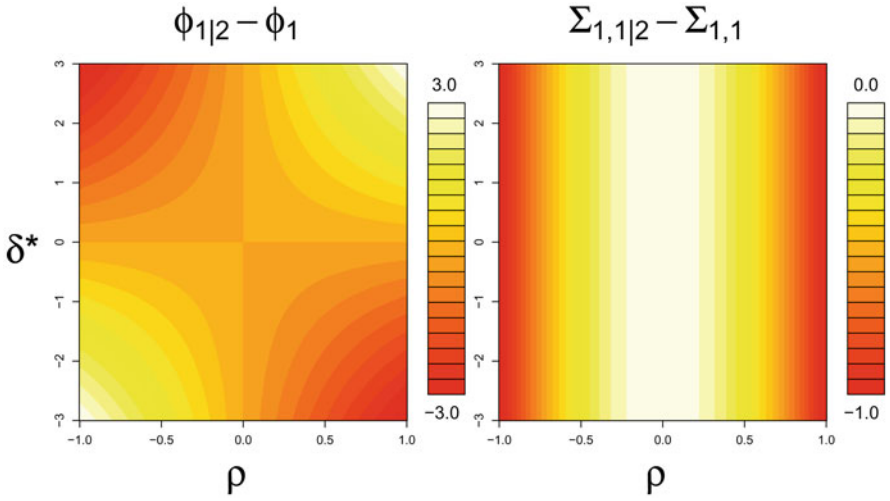


Fig. 10.3 Differences between the conditional and marginal distributions of θ for the hyper parameters ϕ and Σ. For a range of values of ρ and δ^*, we evaluated the difference between the hyper mean parameter $\phi_{1|2}$ (i.e., the mean of the conditional distribution of θ), and ϕ_1 (i.e., the mean of the marginal distribution of θ) shown in the *left* panel, and the difference between the hyper variance parameter $\Sigma_{1,1|2}$ (i.e., the variance of the conditional distribution of θ), and $\Sigma_{1,1}$ (i.e., the variance of the marginal distribution of θ) show in the *right* panel

from this point. The right panel of Fig. 10.3 shows that the difference between the conditional and marginal distributions is small in magnitude when $\rho \approx 0$, otherwise, the difference increases symmetrically about zero. Note that the difference between the two distributions does not depend on δ^*.

Comparing the parameters ϕ and Σ of the conditional and marginal distributions of θ is important because it highlights when—and the degree to which—the added constraint of neural measures is beneficial to behavioral submodel parameters (and vice versa). Figure 10.3 and Eq. 10.4 tell us that the differences observed in these two distributions are most apparent when δ^* and ρ are distant from zero for the mean parameter vector ϕ, but only when ρ is distant from zero for Σ.

Fact 3 is crucial because when $\Sigma_{1,2} = \mathbf{0}$, we gain no additional constraint on the submodel parameters, because θ and δ will be independent, and so no additional information is learned from modeling the neural and behavioral data jointly. From above, $\Sigma_{1,2} = \theta_\sigma \delta_\sigma \rho$, and because $\theta_\sigma > 0$ and $\delta_\sigma > 0$ in nontrivial cases, $\Sigma_{1,2} = \mathbf{0}$ if and only if $\rho = \mathbf{0}$. If follows then, that for the joint modeling framework to be beneficial in constraining submodel parameters, we only require that a single element of ρ be nonzero. In practice, we have found many significantly nonzero correlations between a variety of behavioral and neural measures [23, 46].

10.5 Constraints at Multiple Levels

While the joint models used in Turner et al. [23] connected neural and behavioral submodel parameters at the subject level, it is also possible to establish a connection between these parameters on the individual trial level. To accomplish this, one only needs to define the relationship between the neural and behavioral submodel parameters on a trial-to-trial basis. For the interested reader, we can explain this more formally by reconsidering Eq. 10.3. In the first case, suppose we wish to connect the submodel parameters at the subject level (see Turner et al. [23] for examples). Let $B_{i,j}$ denote the jth subject's behavioral data on the ith trial, and similarly, let $N_{i,j}$ denote the corresponding neural data. Let θ_j and δ_j denote Subject j's behavioral and neural parameters, respectively, and Ω denote the between-subject parameters (i.e., the hyper parameters). Thus, the posterior distribution is

$$p(\theta, \delta, \Omega \mid B, N) \propto \prod_j \left(\prod_i \Big[p(B_{i,j} \mid \theta_j) p(N_{i,j} \mid \delta_j) \Big] p(\theta_j, \delta_j \mid \Omega) \right) p(\Omega). \tag{10.5}$$

To extend the model to the second case where we wish to specify trial-to-trial parameters, let $\theta_{i,j}$ and $\delta_{i,j}$ denote the jth subject's parameters on Trial i. With this new model structure in mind, Eq. 10.5 becomes

$$p(\theta, \delta, \Omega_j \mid B, N) \propto \prod_j \prod_i \Big[p(B_{i,j} \mid \theta_{i,j}) p(N_{i,j} \mid \delta_{i,j}) p(\theta_{i,j}, \delta_{i,j} \mid \Omega_j) \Big] p(\Omega_j), \tag{10.6}$$

where now Ω_j denotes subject-specific hyper parameters. Equation 10.6 expresses the posterior distribution for Subject j only, but to extend the model to multiple subjects, we need only assume a prior distribution over the Ω_js to capture the between-subject variability.

As an example, Turner et al. [46] developed the neural drift diffusion model (NDDM) as a way to connect trial-specific behavioral submodel parameters to trial-specific neural measures. The NDDM is an extension of the drift diffusion model (DDM; [42]) that includes five sources of variability: trial-to-trial variability in the drift rate, nondecision time, start point, and neural activity in particular regions of interest, as well as within-trial variability in the evidence accumulation process. In addition to the trial-to-trial parameters, the model possesses subject-specific parameters, which are held constant across trials. Turner et al. used their model to show how pre-stimulus brain activity (as measured by the BOLD response) could be used to predict the behavioral dynamics in subsequent stimulus information processing, within the confines of the mechanisms posited by the DDM.

Due to the framework's flexibility, there are many other possibilities for developing joint models. For example, one could include multiple neural measures such as those containing simultaneous electroencephalography (EEG) and fMRI recordings. One could then use the spatial characteristics of the two measures to better identify

and amplify the neural signal, and gain even greater constraint on the behavioral model. As another example, one could combine structural properties of a subject's brain such as those obtained by diffusion-weighted imaging (DWI) measurements, with functional, trial-by-trial neural measures. Having a mixture of subject-specific and trial-specific neural measures could inform our cognitive models greatly, and would be easy to implement by assuming different joint modeling structures at different levels of a hierarchical model.

10.6 Incorporation of Singular Measures

Another benefit of joint modeling is the flexible addition of supplementary data that is singular, or containing only a single type of measurement. Adding the singular data will only influence the parameters for a single submodel and will not overwhelm the contribution of the remaining data types in the model. Balancing the relative contributions of the data is advantageous when, for example, the neural data is costly to obtain. As a concrete example, due to budgetary demands, one may only be able to obtain 80 trials of joint behavioral and neural data from a single subject performing a random dot motion task. Following the session, the subject could then perform the same random dot motion task where only behavioral measures are obtained. The data from the two tasks could be combined to provide a better understanding of the subject; however, the additional (singular) behavioral data would only enhance the behavioral submodel parameters and would not overwhelm the neural portion of the model. Such a procedure would allow for greater certainty about the behavioral submodel parameters (i.e., θ), and the overarching (hierarchical) model parameters (i.e., θ_μ and θ_σ in Eq. 10.3).

10.7 Submodel Interchangeability

The final benefit we will discuss in this chapter is submodel interchangeability. In the joint modeling framework, one is not committed to any particular submodel for either the neural or behavioral data. This means that one can choose a submodel on the basis of convenience (e.g., mathematical tractability), theoretical endorsement, or simply personal preference. Being able to easily switch between different models also allows for a direct model comparison by way of fit statistics, and prediction performance—a feature of joint modeling that is similar in spirit to other integrative cognitive model comparison methods [67, 68].

Figure 10.4 illustrates the idea of submodel interchangeability. On the left side, we can choose between various submodels to account for the same behavioral data. On the right side, the submodel for the neural data can remain fixed across models. Under these assumptions, we only need to adjust the likelihood function relating the behavioral submodel parameters to the behavioral data, and the prior distribution

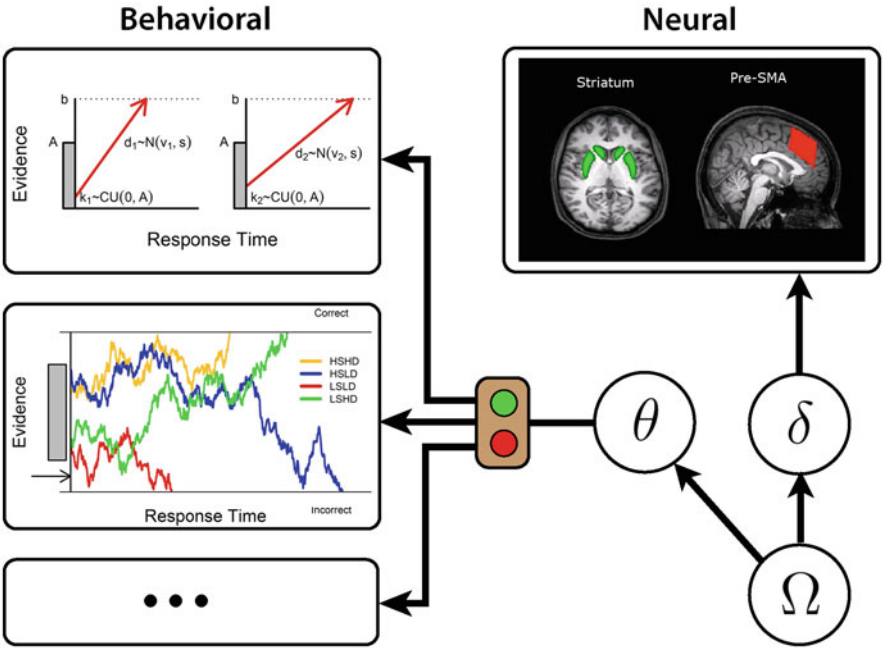


Fig. 10.4 Submodel interchangeability. The joint modeling framework facilitates model comparison because competing submodels can be easily substituted in to account for different aspects of the data

for the behavioral submodel parameters. The only other adjustment that may be necessary is for the size and shape of the hyper parameters connecting the two submodels. For example, the first submodel may only use seven parameters, whereas the second submodel uses eight parameters. In this case, the hyper mean vector ϕ would increase from a length of seven to a length of eight, and the hyper variance-covariance matrix Σ would increase in size from a (7×7) matrix to an (8×8) matrix. Each unique selection for a submodel constitutes a new joint model of the data, and after fitting each model to the data, the relative merits of model fit and generalizability can be properly assessed.

In this section, we will perform such a model comparison between the Linear Ballistic Accumulator (LBA; [69]) model and a DDM [42, 70]. Because the models make very different assumptions about the underlying cognitive processes at work in a choice response time experiment, a comparison of the two models on the basis of model fit would provide support for the assumptions made by the better-fitting model. We note that this model comparison is meant purely for illustrative purposes, and is in no way a definitive claim that one model is better than another. Performing such a task would require many more data sets, and a comparison of multiple variants of each model (see Teodorescu and Usher [71] for such an endeavor). Here, we mean only to compare the models for the first time on the basis of both behavioral and neural data.

10.7.1 Details of the Models

The data are described in Forstmann et al. [16], and consist of 20 young subjects and 14 elderly subjects who participated in a random dot motion task with two alternatives (i.e., detection of mostly leftward or rightward movement). There were three speed emphasis conditions: accuracy, neutral (i.e., self-paced responding), and speed. For the neural data, DWI measures were obtained, which allow one to estimate tract strength, a probabilistic white matter connectivity measure, between different cortico-subcortical brain regions [72]. Based on previous results, four different tract strength measures—between the left and right pre-SMA into the left and right striatum—were obtained. We will use the same neural model as in Turner et al. [23].

For the LBA model, we assume separate thresholds for each speed condition such that $b = \{b^{(1)}, b^{(2)}, b^{(3)}\}$, and separate drift rates for the accumulators corresponding to the correct and incorrect responses, such that $v = \{v^{(1)}, v^{(2)}\}$. We use a bias parameter a such that the upper bound of the start point for each condition is determined by $A^{(k)} = ab^{(k)}$. [2] Finally, we use a nondecision time parameter τ, which is fixed across conditions. To satisfy mathematical scaling properties, we conventionally set the between-trial drift variability $s = 1$.

We will use a modeling strategy for the DDM that is similar to the LBA model. Specifically, we assume separate thresholds for each speed condition such that $\alpha = \{\alpha^{(1)}, \alpha^{(2)}, \alpha^{(3)}\}$. We also assume a bias parameter ω such that the start point for each condition is $z^{(k)} = \omega\alpha^{(k)}$. For the DDM, only one drift rate parameter is needed, which we denote v. Finally, we assume a nondecision time parameter τ, which is fixed across conditions. Note that this model is a drastic simplification of the full DDM, which includes trial-to-trial variability in drift rate, starting point, and nondecision time.

As described above, we apply a transformation so that each parameter has infinite support. Specifically, parameters bounded by zero (e.g., thresholds, nondecision time) were log transformed and parameters bounded by both zero and one (e.g., bias parameters) were logit transformed. The transformation justified the multivariate normality assumption we use on the hyperparameters as in Eq. 10.2. We specify noninformative, dependent priors for ϕ and Σ such that

$$\begin{aligned} \phi \mid \Sigma &\sim \mathbf{MVN}\left(\mu_0, s_0^{-1}\Sigma\right), \text{ and} \\ \Sigma &\sim \mathbf{W}^{-1}(\Phi, d_0), \end{aligned}$$

where $\mathbf{W}^{-1}(a, b)$ denotes the inverse Wishart distribution with dispersion matrix a and degrees of freedom b. We let m denote the length of the parameter ϕ (i.e., $m = 7$ for the DDM and $m = 8$ for the LBA model). We set $s_0 = 1/10$, $d_0 = m$, and $\mu_0 = \mathbf{0}$, a vector of m zeros.

[2] Another alternative is to fix A across conditions [23]. However, we do not use this approach here to maintain consistency with the DDM.

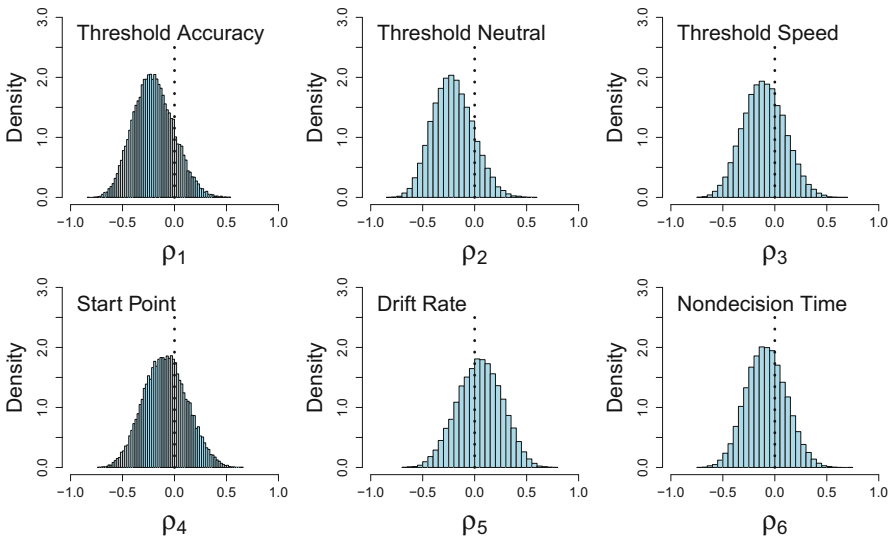


Fig. 10.5 Estimated posterior distributions for the DDM. Each panel shows the estimated correlation parameter of the neural submodel with a particular behavioral submodel parameter: accuracy condition threshold (*top left*), neutral condition threshold (*top middle*), speed condition threshold (*top right*), start point (*bottom left*), drift rate (*bottom middle*), and nondecision time (*bottom right*)

10.7.2 Results

To fit each model, we used a combination of Gibbs sampling and differential evolution with Markov chain Monte Carlo [55, 56]. We ran the algorithm to obtain samples from the joint posterior distributions for 5000 iterations following a burnin period of 5000 iterations with 32 chains. We then thinned the chains by retaining every fifth sample, resulting in 32,000 total samples of the joint posterior distributions.

We then examined the estimated posterior distributions of the parameters relating mean tract strength measurements to the behavioral model parameters. For the LBA model, we observed two differences in the pattern of results reported in Turner et al. [23]. First, the correlations between the drift rates and tract strength was effectively zero, whereas in Turner et al. we observed a strong correlation between the drift rate for the accumulator corresponding to the correct response and tract strength. Second, the correlation between the threshold in the speed condition and tract strength was effectively zero, whereas in Turner et al., we observed a slight positive correlation. We attribute these slight differences in parameter estimates to a slight bias in the model of Turner et al. as a result of constraining the off-diagonal elements to zero. Regardless, both models make very similar predictions for accuracy and RT at both the group and subject levels.

The pattern of correlations between the LBA model parameter and tract strength was similar to the pattern of correlations between the DDM model parameters and tract strength, which we expected based on previous comparisons of these models for behavioral data alone [73, 74]. Figure 10.5 shows the correlation parameters between

the neural submodel parameters and a particular parameter within the (simplified) DDM: accuracy condition threshold (top left), neutral condition threshold (top middle), speed condition threshold (top right), start point (bottom left), drift rate (bottom middle), and nondecision time (bottom right). The correlations with the threshold parameters are similar to the LBA model's account. Specifically, the correlation is negative in the accuracy and neutral conditions, but is essentially zero in the speed condition. The remaining parameters are less clear. There seems to be a slight negative correlation with the start point and nondecision time parameters, and drift rate seems to be uncorrelated with tract strength. However, the estimated posterior distributions for the parameters in each model are highly variable due to a limited number of subjects, and so additional experiments may be required to fully understand these relationships.

A strong negative correlation in the threshold parameters, and a negative correlation of nondecision time in both the DDM and LBA model are consistent with neural explanations of the data. Because two of the tract strength measurements were measures of connectivity between the left and right striatum and the left and right pre-SMA, one might expect that as the strength of the connection increased, it would take less time for the pre-SMA to become activated, and consequently, the amount of time to execute the response would decrease. Because the nondecision time parameter represents (in part) the time to execute the motor response, by this explanation, a negative correlation should be observed. A negative correlation between threshold parameters and tract strength indicates that as tract strength increases, less evidence is required to make a decision, which is indicated in the model as a decrease in the threshold parameter. The observation that the threshold for the speed condition is uncorrelated may reflect a different response strategy than what is used in either the accuracy or neutral conditions.

Perhaps the most important analysis we can perform is to compare the DDM and LBA models. To compare the models, we chose to use conventional Bayesian measures of model fit, which balance all of the important aspects of model comparison: model complexity, number of parameters, and degree of fit to the data [75–77]. We chose the deviance information criterion (DIC; [78]), and the Bayesian predictive information criterion (BPIC; [79]). While the DIC is the more conventional metric in the literature, the DIC has been criticized for not properly penalizing models for their number of parameters [79]. The BPIC was developed as a solution to this problem, however, in practice they tend to produce similar results. For both of these measures, a better fit is indicated by a smaller (i.e., more negative in most cases) value.

In reporting our results, we wished to provide some measure of the variability inherent in each of the fitting statistics. To do so, we calculated the statistic on each chain in our sampler individually, producing a distribution of (32) statistics. We then recorded the mean and standard deviation of this distribution and report them below. The idea behind this is to provide greater support for how the statistics for one model compare to the statistics from the other model.

We performed the model comparison in two ways. First, because the two models were fit hierarchically to the data, we can compare the models on the basis of the full data set. The DIC statistic for the DDM was −6537.14 (14.51) and for the LBA was −9149.58 (6.31). The BPIC value was −6520.32 (18.50) for the DDM and −9089.2

Table 10.1 DDM and LBA model fit statistics to the data of Forstmann et al. (2011). Standard deviations of each fit statistic (across chains) appear in parentheses

	DIC		BPIC	
Subject	DDM	LBA	DDM	LBA
1	−39.93 (0.48)	−262.01 (0.87)	−32.59 (0.77)	−253.82 (1.36)
2	−1704.48 (0.47)	−1737.74 (0.84)	−1698.23 (0.76)	−1730.95 (1.34)
3	−236.25 (0.51)	−570.68 (1.05)	−230.09 (0.8)	−563.71 (1.61)
4	−1604.91 (0.53)	−1526.61 (0.74)	−1598.78 (0.85)	−1519.72 (1.11)
5	−835.7 (0.55)	−831.55 (0.81)	−829.3 (0.84)	−824.02 (1.27)
6	−2099.74 (0.56)	−2204.22 (0.67)	−2093.56 (0.87)	−2197.22 (1.07)
7	−400.73 (0.56)	−400.1 (0.63)	−394.59 (0.86)	−393.06 (1.01)
8	313.23 (0.67)	−14.54 (0.73)	320.07 (1.04)	−6.66 (1.13)
9	817.36 (0.48)	801.76 (1.02)	823.51 (0.72)	809.02 (1.55)
10	−776.59 (0.53)	−1153.34 (1.11)	−769.52 (0.82)	−1146.2 (1.63)
11	−1158.3 (0.51)	−1169.62 (0.81)	−1152.08 (0.77)	−1162.55 (1.28)
12	−121.28 (0.6)	−238.29 (0.95)	−114.26 (0.94)	−230.99 (1.48)
13	−33.8 (0.44)	−215.17 (1.04)	−27.64 (0.7)	−208.79 (1.6)
14	103.96 (0.44)	100.89 (0.8)	110.15 (0.69)	108.81 (1.31)
15	−490.56 (0.45)	−620.58 (1.2)	−484.06 (0.71)	−613.67 (1.78)
16	271.26 (0.58)	2.22 (0.87)	277.27 (0.9)	8.28 (1.36)
17	1.36 (0.51)	−98.52 (0.66)	7.51 (0.8)	−91.22 (1.1)
18	490.5 (0.57)	365.78 (0.72)	497.06 (0.91)	373.29 (1.16)
19	122.09 (0.4)	123.14 (1.33)	128.25 (0.64)	130.61 (1.97)
20	532.8 (0.65)	448.55 (1.23)	540.09 (1.05)	457.47 (1.97)
21	188.12 (0.52)	157.42 (1.96)	194.34 (0.8)	167.42 (3.03)
22	219.3 (0.76)	117.94 (1.19)	226.36 (1.22)	126.51 (1.95)
23	−361.79 (0.63)	−367.21 (0.64)	−355.01 (0.99)	−359.72 (0.98)
24	405.21 (0.48)	261.57 (0.83)	411.54 (0.74)	269.11 (1.38)
Total wins	4	20	4	20

(11.94) for the LBA model. The effective number of parameters (pD) was 16.82 (6.42) for the DDM and 60.38 (5.86) for the LBA model. Thus, the LBA model fit the grouped data better than the DDM, but in a way that was more complex. This could suggest that the version of the DDM we used here was too simple, and that trial-to-trial variability is required to account for these data.

The second comparison we can make is at the individual subject level. Table 10.1 shows the DIC and BPIC values for each of the 24 subjects for each model. In comparing the models, the statistics were precise enough such that only a few subjects

had distributions that overlapped. The table shows that the LBA model fit the data for 20 of the 24 subjects better than did the simplified version of the DDM used here.

10.8 Concluding Comments

In this chapter, we have discussed some of the benefits of the joint modeling approach in greater detail [23]. We began by discussing the technical details of the method and justifying some of its assumptions, while discussing possible alternatives. We then illustrated how to use the model to make predictions for unobserved data, and argued for our generative modeling approach over discriminative ones. Next, we showed how adding more data to a model can better constrain the model under most circumstances, and examined the parameter space for varying degrees of constraint. We then acknowledged that the joint modeling idea can be implemented at different or multiple levels within a hierarchical model, and that adding additional behavioral or neural measures had no effect on the opposing side of the model. Finally, we compared a joint model version of the DDM and the LBA model by fitting them both to the data of Forstmann et al. [16]. Under the myopic model variants we used here, the LBA model provided a better fit than did the DDM.

In closing, the joint modeling framework provides a convenient way for cognitive modelers to conjoin their favorite behavioral model to neural data. The approach does not require extensive knowledge of how the brain operates, nor does it require that one have a priori hypotheses about how neural measures are related to behavioral model parameters. The modeling approach used here provides a means for enforcing a statistically reciprocal relationship between cognitive modeling and cognitive neuroscience [22].

Exercises

1. Can you describe (in greater detail) why discriminative approaches will generally outperform generative approaches in a prediction task?
2. In many cognitive models, we must fix a subset of parameters to some arbitrary value in order to fully identify the remaining parameters. Some would argue that this arbitrary selection is an undesirable property of the model. Might the joint modeling approach help identify these (scaling) parameters?
3. To what extent can the joint modeling framework be further extended? For example, is it possible for the joint modeling framework to be extended from a trial-to-trial basis to a second-by-second basis?

Further Reading

1. The original paper on joint modeling provides a few examples and is more thorough in the technical aspects of the approach [23].
2. A follow-up paper extends joint modeling to the single-trial level for the drift diffusion model [46].
3. Much of Anderson and colleagues' [10–12, 80, 81] approach to combining neural and behavioral measures is more constrained and certainly more mechanistic than our joint modeling approach.

References

1. O'Reilly RC (2001) Neural Comput 13:1199
2. O'Reilly RC (2006) Science 314:91
3. O'Reilly R, Munakata Y (eds) (2000) Computational explorations in cognitive neuroscience: understanding the mind by simulating the brain. MIT Press, Cambridge
4. Mazurek ME, Roitman JD, Ditterich J, Shadlen MN (2003) Cereb Cortex 13:1257
5. Usher M, McClelland JL (2001) Psychol Rev 108:550
6. Shadlen MN, Newsome WT (2001) 86:1916
7. deLange FP, Jensen O, Dehaene S (2010) 30:731
8. de Lange FP, van Gaal S, Lamme VAF, Dehaene S (2011) 9:e1001203
9. O'Connell RG, Dockree PM, Kelly SP (2012) Nat Neurosci 15:1729
10. Anderson JR (2007) How can the human mind occur in the physical universe? Oxford University Press, New York
11. Anderson JR, Carter CS, Fincham JM, Qin Y, Ravizza SM, Rosenberg-Lee M (2008) Cognit Sci 32:1323
12. Anderson JR, Qin Y, Jung KJ, Carter CS (2007) Cognit Psychol 54:185
13. van Vugt MK, Simen P, Nystrom LE, Holmes P, Cohen JD (2012) Front Neurosci 6:1
14. Forstmann BU, Anwander A, Schäfer A, Neumann J, Brown S, Wagenmakers EJ, Bogacz R, Turner R (2010) Proc Natl Acad Sci 107:15916
15. Forstmann BU, Dutilh G, Brown S, Neumann J, von Cramon DY, Ridderinkhof KR, Wagenmakers EJ (2008) Striatum and pre–SMA facilitate decision–making under time pressure. Proc Natl Acad Sci 105:17538
16. Forstmann BU, Tittgemeyer M, Wagenmakers EJ, Derrfuss J, Imperati D, Brown S (2011) J Neurosci 31:17242
17. Ratcliff R, Philiastides MG, Sajda P (2009) Proc Natl Acad Sci U S A 106:6539
18. Philiastides MG, Ratcliff R, Sajda P (2006) J Neurosci 26:8965
19. Ho T, Brown S, Serences J (2009) J Neurosci 29:8675
20. Liu T, Pleskac TJ (2011) J Neurophysiol 106:2383
21. Tosoni A, Galati G, Romani GL, Corbetta M (2008) Nat Neurosci 11:1446
22. Forstmann BU, Wagenmakers EJ, Eichele T, Brown S, Serences JT (2011) Reciprocal relations between cognitive neuroscience and formal cognitive models: Opposites attract? Trends Cognit Sci 15:272
23. Turner BM, Forstmann BU, Wagenmakers EJ, Brown SD, Sederberg PB, Steyvers M (2013) NeuroImage 72:193
24. Gershman SJ, Blei DM, Pereira F, Norman KA (2011) Neuroimage 57:89
25. Guo Y, Bowman FD, Kilts C (2008) Hum Brain Mapp 29:1092
26. Kershaw J, Ardekani BA, Kanno I (1999) IEEE Trans Med Imaging 18:1138
27. Quirós A, Diez RM, Gamerman D (2010) NeuroImage 49:442

28. Van Gerven MAJ, Cseke B, de Lange FP, Heskes T (2010) NeuroImage 50:150
29. Wu W, Chen Z, Gao S, Brown EN (2011) NeuroImage 56:1929
30. Dennis S, Lee M, Kinnell A (2008) J Math Psychol 59:361
31. Lee MD (2008) Psychon Bull Rev 15:1
32. Lee MD (2011) J Math Psychol 55:1
33. Rouder JN, Lu J (2005) Psychon Bull Rev 12:573
34. Rouder JN, Lu J, Speckman P, Sun D, Jiang Y (2005) Psychon Bull Rev 12:195
35. Rouder JN, Sun D, Speckman P, Lu J, Zhou D (2003) Psychometrika 68:589
36. Shiffrin RM, Lee MD, Kim W, Wagenmakers EJ (2008) Cognit Sci 32:1248
37. Oravecz Z, Tuerlinckx F, Vandekerckhove J (2009) Psychometrika 74:395
38. Vandekerckhove J, Tuerlinckx F, Lee MD (2011) Psychol Methods 16:44
39. Lee MD, Wagenmakers EJ (2012) A course in Bayesian graphical modeling for cognitive science. http://www.ejwagenmakers.com/BayesCourse/BayesBookWeb.pdf. Accessed 1 Jan 2012
40. Green DM, Swets JA (1966) Signal detection theory and psychophysics. Wiley, New York
41. Dennis S, Humphreys MS (2001) Psychol Rev 108:452
42. Ratcliff R (1978) Psychol Rev 85:59
43. Frank LR, Buxton RB, Wong EC (1998) Magn Reson Med 39:132
44. Flandin G, Penny WD (2007) NeuroImage 34:1108
45. Friston K (2002) NeuroImage 16:513
46. Turner BM, van Maanen L, Forstmann BU (2014) Informing cognitive abstractions through neuroimaging: The Neural Drift Diffusion Model. In press at Psychological Review.
47. Eichele T, Debener S, Calhoun VD, Specht K, Engel AK, Hugdahl K, von Cramon DY, Ullsperger M (2008) Proc Natl Acad Sci U S A 16:6173
48. Gelman A, Carlin JB, Stern HS, Rubin DB (2004) Bayesian data analysis. Chapman and Hall, New York
49. Christensen R, Johnson W, Branscum A, Hanson TE (2011) Bayesian ideas and data analysis: an introduction for scientists and statisticians. CRC Press, Taylor and Francis Group, Boca Ranton
50. Lunn D, Thomas A, Best N, Spiegelhalter D (2000) Stat Comput 10:325
51. Plummer M (2003) Proceedings of the 3rd International Workshop on Distributed Statistical Computing
52. Robert GO, Sahu S (1997) J Royal Stat Soc B 59:291
53. Liu JS, Sabatti C (2000) Biometrika 87:353
54. Hoffman MD, Gelman A (2011) The no-u-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. Manuscript submitted for publication
55. ter Braak CJF (2006) Stat Comput 16:239
56. Turner BM, Sederberg PB, Brown SD, Steyvers M (2014) Psych Methods 18:368–384
57. Wilkinson RD (2011) Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. Manuscript submitted for publication
58. Wood S (2010) Nature 466:1102
59. Turner BM, Sederberg PB (2012) J Math Psychol 56:375
60. Turner BM, Sederberg PB (2014) Psychonomic Bulletin and Review 21:227–250
61. Navarro DJ, Fuss IG (2009) J Math Psychol 53:222
62. Bishop CM, Lasserre J (2007) Bayesian Stat 8:3
63. Norman KA, Polyn SM, Detre GJ, Haxby JV (2006) Trends Cognit Sci 10:1
64. Polyn SM, Natu VS, Cohen JD, Norman KA (2005) Science 310:1963
65. Johnson RA, Wichern DW (2007) Applied multivariate statistical analysis. Pearson Prentice Hall, Upper Saddle River
66. Rice JA (2007) Mathematical statistics and data analysis. Duxbury Press, Belmont
67. Purcell B, Heitz R, Cohen J, Schall J, Logan G, Palmeri T (2010) Psychol Rev 117:1113
68. Mack ML, Preston AR, Love BC (2013) Decoding the brain's algorithm for categorization from its neural implementation. In press at Current Biology
69. Brown S, Heathcote A (2008) Cognit Psychol 57:153

70. Feller W (1968) An introduction to probability theory and its applications, vol 1. Wiley, New York
71. Teodorescu AR, Usher M (2013) Psychol Rev
72. Behrens T, Johansen-Berg H, Woolrich MW, Smith SM, Wheeler-Kingshott CA, Boulby PA, Barker GJ, Sillery EL, Sheehan K, Ciccarelli O, Thompson AJ, Brady JM, Matthews PM (2003) Nat Neurosci 6:750
73. Donkin C, Heathcote A, Brown S (2009) In: Howes A, Peebles D, Cooper R (eds) 9th International Conference on Cognitive Modeling—ICCM2009. Manchester, UK
74. Donkin C, Brown S, Heathcote A, Wagenmakers EJ (2011) Psychon Bull Rev 18:61
75. Myung IJ (2000) J Math Psychol 44:190
76. Myung IJ, Forster M, Browne MW (2000) J Math Psychol 44:1
77. Pitt MA, Myung IJ, Zhang S (2002) Psychol Rev 109:472
78. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A (2002) J Royal Stat Soc B 64:583
79. Ando T (2007) Biometrika 94:443
80. Anderson JR, Betts S, Ferris JL, Fincham JM (2010) Proc Natl Acad Sci U S A 107:7018
81. Anderson JR, Fincham JM, Schneider DW, Yang J (2012) NeuroImage 60:633

Chapter 11
Predictive Coding in Sensory Cortex

Peter Kok and Floris P. de Lange

Abstract In recent years, predictive coding has become an increasingly influential model of how the brain processes sensory information. Predictive coding theories state that the brain is constantly trying to predict the inputs it receives, and each region in the cortical sensory hierarchy represents both these predictions and the mismatch between predictions and input (prediction error). In this chapter, we review the extant empirical evidence for this theory, as well as discuss recent theoretical advances. We find that predictive coding provides a good explanation for many phenomena observed in perception, and generates testable hypotheses. Furthermore, we suggest possible avenues for further empirical testing and for broadening the perspective of the role predictive coding may play in cognition.

11.1 Introduction

In recent years, predictive coding has become an increasingly influential model of how the brain processes sensory information [1–3]. This model challenges the traditional view of sensory cortex as a unidirectional hierarchical system that passively receives sensory signals and extracts increasingly complex features as one progresses up the hierarchy. Instead, predictive coding theories state that the brain is constantly trying to predict the inputs it receives, and each region in the sensory hierarchy represents both these predictions and the mismatch between predictions and input (prediction error). Moreover, regions in the sensory hierarchy continually interact, informing each other about what they expect the other region is observing, and how this expectation matches their input.

In this chapter, we will review recent theoretical and empirical advances in the field of predictive coding. First, we will outline the motivations behind predictive coding, and discuss its principal features (§ 2). Then, we will review empirical evidence from cognitive neuroscience for several basic tenets of predictive coding (§ 3), and discuss one of them—the implementation of attention in predictive coding—in more detail

P. Kok (✉) · F. P. de Lange
Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen,
6500 HB, Nijmegen, The Netherlands
e-mail: p.kok@donders.ru.nl

B. U. Forstmann, E.-J.Wagenmakers (eds.), *An Introduction to Model-Based Cognitive Neuroscience,* DOI 10.1007/978-1-4939-2236-9_11

(§ 4). We will end with a discussion of the limitations of current empirical foundations of predictive coding, and suggest future directions to strengthen these foundations and extend the perspective on predictive coding to other cognitive domains (§ 5).

11.2 Predictive Coding

11.2.1 Perception as Inference

One of the major motivations behind the development of predictive coding models of sensory processing has been the observation that perception is not solely determined by the input to our eyes, but is strongly influenced by our expectations. Over a century ago, Helmholtz [4] cast perception as a process of unconscious inference, wherein perception is determined by both sensory inputs and our prior experience with the world. For example, many perceptual illusions can be explained as the result of our prior knowledge of the world influencing perceptual inference [5, 6]. When we are presented with four 'Pac-Man' figures arranged in a certain way, we perceive an illusory square (Kanizsa square; Fig. 11.1a). Presumably, the brain infers that the most likely cause for such an input, given its prior experience of the world, is a white square overlaying four black circles. Note that occlusion is a ubiquitous feature of the visual world, and inferring the presence of whole objects despite this is key to successful perceptual inference. Furthermore, priors provided by the larger context can help disambiguate local details. For example, the same figure can be perceived as the letter '*B*' or the number '13', depending on the surrounding figures ('*A*' and '*C*' or '12' and '14', respectively; Fig. 11.1b). Finally, prior knowledge can improve perception by 'explaining away' predictable features (e.g., stripy leaves), leaving unexpected (and potentially vitally important) features to stand out (a tiger!) (Fig. 11.1c). That is, without prior knowledge of the world the image in Fig. 11.1c might look like a mass of incoherent lines, while recognising the majority of the lines as parts of plants allows 'subtracting' them from the image. Any features that cannot be explained away as part of the plants (the stripes on the tiger's back and head) will be all the more salient. In recent years, the idea of perception as inference has enjoyed a revival, benefitting from converging ideas from computer vision research and neuroscience [3, 7–10].

11.2.2 Coding Scheme

One model of sensory processing that describes perception as fundamentally inferential is predictive coding [1–3, 11]. In this model (Fig. 11.2a), each cortical sensory region contains two functionally distinct sub-populations of neurons. Prediction (P) units represent the hypothesis that best explains the input the region receives, while prediction error (PE) units represent that part of the input that is not explained by the current hypothesis, i.e. the mismatch between input and prediction. Connected

Fig. 11.1 Examples of perceptual inference. **a** Kanizsa *square*: four 'Pac-Man' figures or a *white* square overlaying *black* circles? **b** Context resolves ambiguity: is the figure in the centre the letter '*B*' or the number '13'? **c** Prior knowledge improves processing of noisy sensory inputs: 'explaining away' the leaves makes the tiger stand out more

regions in the cortical hierarchy interact recurrently in a joint effort to find the world model that best explains the sensory inputs in the P units, and thereby reduce the activity of the PE units. This interaction takes place as follows: (1) The PE in one region serves as input to the next region in the cortical hierarchy, triggering that region to select a hypothesis that better matches its input. Note that the representational content of PEs (as opposed to an unspecific "surprise" signal) allows for selection of specific hypotheses in the higher order region. (2) The (newly) selected higher order hypothesis is subsequently sent back as a prediction to the lower order region, where it is compared to the current lower level hypothesis, and (3) the mismatch is represented as the (new) prediction error. The above describes one cycle of hypothesis testing in the predictive coding framework. This is an iterative process, culminating in a state in which PE units are all silenced and an accurate representation of the current sensory world is represented by activity in the relevant P units.

Since top-down predictions suppress expected sensory input (i.e., reduce prediction error), expected stimuli lead to relatively little neuronal firing. Such a coding scheme has several advantages. First, it is metabolically efficient. Second, it makes unexpected (and potentially highly relevant) stimuli more salient: if you 'explain away' the striped leaves, the crouching tiger stands out even more (Fig. 11.1c). In fact, it has been proposed that saliency might be equated to the strength of the prediction error [12, 13]. Third, while expected stimuli result in reduced firing in the PE units, the stimulus *representation in the P units* is enhanced [14]. A valid prediction leads to the proper hypothesis being selected prior to sensory input, and since this hypothesis quickly silences these sensory inputs (prediction error) when they arrive, alternative hypotheses are not given a chance to compete (Fig. 11.2b). (Note that this pre-selection does not prevent potentially relevant unexpected inputs from being processed, since such inputs will lead to a large PE, attracting attention and triggering selection of alternative hypotheses, see Fig. 11.2c). In other words, pre-selection of the valid hypothesis makes the stimulus representation more unequivocal, or sharper. Such use of prior expectations helps us make sense of the ambiguous and noisy sensory inputs we receive in everyday life [15]. For this aspect of perceptual inference, the hierarchical nature of predictive coding is crucial [9, 16]. Inference on fine scale

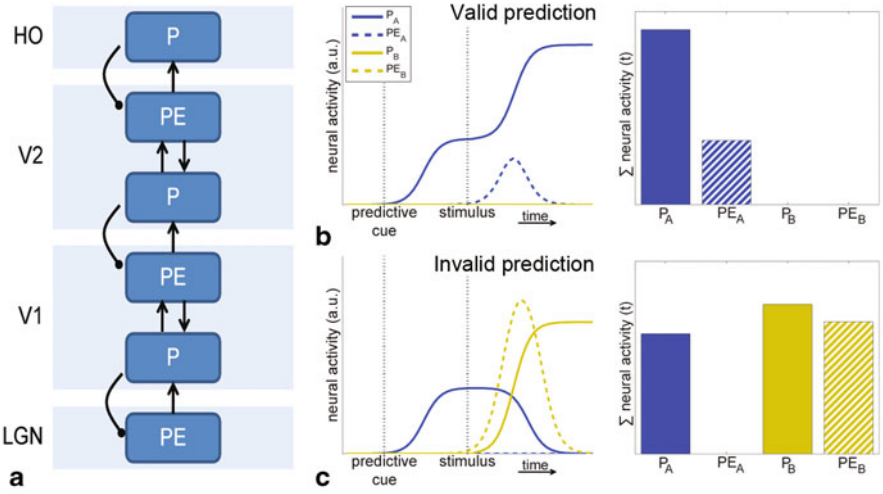


Fig. 11.2 Predictive coding. **a** Schematic predictive coding architecture, with PE units providing excitatory feedforward input, and *P* units providing inhibitory feedback. **b** Hypothesised neural activity in two populations of *P* and PE units, each representing a different hypothesis ('*A*' and '*B*', respectively). Here, stimulus *A* is predicted, and subsequently presented (valid prediction). *Left* panel illustrates schematic timecourse of activity, *right* panel provides integral over time (i.e., proxy of BOLD amplitude). **c** Here, stimulus *A* is predicted, but *B* is presented. Activity is higher overall (particularly in PE units), but less unequivocal (in *P* units)

low level features (black and white stripes, in a seemingly random arrangement) benefits from high level representations (a tiger in a bush). In turn, high level representations can be refined by the high resolution information present in lower order visual areas, e.g., progressing from a coarse representation ('a face') to one reflecting the identity and emotional expression of that face [17].

In a slightly different take on hierarchical inference, Lee and Mumford [9] proposed a model wherein hypotheses at one level reinforce consistent hypotheses at the lower level. In their approach, multiple hypotheses are kept alive at each level of the cortical hierarchy, and excitatory feedback helps the most likely lower level hypothesis to win the competition. In other words, excitatory feedback collapses the lower level hypothesis space and thereby reduces the overall level of neuronal activity. Strictly taken, this is not a predictive coding model (there is no explicit error representation), but it shares many of its key features (hierarchical perceptual inference) as well as empirical predictions (valid top-down hypotheses lead to reduced activity but improved representations).

11.2.3 Neural Implementation

Several different proposals for the neural architecture underlying predictive coding have been made [1, 3, 18, 19]. All these accounts hypothesise the existence of

separate sub-populations of P and PE neurons, and suggest that these neurons reside in different cortical layers. A major difference lies in the type of information cortical areas exchange: in classical predictive coding schemes [1–3] PE units are the source of feedforward and the target of feedback connections, while in Spratling's PC/BC model [18] errors are processed intracortically, and P units are reciprocally connected between regions. These schemes result in different predictions regarding the location of the sub-populations of P and PE units, based on known interlaminar and intercortical connectivity patterns [19]. Feedforward connections mainly arise from layers 2/3 and send input to layer 4 of the next higher-order region in the hierarchy, while feedback is sent from layers 5/6 to the agranular layers of the lower-order region [20–22]. Therefore, if feedforward connections carry prediction errors, PE units would be expected to reside in layers 2/3, while feedback-sending *P* units would reside in layers 5/6 [23]. In the PC/BC model, separate populations of P units would reside in layers 2/3 and 5/6, sending forward and backward predictions, respectively, while PE units reside in layer 4, which does not have interregional outputs. Note that such a separation of forward and backward messages seems necessary for hierarchical inference [9, 19], since these messages need to be tailored to higher-order (larger, more complex receptive fields) and lower-order (smaller, simpler receptive fields) regions, respectively. While these schemes differ in terms of the details of neural implementation, they are computationally equivalent [18].

11.3 Empirical Evidence for Predictive Coding

The recurrent interaction between prediction and prediction error units that characterises predictive coding models leads to several hypotheses than can be empirically tested. For example, since perception reflects an integration of top-down expectations and bottom-up sensory input, the same sensory input should lead to different responses depending on the strength and validity of the expectation. Specifically, the amplitude of the stimulus-evoked response should be lower, the more expected the input is (i.e., the less prediction error there is). Also, top-down expectations may activate hypotheses in sensory regions in the absence of sensory input. Further empirical validation of predictive coding may come from assessing its neural substrate: are there separate units coding predictions and prediction errors? We will discuss each of these points in turn, reviewing evidence from neuroimaging, electrophysiology, and physiology.

11.3.1 Encoding of Surprise in the Brain

One of the most robust modulations of sensory responses is repetition suppression (RS): when a stimulus is repeated, the neural response to the second stimulus is reduced compared to the first. This effect holds across a range of modalities, stimulus

properties, and brain areas, and has been considered the result of stimulus-induced neural adaptation [24]. However, if prediction is indeed a fundamental feature of sensory processing, RS may (partly) reflect the fact that the initial presentation of the stimulus induces an expectation of that same stimulus reappearing [25]. To test this hypothesis, Summerfield et al. [26] used functional magnetic resonance imaging (fMRI) to compare the neural response to stimulus repetitions and alternations, in two different contexts. In one context, a face stimulus was likely to be repeated, while in the other it was likely to be followed by a different face. These researchers showed that when stimulus repetitions were likely (i.e., expected), repeated stimuli led to a strongly reduced neural response compared to alternations (strong RS). When repetitions were unlikely however, the RS effect was strongly reduced, suggesting that RS at least partly reflects predictability. Since this study used fMRI to investigate neural activity, the time course of the RS effect (and its modulation by predictability) could not be resolved. Therefore, it was unclear whether predictability had an immediate suppressive effect on the expected sensory signal (prediction error suppression), or whether surprising events (alternations in one context, repetitions in the other) resulted in a reorienting of attention, with the reported effects of predictability reflecting later attentional modulations. In an effort to distinguish between these possibilities, Todorovic et al. [27] used magneto-encephalography (MEG) to investigate the time course of the effects of predictability on RS in auditory cortex. They found that predictability affected early stimulus-evoked components in auditory cortex, from 100 ms post-stimulus onwards (Fig. 11.3a; see also 28, for similar findings in monkey inferotemporal cortex using visual stimuli). Such early modulations are not in line with a late attention effect, but rather suggest predictive suppression of sensory signals. Furthermore, in a follow-up study, Todorovic and De Lange [29] reported dissociable time courses for the effects of repetition (i.e., stimulus-induced adaptation) and predictability, suggesting that prediction has suppressive effects independent of those of bottom-up adaptation. These and other studies [30–32] clearly show that prediction suppresses expected sensory signals.

Other studies have investigated violations of more high-level sensory predictions. One example is apparent motion: Static visual stimuli presented successively at separate spatial locations induce the illusory perception of motion between these locations. Areas of the primary visual cortex (V1) that correspond retinotopically to visual stimulation along the trajectory of illusory motion, but that are not directly stimulated by the static stimuli, have been shown to be active during perception of apparent motion [33]. Presumably, this is caused by higher level motion sensitive areas with larger receptive fields (i.e., MT/V5) inferring a moving stimulus and sending predictions of this inferred stimulus back to the corresponding locations in V1 [34–36]. Interestingly, the study by Ahmed et al. [36] was performed in anaesthetised ferrets, suggesting that this predictive feedback is not dependent on attention, but rather reflects automatic processes inherent to sensory processing in the visual cortical hierarchy. In a recent study, Alink et al. [37] reasoned that if these feedback signals indeed reflect predictions, they should affect the processing of stimuli presented along the apparent motion trajectory. Specifically, stimuli presented in temporal alignment with the inferred motion path should evoke less prediction error

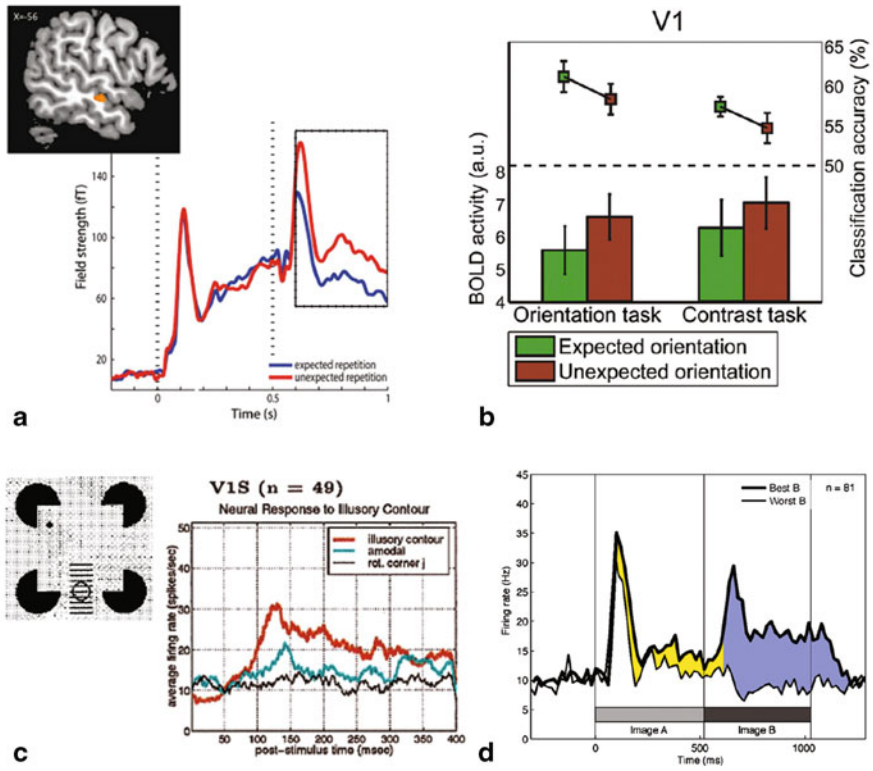


Fig. 11.3 Empirical evidence for predictive coding. **a** Unexpected stimulus repetitions evoke more activity in auditory cortex than expected repetitions. Reprinted from [27] with permission from the authors. **b** Grating stimuli with an expected orientation evoke less activity in V1 (*green* and *red* bars), but this activity contains more orientation information (*green* and *red* squares). This effect is independent of feature attention: it holds both when orientation is task relevant (leftmost bars) and when it is task irrelevant (rightmost bars). Reprinted from [14] with permission from the authors. **c** Illusory contours evoke activity in V1 cells with a receptive field on the contour, presumably as a result of feedback from higher order regions (Reprinted from [59]). **d** Predictive activity in macaque temporal cortex. After paired-association learning, neurons fire more strongly when the first stimulus of a pair (*A*) predicts that the second stimulus (*B*) will be their preferred stimulus ('Best *B*', thick line), than when stimulus A predicts a non-preferred stimulus B ('Worst B', thin line). Increased firing to preferred stimuli is present not only after stimulus presentation (*blue* shading), but already before stimulus presentation (*yellow* shading). Reprinted from [28]

than stimuli that are not temporally aligned. Indeed, Alink et al. [37] found that a visual stimulus presented along the apparent motion path in temporal alignment with the apparent motion evoked a reduced activation in V1, compared to when it was not temporally aligned. Presumably, such a non-aligned stimulus violates top-down predictions and therefore causes a larger prediction error.

Predictions operate not only within, but also between sensory modalities. An everyday example of this is the perception of audiovisual speech. In natural speech,

visual inputs (i.e., mouth movements) precede auditory input by 100–300 ms [38]. This allows the brain to make predictions about the auditory speech signals before their arrival. Indeed, presence of visual speech signals improves speech perception [39], and speeds up cortical processing of auditory speech [40, 41]. Furthermore, when visual and auditory signals mismatch, this can result in distorted or merged percepts [42]. For example, simultaneous presentation of auditory 'ba' and visual 'ga' signals is perceived as 'da'. If visual speech signals indeed result in predictions being sent to auditory sensory areas, mismatch of visual and auditory signals should lead to increased PE in auditory cortex, compared to matching visual and auditory signals. This is indeed what was found by Arnal et al. [41, 43], who used both fMRI and MEG to characterise the response to audiovisual mismatches. Their results show an increased response to incongruent audiovisual stimuli in the superior temporal sulcus—a multisensory region—as well as increased gamma activity in the auditory cortex [43]. Both these responses scaled with the amount of predictive information contained by the visual stimulus: the more informative the visual stimulus was regarding the syllable being pronounced, the stronger the PE when the subsequent auditory stimulus did not match.

Studies on audiovisual speech exploit predictions learned over a lifetime. Recent studies have also shown effects of predictions across modalities when these predictions are learned over the course of the experiment [14, 44, 45]. For example, Den Ouden et al. [44] presented auditory cues that predicted with 80 % likelihood that a visual stimulus would appear. When a visual stimulus was preceded by such a cue, the activity it evoked in V1 was reduced compared to when it was not preceded by a predictive cue. Remarkably, the *omission* of a predicted visual stimulus also evoked more activity in V1 than the omission of a non-predicted stimulus. In this study, both the auditory and visual stimuli were completely irrelevant to participants' task. These results demonstrate that predictions are learned rapidly, even when irrelevant to the task at hand, and affect sensory responses at the earliest stages of cortical processing. In line with this, we [14] found that auditory cues that predicted the features of a visual stimulus led to reduced activity in V1. Specifically, when the pitch of a preceding auditory tone correctly predicted the orientation of a subsequent grating stimulus, the response to this grating in V1 was reduced, compared to when the prediction was invalid. Furthermore, this study investigated not only the amplitude of the neural response evoked by the stimuli, but also used multivariate pattern analysis (MVPA) methods to probe the amount of information contained in the neural signal. Interestingly, we found that a valid orientation prediction led to a *decrease* in the amplitude of the neural signal in V1, but to an *increase* in the amount of information about the grating orientation in the signal (Fig. 11.3b). This is exactly the pattern of results that is predicted by predictive coding theories of perception (cf. Fig. 11.2b, c): valid predictions lead to selection of the proper hypothesis prior to sensory input, allowing this hypothesis to quickly suppress the sensory signal when it comes in (prediction error suppression), thereby preventing activation of alternative hypotheses (representational sharpening). These results suggest that the population level neural signals measured in humans (with fMRI or EEG/MEG) are a mixture of prediction and prediction error signals [46].

While all these studies exemplify suppressive effects of sensory predictions on the basis of learned contingencies between events in the external world, potentially the largest source of sensory prediction is derived internally, from our motor system. Namely, any movement gives rise to largely predictable sensory input, which according to the predictive coding framework should therefore be suppressed. Some of the clearest demonstrations of this phenomenon so far have come from fish [47]. Many fish are equipped with electroreceptors, allowing them to detect nearby objects (e.g., other fish) through changes in the electric field around them. However, these fishes' own movements (and in some species, self-generated electric currents) also cause disturbances in the electric field around them, such that detecting non-self objects would benefit from suppressing such self-generated signals. Indeed, several types of predictive signals (arising from corollary discharge, proprioception, and higher level electrosensory regions) have been shown to evoke *negative images* of the predicted sensory input in the electrosensory organs of these fish [47]. When the predicted inputs arrive, they are cancelled out by these negative images, enhancing sensitivity to non-self generated signals. These predictive signals have been shown to be highly plastic—when paired with an artificially generated stimulus they adapt within minutes—and highly precise in terms of timing, amplitude, and spatial location. Similar predictive suppression mechanisms have been observed in humans [48–52].

As noted above, a crucial feature of predictive coding is its hierarchical nature: valid high-level hypotheses can enhance representations through reducing prediction error in lower-order regions. Murray and colleagues [53, 54] have shown that when stimuli have lower level features that can be grouped into a higher order shape there is increased activity in shape-selective area LOC, but decreased activity in V1, compared to stimuli for which no such grouping takes place. The researchers ensured that the stimuli were matched for low-level features, precluding an explanation in terms of physical differences between the conditions. Presumably, the inferences of high level areas are subtracted from the incoming sensory signals in lower order areas, leading to reduced activity in V1 whenever such a high level hypothesis is generated.

11.3.2 Encoding of Predictions in the Brain

In a predictive coding framework of perception, prior expectations may be hypothesised to activate representations of predicted stimuli prior to sensory input [55]. One way to test this is to probe activity in sensory cortex when a stimulus is predicted, but no bottom-up input is subsequently provided. In line with this, recent studies have shown increased responses to unexpectedly omitted stimuli in early sensory cortex [44, 56], as early as 100 ms after the stimulus was predicted to appear [27, 30]. Recently, we used multivariate methods to probe the representational content of such omission responses [57]. In this study, we presented subjects with auditory cues (high or low pitch) that predicted the orientation of an upcoming grating stimulus (clockwise or anticlockwise). In 25 % of trials, the grating stimulus was omitted. In these trials, only a prediction-inducing auditory tone was presented. Interestingly,

the pattern of activity evoked in V1 on these omission trials was similar to the pattern evoked by the predicted stimulus (e.g., a clockwise grating). In other words, neural activity in V1 evoked solely by predictions, in the absence of visual input, contained information about the grating orientation that was predicted.

Further evidence for representation-specific signals in early visual cortex in the absence of input comes from a study that presented subjects with naturalistic images of which one quadrant was occluded [58]. These authors used multivariate methods to show that the non-stimulated region of V1 (i.e., retinotopically corresponding to the occluded quadrant) contained information about the presented naturalistic image. Control analyses showed that this could not be explained by the spreading of activity (through lateral connections) from stimulated regions within V1. In a hierarchical inference framework, these results would be taken to reflect predictive feedback from a higher-order region containing a representation of the naturalistic scene as a whole (e.g., the whole car), informing lower-order areas which fine spatial features to expect (e.g., the angle of the tail light).

Single neuron recordings in monkeys have also provided empirical support for neuronal responses to absent but predicted input. Lee and Nguyen [59] presented monkeys with Kanizsa figures (Fig. 11.1a) in such a way that the illusory contours were positioned in the receptive fields of the V1 and V2 neurons they recorded from. Interestingly, both V1 and V2 neurons responded to these illusory contours (Fig. 11.3c). Moreover, V2 neurons responded consistently earlier in time, suggesting a feedback mechanism from V2 to V1. Similarly to the apparent motion example discussed earlier, this can be understood to result from inferences about the presence of a white square occluding the black circles within higher-order visual regions with receptive fields that encompass the whole figure. These inferences are subsequently sent as top-down predictions to those lower-order neurons that are expected to detect the (illusory) sides of the square [60]. It should be noted, however, that some previous studies have observed illusory contour responses in V2, but not in V1 [61]. The presence of predictive feedback to V1 may depend on such factors as stimulus size, attention, and experience with the stimulus.

It should be noted that while the above studies clearly demonstrate representation-specific activity in the absence of sensory input, they cannot strictly distinguish between prediction (P unit) and prediction error (PE unit) activity. Since PE reflects the mismatch between the prior expectation (a specific stimulus) and the input (an empty screen), unexpected omissions could conceivably cause activity in PE units. Evidence of true 'predictive' activity would require demonstrating representation-specific activity *prior* to sensory input. Such evidence is provided by paired-association studies in the anterior temporal cortex of macaques [28, 62, 63]. In these studies, monkeys were exposed to pairs of stimuli that were sequentially presented. Learning which stimuli form pairs allowed the monkeys to predict the identity of the second member of a pair upon presentation of the first member. Indeed, after learning, neurons that respond strongly to the second member of a pair already start firing upon presentation of the first member of the pair, i.e., as soon as the identity of the second member can be predicted (Fig. 11.3d). This predictive firing increases until the second stimulus appears, and is higher when monkeys correctly identify

the upcoming stimulus [62]. Furthermore, Erickson and Desimone [63] found that in sessions in which monkeys showed behavioural evidence of association learning, neural activity during the delay period between the two stimuli correlated less strongly with the response to the first stimulus, and more with the response to the second (upcoming) stimulus. In other words, delay activity became more predictive. Meyer and Olson [28] showed that when such a prediction is violated, that is, when the first stimulus is followed by one it has not previously been paired with, neural activity to the second stimulus is increased, suggesting a prediction error response. Similar findings of pair association in the medial temporal lobe have been reported in humans using fMRI [64].

11.3.3 Integration of Predictions and Inputs

In predictive coding theories, P and PE units do not function independently of each other. Rather, the error response in the PE units influences the hypotheses selected in the *P* units, and vice versa. Therefore, the final hypothesis the brain settles on (the posterior) reflects an integration of prior expectations and sensory input. In other words, if you expect *A*, and get bottom-up input *B*, your percept (posterior) should be somewhere in between. The relative weights of prior and input depend on their precisions: when the input is strong and unequivocal, the prior will have little effect, but if the input is ambiguous, the posterior will be largely determined by the prior [65, 66]. The integration of prior and input has been demonstrated convincingly in behaviour [10, 67, 68], yet neural evidence has been mostly lacking. The main question is whether there is already integration of bottom-up inputs and top-down expectations in sensory regions, as predicted by predictive coding theories, or whether integration takes place in downstream association areas that are proposed to be involved in perceptual decision-making, such as parietal and prefrontal cortex [69]. In line with the former, Nienborg and Cumming [70] have shown that sensory responses in macaque early visual cortex (V2) are dynamic, shifting from the representation of the bottom-up stimulus to the perceptual choice (posterior) within seconds. In a recent study in humans, Serences and Boynton [71] presented participants with ambiguous (0 % coherent) moving dot stimuli and forced them to report whether the dots moved toward the top right or bottom left of the screen. Multivariate analysis methods revealed that motion sensitive area MT + contained information about the (arbitrarily) chosen direction of motion. These studies suggest that early visual areas represent the posterior rather than just the bottom-up input. While the source of the arbitrary choice (i.e., the prior) is unknown and undetermined in the study by Serences and Boynton [71], future studies may test the integration of prior and bottom-up stimulus more directly by explicitly manipulating the prior and probing its effects on stimulus representations in visual cortex.

11.3.4 Separate Units Coding Predictions and Prediction Errors

Although many findings of prediction and prediction error effects in cortex have been reported, there is, somewhat surprisingly, a conspicuous lack of direct evidence for separate populations of neurons encoding predictions (P units) and errors (PE units) [72]. However, some conjectures can be made.

Miller and Desimone [73] recorded from single neurons in IT cortex while monkeys performed a delayed match-to-sample task. In this task, monkeys were presented with a sample stimulus, and had to respond when any of the following test stimuli matched the sample. Roughly half of the IT cells recorded showed differential responses to stimuli that matched, compared to stimuli that did not match the sample. Of these, 62 % were *suppressed* by test stimuli that matched the sample, while 35 % showed an *enhanced* response. These effects were present right from the onset of visual responses in IT, about 80–90 ms after stimulus presentation. Only 3 % of cells showed mixed effects, i.e., suppression by some stimuli and enhancement by others, leading the authors to argue that the two classes of cells appear to be distinct. The behaviour of these two classes of cells is reminiscent of PE (suppressed response to matches) and P (enhanced response to matches) units, respectively, though effects of stimulus predictability were not explicitly tested in this study. Woloszyn and Sheinberg [74] also provided evidence for two functionally distinct sub-populations in IT. They found that the maximum response and stimulus-selectivity of excitatory cells were increased for familiar compared to novel stimuli (potentially reflecting enhanced representation in P units), while inhibitory interneurons responded more strongly to *novel* stimuli than to familiar ones (potentially reflecting a larger PE response).

Arguments for a separate population of prediction error neurons have also been inspired by so-called extra-classical receptive field effects in early visual cortex [1]. Certain neurons fire less when a stimulus extends beyond their receptive field [75]. Furthermore, such suppressive surround effects are stronger when the surround is a (predictable) continuation of the centre stimulus, e.g., a continuous line segment or a grating with an iso-oriented surround, compared to when the surround is non-continuous (e.g., a cross-oriented grating) [76–78]. A predictive coding framework can readily explain such responses; a large, continuous stimulus is represented well by a *P* unit in a higher-order area (e.g., V2), which then sends a prediction to the relevant lower-order (e.g., V1) error neurons, suppressing their response [1]. Indeed, extra-classical receptive field effects have been shown to (partly) depend on feedback from higher-order areas [79, 80]. Hupé et al. [80] showed that feedback from area MT leads to surround suppression in V1, as well as increased responses to stimuli confined to the classical receptive field. In other words, when feedback can successfully predict the lower-order response its effect is inhibitory, but when it cannot it is excitatory.

One (somewhat counterintuitive) feature of the classical predictive coding scheme is that P units in one region send inhibitory feedback to the PE units one step down in the cortical hierarchy. However, in the cortex, interregional feedback connections are

predominantly excitatory [but see [19, 80, 81]. It is possible that feedback may indirectly target inhibitory interneurons, achieving a net inhibition, as has been observed in surround suppression [79, 80]. Furthermore, there are alternative implementations of predictive coding that do not rely on inhibitory intercortical feedback. In the work of Spratling [18, 83], for example, excitatory feedback is sent from P units in one region to P units in the region below it in the cortical hierarchy. In other words, feedback directly reinforces lower order hypotheses that are consistent with the higher order hypothesis. Here, error suppression is an intracortical phenomenon, consistent with intraregional 'back projections' (i.e., from infragranular to supragranular and granular layers) targeting predominantly inhibitory interneurons [84, 85].

In sum, while there is no direct unequivocal evidence for the existence of separate populations of P and PE units, there is suggestive evidence that different layers within each cortical column may implement these distinct computational roles.

11.4 Predictive Coding and Attention

Traditionally, theories of attention and predictive coding have been seen as diametrically opposed [72, 86]. While predictive coding posits that neural responses to expected stimuli should be suppressed, many studies have reported *increased* neural responses to stimuli appearing at expected locations [87, 88]. This increase in activity has been attributed to spatial attention. In fact, studies of visuospatial attention have traditionally used probabilistic cues that predict the upcoming target location as a means of manipulating attention [89, 90]. However, belying this apparent tension between theories of attention and prediction, attention fits quite naturally into a predictive coding framework that takes the relative precision of predictions and sensory input into account [91, 92]. In what follows, we will outline an account of attention in predictive coding, and review empirical evidence for this theory.

11.4.1 Attention and Precision

In the real world, the reliability of sensory signals is changeable: keeping track of a ball is a lot easier in the light of day than at dusk. Perceptual inference must take the precision (inverse variance) of sensory signals (i.e., prediction errors) into account [2, 91, 93]. It is imperative to know whether sensory signals fail to match our prior expectations because they contain information that disproves our current model of the world (e.g., we see and hear a giant luminescent hound), or because the sensory signals are simply too noisy (we hear a dog howl but see only mist). While the former should lead us to update our beliefs (a demon hound!), the latter should not. Specifically, PEs should be weighted by their precision (i.e., reliability), leading to less weight being attributed to less reliable sensory information. In terms of hierarchical inference, perception should be dominated by sensory signals when

their precision is high, and by top-down expectations when their precision is low, e.g., when sensory signals are ambiguous [65, 66].

The precision of sensory signals has long been acknowledged to be a matter of importance for models of perceptual inference [2, 93], and recent predictive coding models incorporate it explicitly [91, 92]. In these models, the brain estimates not only PEs themselves, but also their precision. PEs are subsequently weighted by their precision through modulation of synaptic gain of the PE units. One hypothesis suggests that attention is the process whereby the brain optimises precision estimates [91, 92]. By increasing the precision of specific PEs, attention increases the weight these errors carry in perceptual inference. Mechanistically, this is equivalent to proposals of attention increasing synaptic gain (precision) of specific sensory neurons (PE units) [94–96]. Note that in predictive coding models, sensory signals and PE are equivalent, since these errors are the only sensory information that is yet to be explained. Therefore, while casting attention as modulating the precision of PEs may at first glance seem a radically different view of its function, it is in fact fully consistent with contemporary theories of attention. Indeed, in this account, spatial attention is proposed to increase the precision of information coming from a certain region of visual space, similar to the effect of pointing a flashlight in that direction, making spotlight metaphors of attention an intuitive way to think about its functional role [92, 97]. Furthermore, biased competition, a highly influential theory of attention [94], can be shown to emerge from a predictive coding framework in which attention is cast as optimising precision [92].

Crucially, prediction and precision (attention) are not independent. Instead, precision depends on the expected states of the world [92]: expectation of a stimulus on the left side of the visual field leads to expectation of high precision sensory signals at that location. Spatial attention might enhance these sensory signals further by boosting the precision (synaptic gain) at that location. This suggests that attention can be seen as a selective enhancement of sensory data that have high precision (high signal-to-noise ratio) in relation to the brain's current predictions [98]. Mechanistically, this means that attention does not simply boost the synaptic gain of PE units indiscriminately. Rather, it boosts the gain of PE units receiving input from P units (in the same hierarchical level and the level above) that are currently active. Therefore, attending to a region of space where a stimulus is not expected (in this example, to the right) would be relatively ineffective. Put simply, there should be no strong expectation of a high precision sensory signal when a stimulus is not expected to appear.

In sum, recent predictive coding theories propose that prediction is concerned with *what* is being represented, while attention is the process of optimising the *precision* of representations [91]. We will now turn to a discussion of empirical evidence for this proposal.

11.4.2 Empirical Evidence

The account outlined above may reconcile some seemingly contradictory findings in the literature. Generally speaking, it seems that expectation is associated with reduced sensory signals when stimuli are task irrelevant (unattended), but enhanced responses when stimuli are relevant (attended) [99]. For instance, two recent studies found opposite effects of predictive motion trajectories on stimulus processing. Doherty et al. [87] had participants view a red ball moving across the screen in either a regular (predictable) or irregular (unpredictable) trajectory. At some point the ball disappeared behind an occluder, and when it reappeared participants were required to detect a black dot on the surface of the ball as soon as possible. The different types of trajectories meant that the reappearance of the dot could be either predictable (in space and time) or unpredictable. Using EEG, these authors found that predictability enhanced the neural response in early sensory regions. In contrast, Alink et al. [37] found a *reduced* neural response in V1 in response to a stimulus that was congruent with a predictable trajectory (see § 11.3.1 for a more detailed discussion of this study), compared to a stimulus that was incongruent with the trajectory. In their design, subjects did not perform a task that involved the stimuli, instead stimulus presentations were fully irrelevant.

These studies provide a suggestion for a potential interaction between attention and prediction [for a review, see 98]. However, there are also notable differences between these two studies in terms of experimental paradigm (e.g., real vs. apparent motion) and methodology (EEG vs. fMRI). A direct study of the (potential) interaction of prediction and attention has been lacking.

In a recent fMRI study, we [56] manipulated both visuospatial attention (which side of the screen is task-relevant) and visuospatial prediction (on which side of the screen the stimulus is likely to appear). Spatial attention was manipulated on a trial-by-trial basis, by means of a cue that pointed either to the left or to the right. Unlike in typical Posner cueing tasks, in which attention is biased towards one visual hemifield by increasing the probability of the target appearing on that side, in this experiment, the attention manipulation was not probabilistic. The key difference is that in the former, both visual hemifields are potentially task-relevant, with one more likely to be so than the other, while in our study only one hemifield was task-relevant in a particular trial. If a subsequent grating stimulus appeared in the hemifield indicated by the cue, subjects were asked to do an orientation identification task on the grating stimulus. If instead the grating appeared on the other side of the screen, subjects could simply ignore it. Prediction was manipulated in mini-blocks: in each block of eight trials, subjects were told that stimuli were (a) 75 % likely to appear on the left, (b) 75 % likely to appear on the right, (c) 50 % likely to appear on either side (neutral cue). Thereby, attention and prediction were orthogonally manipulated. We reasoned that there were two likely explanations for the seemingly contradictory effects of expectation reported in the literature. One possible explanation is that attention and prediction have opposing main effects, enhancing and suppressing sensory signals, respectively. If the enhancing effect of attention outweighs the suppressive

effect of prediction, this would explain the seemingly enhancing effect of expectation in attentional cueing experiments. Alternatively, attention and prediction might operate synergistically to optimise perceptual inference, with attention boosting the precision (synaptic gain) of PE units that are expected to receive input based on current predictions. Hereby, if attention and prediction are congruent (i.e., a stimulus is expected in the task-relevant hemifield), attention would boost (the precision of) the expected sensory signal. However, if attention and prediction are incongruent (a stimulus is expected in the task-irrelevant hemifield), attention would be relatively ineffective in boosting the sensory signal; there should be no strong expectation that an unpredicted signal is going to be precise. This account would therefore predict an interactive effect of attention and prediction on sensory signals (see Fig. 11.4a).

The data provided support for the latter hypothesis (Fig. 11.4b). When stimuli were task-irrelevant (unattended), predicted stimuli evoked a reduced neural response in V1 compared to unpredicted stimuli. However, when stimuli were task-relevant (attended), this pattern reversed: here, predicted stimuli evoked an enhanced neural response. This interaction is in line with predictive coding models casting attention as optimising precision estimates during perceptual inference [91, 92]. Furthermore, when a stimulus was predicted in the task-relevant hemifield (i.e., there was a strong and precise prediction), we observed an increased response in V1 when this stimulus was omitted (Fig. 11.4c). As discussed in § 11.3.2, this might reflect either the prediction itself, or a prediction error response. In either case, this effect is in line with predictive coding, but hard to reconcile with bottom-up attention accounts of 'stimulus surprise', since there was no stimulus to grab attention in these trials (or, rather, a stimulus appeared in the opposite hemifield).

Further support for attention increasing the precision of sensory signals (prediction errors) comes from studies showing that fluctuations in the amplitude of the neural signal in visual areas covary with detection task performance, both pre- [100] and post-stimulus [101]. In other words, activity in these regions is higher when people correctly detect or reject the presence of a stimulus than when they incorrectly report or miss it (although amplitude has also been seen to covary with subjective perception rather than performance accuracy; [102]).

Boosting specific sensory signals results in a gain in signal-to-noise ratio (precision) for those signals. Such a gain could also be achieved by reducing the neural noise in sensory areas. In fact, single cell recordings in macaques have revealed a decrease in neural noise correlations as the result of attention [103, 104]. Furthermore, a recent behavioural study that applied sophisticated signal detection theory analyses showed that whereas prediction increases the baseline activity of stimulus-specific units (*P* units?), attention suppresses internal noise during signal processing [55]. In order to optimally boost the signal-to-noise ratio of selected sensory signals, attention may both increase the amplitude of specific prediction errors, as well as suppress noise fluctuations arising from non-selected sources.

In sum, the empirical findings discussed above are in line with predictive coding models incorporating precision estimates, and casting attention as the process of optimising those estimates. This framework resolves the apparent tension between

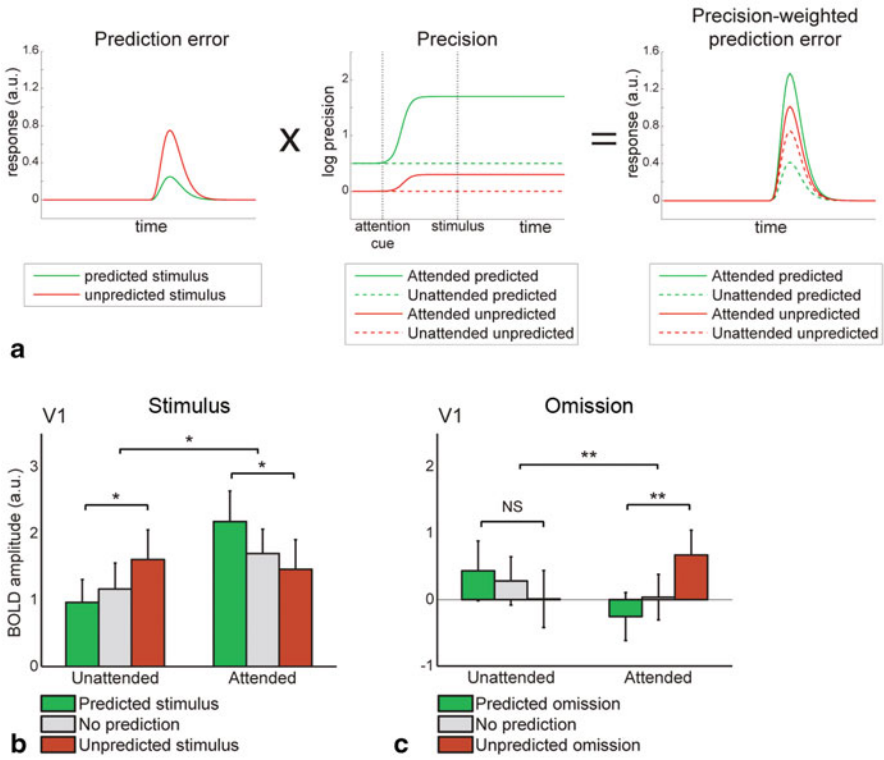


Fig. 11.4 Attention and precision. **a** *Left* panel: The hypothetical prediction error responses to physically identical stimuli, preceded by either a valid (*green*) or invalid (*red*) prediction cue. *Middle* panel: In recent predictive coding models, attention increases the precision (synaptic gain) of prediction errors. This enhancement of precision by attention occurs in relation to current predictions, reflected here by the fact that attention hardly increases precision when no stimulus is predicted to occur. The order of magnitude of the precision values displayed here was based on figures in Feldman and Friston [91], the exact values were chosen arbitrarily, and their evolution over time was simplified. *Right* panel: Prediction errors are weighted by their precision, calculated here as a simple multiplication of prediction error (*left* panel) and precision (*middle* panel). The fact that attention enhances precision *in relation to current predictions* leads to an interactive effect of prediction and attention on the amplitude of the prediction error response. **b** When stimuli are unattended (task irrelevant), predicted stimuli evoke a reduced response in V1 compared to unpredicted stimuli. On the other hand, when stimuli are attended, predicted stimuli evoked a larger response than unpredicted stimuli. This is exactly the interaction between attention and prediction that is hypothesised by recent predictive coding models, see **a**. **c** In visual cortex corresponding to the visual field where no stimulus appeared, i.e. ipsilateral to the stimulus, unpredicted omission of a stimulus in the attended visual field evoked a larger response in V1 than predicted omission of a stimulus. Figures reprinted from [56], with permission from the authors

theories of attention and predictive coding [18], and explains the seemingly contradictory findings in the literature regarding the effects of expectation on neural activity [72, 99].

11.5 Concluding Remarks

In this chapter, we reviewed recent theoretical and empirical advances in the field of predictive coding. Although we have shown that predictive coding makes predictions that can be tested by cognitive neuroscience, and which have been supported by the extant data reasonably well, we would like to stress that more evidence is needed. Particularly, direct evidence for separate sub-populations of P and PE units is lacking. Since these two sub-populations are proposed to co-exist in every (sensory) cortical region, high-resolution methods are required to simultaneously sample neural activity from multiple sites at high spatial resolution. Specifically, given the speculations on different laminar distributions of P and PE units (see § 11.2.3), multicontact laminar electrodes [e.g., [104] or high-resolution laminar fMRI [106] could provide such evidence. So far, there have been no studies using these methods that have focused on the effects of prediction on sensory responses. Under a predictive coding framework, it may be hypothesised that, preceding stimulus onset, expectation would lead to activity in cortical layers containing P units, while after stimulus onset, activity would scale with prediction error in layers dominated by PE units. At the level of single neurons, P and PE units are predicted to be reciprocally connected, with the strength of the excitatory forward connection between individual PE and P units being equal to the strength of the inhibitory backward connection between these same neurons [1, 18]. In V1, it seems conceivable that simple and complex cells [75] could be interpreted as PE and P units, respectively. If this is true, complex cells are expected to inhibit the simple cells that provide them with input. This is a testable hypothesis. In the coming years, studies testing these hypotheses will provide us with much needed answers regarding the possible implementation of predictive coding in the human cortex.

So far, studies of predictive coding have mostly focused on the effects of prediction on the processing of sensory inputs. However, the model also provides a natural explanation for top-down activations of representations in the absence of sensory input. For example, processes like working memory and mental imagery (and even dreaming) might reflect activating part of one's internal model of the (visual) world [2]. These activations would come about through a different flow of information, compared to stimulus-driven activations: whereas the latter would arrive as input into layer 4 and sent onwards to supra- and infragranular layers, the former would bypass layer 4 and directly target agranular layers [107]. Crucially, these opposite flows of information could result in identical representations being activated (in agranular layers). Indeed, recent neuroimaging studies suggest that working memory [108], mental imagery [109, 110], and even dreaming [111] share sensory representations with perception. Such offline activations of the brain's internal model could serve several purposes, such as simulating scenario's not (yet) encountered but consistent with the model (e.g., mental rehearsal), and consolidating synaptic connections between representations within and across different levels of the cortical hierarchy. Speculatively, dreams may subserve both these functions.

Future work might also focus on the link between the neuronal substrate of predictive coding and subjective perception. It seems natural to assume that the contents of perception reflect the current hypothesis represented in the P units across the cortical hierarchy. Might the intensity (e.g., brightness, contrast, duration) of the percept then scale with the prediction error [25]? This account would predict that valid expectations lead to percepts that are 'sharper' (improved representation in P units) but less intense (reduced PE), in line with neural effects of expectation in sensory cortex [14]. Indeed, oddball stimuli (that is, unexpected deviants) are perceived as being of longer duration than standards [25, 112, 113]. Also, this account can explain the fact that representations activated by top-down processes such as working memory and imagery are not perceived as vividly as those activated during normal perception; presumably the former bypass PE units and directly activate P units. Furthermore, since attention is proposed to boost the synaptic gain of PE units (see § 4), the increase in perceived contrast observed as a result of attention fits naturally in this framework [114]. Finally, psychosis has been conjectured to involve aberrantly increased prediction errors, and indeed patients report more intense percepts (brighter colours, louder sounds) in early stages of the disease [115]. In fact, it is interesting to note that many positive and negative symptoms of syndromes like schizophrenia [116–118], psychosis [115], and autism [119, 120] can be explained in terms of specific failures of predictive coding mechanisms.

In sum, predictive coding provides a good explanation for many phenomena observed in perception, and generates testable predictions. In this chapter, we have reviewed existing empirical evidence for some of these predictions, as well as outlined possible future directions for further empirical testing and for broadening the perspective of the role predictive coding may play in cognition.

Exercises

1. Does the suppressed response in V1 to predicted stimuli [37, 44, 53] mean that there is less stimulus information in V1 for such stimuli? Why/Why not?
2. In what respect are the neural effects of prediction and attention opposite to each other, and in what respect are they similar?
3. Come up with an experiment that could potentially falsify predictive coding.
4. Given that top-down predictions silence prediction errors in lower-order regions; does predictive coding require inhibitory feedback between cortical regions? Read the paper by Spratling [18] and prepare a 15 min presentation on the differences between the physiological implementation implied by "classical" predictive coding models [1, 3] and that implied by Spratling's PC/BC model.
5. During hallucinations, schizophrenic and psychotic patients perceive things that are not actually there. Autistic patients, on the other hand, sometimes seem to perceive things more precisely or truthfully than non-autistics. How could these symptoms be understood in terms of predictive coding?
6. Read the Corlett et al. [115] paper, and prepare a 30 min presentation on the relationship between predictive coding and psychosis.

Further Reading

1. For an introduction to the principles of predictive coding and a global perspective of its implications for cognition and action, see the recent review by Andy Clark [121].
2. Friston [3] offers a comprehensive and mathematical description of predictive coding, including a proposal for its neuronal implementation.
3. Summerfield and Egner [72] review the commonalities and differences between theories of predictive coding and attention.
4. In a clearly written and succinct paper, Spratling [18] presents the computational principles of predictive coding and biased competition, and shows that—under certain assumptions—they are equivalent.
5. Lee and Mumford [9] offer a slightly different take on hierarchical inference during perception that shares many of the principles of predictive coding.

References

1. Rao RPN, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci 2(1):79–87
2. Mumford D (1992) On the computational architecture of the neocortex. II. The role of cortico-cortical loops. Biol Cybern 66(3):241–251
3. Friston KJ (2005) A theory of cortical responses. Philos Trans R Soc Lond B Biol Sci 360(1456):815–836
4. Helmholtz H (1867) Handbuch der physiologischen optik. L. Voss, Leipzig
5. Weiss Y, Simoncelli EP, Adelson EH (2002) Motion illusions as optimal percepts. Nat Neurosci 5(6):598–604
6. Gregory RL (1997) Knowledge in perception and illusion. Phil Trans R Soc Lond B 352:1121–1128
7. Knill DC, Richards W (Eds) (1996) Perception as Bayesian inference. Cambridge University Press: Cambridge
8. Yuille A, Kersten D (2006) Vision as Bayesian inference: analysis by synthesis? Trends Cogn Sci 10(7):301–308
9. Lee TS, Mumford D (2003) Hierarchical Bayesian inference in the visual cortex. J Opt Soc Am A Opt Image Sci Vis 20(7):1434–1448
10. Kersten D, Mamassian P, Yuille A (2004) Object perception as Bayesian inference. Annu Rev Psychol 55:271–304
11. Den Ouden HEM, Kok P, De Lange FP (2012) How prediction errors shape perception, attention, and motivation. Front Psychol 3:548
12. Spratling MW (2012) Predictive coding as a model of the V1 saliency map hypothesis. Neural Networks 26:7–28
13. Pearce JM, Hall G (1980) A model for Pavlovian learning: variations on the effectiveness of conditioned but not of unconditioned stimuli. Psychol Rev 87(6):532–552
14. Kok P, Jehee JFM, De Lange FP (2012) Less is more: expectation sharpens representations in the primary visual cortex. Neuron 75:265–270
15. Bar M (2004) Visual objects in context. Nat Rev Neurosci 5(8):617–629
16. Ahissar M, Hochstein S (2004) The reverse hierarchy theory of visual perceptual learning. Trends Cogn Sci 8(10):457–464
17. Sugase Y et al (1999) Global and fine information coded by single neurons in the temporal visual cortex. Nature 400:869–873

18. Spratling MW (2008) Reconciling predictive coding and biased competition models of cortical function. Front Comput Neurosci 2:4
19. Bastos AM et al (2012) Canonical microcircuits for predictive coding. Neuron 76:695–711
20. Felleman DJ, Van Essen DC (1991) Distributed hierarchical processing in the primate cerebral cortex. Cereb Cortex 1(1):1–47
21. Maunsell JHR, Van Essen DC (1983) The connections of the middle temporal visual area (MT) and their relationship to a cortical hierarchy in the macaque monkey. J Neurosci 3(12):2563–2586
22. Rockland K, Pandya D (1979) Laminar origins and terminations of cortical connections of the occipital lobe in the rhesus monkey. Brain Res 179:3–20
23. Rao RPN, Sejnowski, TJ (2002) Predictive coding, cortical feedback, and spike-timing dependent plasticity. In: Rao RPN, Olshausen BA, Lewicki MS (eds) Probabilistic models of the brain: perception and neural function. MIT Press, Cambridge, pp 297–315
24. Grill-Spector K, Henson R, Martin A (2006) Repetition and the brain: neural models of stimulus-specific effects. Trends Cogn Sci 10(1):14–23
25. Pariyadath V, Eagleman D (2007) The effect of predictability on subjective duration. PLoS ONE 2(11):e1264
26. Summerfield C et al (2008) Neural repetition suppression reflects fulfilled perceptual expectations. Nat Neurosci 11(9):1004–1006
27. Todorovic A et al (2011) Prior expectation mediates neural adaptation to repeated sounds in the auditory cortex: an MEG study. J Neurosci 31(25):9118–9123
28. Meyer T, Olson CR (2011) Statistical learning of visual transitions in monkey inferotemporal cortex. Proc Natl Acad Sci U S A 108(48):19401–19406
29. Todorovic A, De Lange FP (2012) Repetition suppression and expectation suppression are dissociable in time in early auditory evoked fields. J Neurosci 32(39):13389–13395
30. Wacongne C et al (2011) Evidence for a hierarchy of predictions and prediction errors in human cortex. Proc Natl Acad Sci U S A 108(51):20754–20759
31. Wacongne C, Changeux JP, Dehaene S (2012) A neuronal model of predictive coding accounting for the mismatch negativity. J Neurosci 32(11):3665–3678
32. Schröger E, Wolff C (1996) Mismatch response of the human brain to changes in sound location. Neuroreport 7:3005–3008
33. Muckli L et al (2005) Primary visual cortex activity along the apparent-motion trace reflects illusory perception. PLoS Biol 3(8):e265
34. Sterzer P, Haynes JD, Rees G (2006) Primary visual cortex activation on the path of apparent motion is mediated by feedback from hMT + /V5. Neuroimage 32:1308–1316
35. Wibral M et al (2009) The timing of feedback to early visual cortex in the perception of long-range apparent motion. Cereb Cortex 19:1567–1582
36. Ahmed B et al (2008) Cortical dynamics subserving visual apparent motion. Cereb Cortex 18:2796–2810
37. Alink A et al (2010) Stimulus predictability reduces responses in primary visual cortex. J Neurosci 30(8):2960–2966
38. Chandrasekaran C et al (2009) The natural statistics of audiovisual speech. PLoS Comput Biol 5(7):e1000436
39. Sumby WH, Pollack I (1954) Visual contribution to speech intelligibility in noise. J Acoust Soc Am 26(2):212–215
40. Van Wassenhove V, Grant KW, Poeppel D (2005) Visual speech speeds up the neural processing of auditory speech. Proc Natl Acad Sci U S A 102(4):1181–1186
41. Arnal LH et al (2009) Dual neural routing of visual facilitation in speech processing. J Neurosci 29(43):13445–13453
42. McGurk H, MacDonald J (1976) Hearing lips and seeing voices. Nature 264:746–748
43. Arnal LH, Wyart V, Giraud A (2011) Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. Nat Neurosci 14(6):797–801
44. Den Ouden HEM et al (2009) A dual role for prediction error in associative learning. Cereb Cortex 19:1175–1185

45. Den Ouden HEM et al (2010) Striatal prediction error modulates cortical coupling. J Neurosci 30(9):3210–3219
46. Egner T, Monti JM, Summerfield C (2010) Expectation and surprise determine neural population responses in the ventral visual stream. J Neurosci 30(49):16601–16608
47. Bell CC (2001) Memory-based expectations in electrosensory systems. Curr Opin Neurobiol 11:481–487
48. Blakemore SJ, Wolpert DM, Frith CD (1998) Central cancellation of self-produced tickle sensation. Nat Neurosci 1(7):635–640
49. Houde JF et al (2002) Modulation of the auditory cortex during speech: an MEG study. J Cognitive Neurosci 14(8):1125–1138
50. Martikainen MH, Kaneko K, Hari R (2005) Suppressed responses to self-triggered sounds in the human auditory cortex. Cereb Cortex 15:299–302
51. Shergill SS et al (2013) Modulation of somatosensory processing by action. Neuroimage 70:356–362
52. Crapse TB, Sommer MA (2008) Corollary discharge across the animal kingdom. Nat Rev Neurosci 9:587–600
53. Murray SO et al (2002) Shape perception reduces activity in human primary visual cortex. Proc Natl Acad Sci U S A 99(23):15164–15169
54. Fang F, Kersten D, Murray SO (2008) Perceptual grouping and inverse fMRI activity patterns in human visual cortex. J Vision 8(7):2–9
55. Wyart V, Nobre AC, Summerfield C (2012) Dissociable prior influences of signal probability and relevance on visual contrast sensitivity. Proc Natl Acad Sci U S A 109(9):3593–3598
56. Kok P et al (2012) Attention reverses the effect of prediction silencing sensory signals. Cereb Cortex 22(9):2197–2206
57. Kok P, Failing FM, De Lange FP (2014) Prior expectations evoke stimulus templates in the primary visual cortex. J Cogn Neurosci 26(7):1546–1554
58. Smith FW, Muckli L (2010) Nonstimulated early visual areas carry information about surrounding context. Proc Natl Acad Sci U S A 107(46):20099–20103
59. Lee TS, Nguyen M (2001) Dynamics of subjective contour formation in the early visual cortex. Proc Natl Acad Sci U S A 98(4):1907–1911
60. Kok P, De Lange FP (2014) Shape perception simultaneously up- and downregulates neural activity in the primary visual cortex. Curr Biol 24:1531–1535
61. Von Der Heydt R, Peterhans E, Baumgartner G (1984) Illusory contours and cortical neuron responses. Science 224(4654):1260–1262
62. Sakai K, Miyashita Y (1991) Neural organization for the long-term memory of paired associates. Nature 354:152–155
63. Erickson CA, Desimone R (1999) Responses of macaque perirhinal neurons during and after visual stimulus association learning. J Neurosci 19(23):10404–10416
64. Schapiro AC, Kustner LV, Turk-Browne NB (2012) Shaping of object representations in the human medial temporal lobe based on temporal regularities. Curr Biol 22(17):1622–1627
65. Sterzer P, Frith C, Petrovic P (2008) Believing is seeing: expectations alter visual awareness. Curr Biol 18(16):R697–R698
66. Denison RN, Piazza EA, Silver MA (2011) Predictive context influences perceptual selection during binocular rivarly. Front Human Neurosci 5:166
67. Chalk M, Seitz AR, Seriès P (2010) Rapidly learned stimulus expectations alter perception of motion. J Vision 10(8):1–18
68. Sotiropoulos G, Seitz AR, Seriès P (2011) Changing expectations about speeds alters perceived motion direction. Curr Biol 21(21):R883–R884
69. Gold JI, Shadlen MN (2007) The neural basis of decision making. Annu Rev Neurosci 30:535–74
70. Nienborg H, Cumming BG (2009) Decision-related activity in sensory neurons reflects more than a neuron's causal effect. Nature 459:89–92
71. Serences JT, Boynton GM (2007) The representation of behavioral choice for motion in human visual cortex. J Neurosci 27(47):12893–12899

72. Summerfield C, Egner T (2009) Expectation (and attention) in visual cognition. Trends Cogn Sci 13(9):403–409
73. Miller EK, Desimone R (1994) Parallel neuronal mechanisms for short-term memory. Science 263(5146):520–522
74. Woloszyn L, Sheinberg DL (2012) Effects of long-term visual experience on responses of distinct classes of single units in inferior temporal cortex. Neuron 74:193–205
75. Hubel DH, Wiesel TN (1968) Receptive fields and functional architecture of monkey striate cortex. J Physiol 195(1):215–243
76. Jones HE, Wang W, Sillito AM (2002) Spatial organization and magnitude of orientation contrast interactions in primate V1. J Neurophysiol 88:2796–2808
77. Knierim JJ, Van Essen DC (1992) Neuronal responses to static texture patterns in area V1 of the alert macaque monkey. J Neurophysiol 67:961–980
78. Sillito AM et al (1995) Visual cortical mechanisms detecting focal orientation discontinuities. Nature 378:492–496
79. Angelucci A, Bullier J (2003) Reaching beyond the classical receptive field of V1 neurons: horizontal or feedback axons? J Physiology-Paris 97:141–154
80. Hupé JM et al (1998) Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons. Nature 394:784–787
81. Johnson RR, Burkhalter A (1996) Microcircuitry of forward and feedback connections within rat visual cortex. J Comp Neurol 368:383–398
82. Johnson RR, Burkhalter A (1997) A polysynaptic feedback circuit in rat visual cortex. J Neurosci 17(18):7129–7140
83. Spratling MW (2008) Predictive coding as a model of biased competition in visual attention. Vision Res 48:1391–1408
84. Thomson AM, Bannister AP (2003) Interlaminar connections in the neocortex. Cereb Cortex 13:5–14
85. Olsen SR et al (2012) Gain control by layer six in cortical circuits of vision. Nature 483:47–52
86. Koch C, Poggio T (1999) Prediction the visual world: silence is golden. Nat Neurosci 2(1):9–10
87. Doherty JR et al (2005) Synergistic effect of combined temporal and spatial expectations on visual attention. J Neurosci 25(36):8259–8266
88. Chaumon M, Drouet V, Tallon-Baudry C (2008) Unconscious associative memory affects visual processing before 100 ms. J Vis 8(3):1–10
89. Posner MI (1980) Orienting of attention. Q J Exp Psychol 32(1):3–25
90. Mangun GR, Hillyard SA (1991) Modulations of sensory-evoked brain potentials indicate changes in perceptual processing during visual-spatial priming. J Exp Psychol 17(4):1057–1074
91. Friston KJ (2009) The free-energy principle: a rough guide to the brain? Trends Cogn Sci 13(7):293–301
92. Feldman H, Friston KJ (2010) Attention, uncertainty, and free-energy. Front Human Neurosci 4:215
93. Rao RPN, Ballard DH (1997) Dynamic model of visual recognition predicts neural response properties in the visual cortex. Neural Comput 9:721–763
94. Desimone R, Duncan J (1995) Neural mechanisms of selective visual attention. Annu Rev Neurosc 18:193–222
95. Maunsell JHR, Treue S (2006) Feature-based attention in visual cortex. Trends Neurosci 29(6):317–322
96. Reynolds JH, Heeger DJ (2009) The normalization model of attention. Neuron 61(2):168–185
97. Posner MI, Snyder CR, Davidson BJ (1980) Attention and the detection of signals. J Exp Psychol 109(2):160–174
98. Hohwy J (2012) Attention and conscious perception in the hypothesis testing brain. Front Psych 3:96
99. Rauss K, Schwartz S, Pourtois G (2011) Top-down effects on early visual processing in humans: a predictive coding framework. Neurosci Biobehav Rev 35:1237–1253

100. Hesselmann G et al (2010) Predictive coding or evidence accumulation? False inference and neuronal fluctuations. PLoS ONE 5:e9926
101. Ress D, Backus BT, Heeger DJ (2000) Activity in primary visual cortex predicts performance in a visual detection task. Nat Neurosci 3(9):940–945
102. Ress D, Heeger DJ (2003) Neuronal correlates of perception in early visual cortex. Nat Neurosci 6(4):414–420
103. Mitchell JF, Sundberg KA, Reynolds JH (2009) Spatial attention decorrelates intrinsic activity fluctuations in macaque area V4. Neuron 63:879–888
104. Cohen MR, Maunsell JHR (2009) Attention improves performance primarily by reducing interneuronal correlations. Nat Neurosci 12(12):1594–1601
105. Lakatos P et al (2008) Entrainment of neuronal oscillations as a mechanism of attentional selection. Science 320(5872):110–113
106. Koopmans PJ, Barth M, Norris DG (2010) Layer-Specific BOLD Activation in Human V1. Hum Brain Mapp 31:1297–1304
107. Takeuchi D et al (2011) Reversal of interlaminar signal between sensory and memory processing in monkey temporal cortex. Science 331:1443–1447
108. Harrison SA, Tong F (2009) Decoding reveals the contents of visual working memory in early visual areas. Nature 458:632–635
109. Lee SH, Kravitz DJ, Baker CI (2012) Disentangling visual imagery and perception of real-world objects. Neuroimage 59:4064–4073
110. Albers AM et al (2013) Shared representations for working memory and mental imagery in early visual cortex. Curr Biol 23:1427–1431
111. Horikawa T et al (2013) Neural decoding of visual imagery during sleep. Science 340:639–642
112. Schindel R, Rowlands J, Arnold DH (2011) The oddball effect: perceived duration and predictive coding. J Vision 11(2):17
113. Tse PU et al (2004) Attention and the subjective expansion of time. Percept Psychophys 66(7):1171–1189
114. Carrasco M, Ling S, Read S (2004) Attention alters appearance. Nat Neurosci 7(3):308–313
115. Corlett PR et al (2011) Glutamatergic model psychoses: prediction error, learning, and inference. Neuropsychopharmacol 36:294–315
116. Fletcher PC, Frith CD (2009) Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. Nat Rev Neurosci 10(1):48–58
117. Eagleman DM, Pariyadath V (2009) Is subjective duration a signature of coding efficiency? Phil Trans R Soc B 364:1841–1851
118. Blakemore SJ et al (2000) The perception of self-produced sensory stimuli in patients with auditory hallucinations and passivity experiences: evidence for a breakdown in self-monitoring. Psychol Med 30:1131–1139
119. Pellicano E, Burr D (2012) When the world becomes 'too real': a Bayesian explanation of autistic perception. Trends Cogn Sci 16(10):504–510
120. Van de Cruys S et al (2013) Weak priors versus overfitting of predictions in autism: reply to Pellicano and Burr (TICS, 2012). Iperception 4:95–97
121. Clark A (2013) Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behav Brain Sci 36:181–253

Chapter 12
Using Human Neuroimaging to Examine Top-down Modulation of Visual Perception

Thomas C. Sprague and John T. Serences

Abstract Both univariate and multivariate analysis methods largely have focused on characterizing how measurements from neural firing rates, EEG electrodes, or fMRI voxels change as a function of stimulus parameters or task demands –they focus on characterizing changes in neural signals. However, in cognitive neuroscience we are often interested in how these changes in neural signals collectively modify representations of information. We compare methods whereby activation patterns across entire brain regions can be used to reconstruct representations of information to more traditional univariate and multivariate analysis approaches. We highlight findings using these methods, focusing on how a representation-based analysis approach yields novel insights into how information is encoded, maintained and manipulated under various task demands.

12.1 Introduction: Observation-Based vs Representation-Based Approaches to Analyzing Neural Signals

There are two ways of inferring the relationship between changes in stimulus parameters and neural activity: you can either ask how a stimulus affects the response of a neuron, or you can ask how the response of a neuron influences the representation of a stimulus. If there is a one-to-one mapping between neural responses and stimulus parameters such that the spike rate of a particular neuron, and only that particular neuron, maps to, and only to, one particular stimulus (single unit doctrine, [1]), then these two approaches are essentially identical. However, if ensembles of neurons interact in complex ways and this one-to-one mapping breaks down, then these two approaches can yield dramatically different insights into the relationship between

T. C. Sprague (✉)
Neurosciences Graduate Program, University of California,
San Diego, La Jolla, CA 92093-0109, USA
e-mail: tsprague@ucsd.edu

J. T. Serences
Department of Psychology, University of California, San Diego, La Jolla, CA 92093-0109, USA
e-mail: jserences@ucsd.edu

B. U. Forstmann, E.-J. Wagenmakers (eds.), *An Introduction to Model-Based Cognitive Neuroscience,* DOI 10.1007/978-1-4939-2236-9_12

neural activity and the manner in which that neural activity represents the external stimuli that support our actions, thoughts, and behaviors.

For much of the twentieth century studies adopted the first approach—what we'll refer to here as an *observation-based* approach—that focused on assessing how experimental manipulations changed brain activity. For example, scientists measured spike rates of single neurons, electroencephalogram (EEG) waveforms, positron emission tomography (PET) signals, and blood flow in humans (blood oxygenation level dependent functional magnetic resonance imaging, or BOLD fMRI) while subjects viewed different stimuli or performed different tasks. This general endeavor has been and continues to be tremendously successful and has resulted in the identification of fundamental properties of neural systems (e.g., neurons in the early visual system respond to basic visual features in the environment, like edges, [2]), and the ability to measure these signals in humans has raised exciting possibilities such as the existence of specialized "modules" that process different categories of stimulus attributes (such as "face" and "place" areas, [3–5]). Thus, the major advantage of this observation-based approach is its ability to describe *what* alters the response of neural "units" (cells, voxels, electrodes, etc.), *where* these units are located in the nervous system, and *how* activity in these units changes with stimulus properties and task demands. Furthermore, this approach excels at identifying the progression of neural computations across cortical systems in response to the same set of complex stimulus features or task demands (for example, documenting the cascade of visual responses in ventral visual cortex for facial features, facial identity, and facial viewpoint, [6, 7]).

In contrast, recent work has started to adopt the second approach of asking how neural responses—or combinations of neural responses—give rise to changes in the way in which a particular stimulus is represented and how that representation changes as a function of experimental manipulations. We refer to this as the *representation-based* approach because the goal is to understand how stimulus-specific information is *represented* by populations of neurons, rather than to simply document how experimental manipulations change the observed neural responses *per se*. This is a fundamentally different approach to understanding functions of the brain. Instead of amassing observation-based data that catalogs how a particular cell responds to a particular stimulus or task manipulation, the representation-based approach explicitly models the inverse relationship between neural activity and stimulus features in order to reconstruct an 'image' of the stimulus based on patterns of neural responses. For example, an experimenter might want to know how a change in the firing rate of a visual neuron influences how well a relevant stimulus feature is represented. Is the feature represented as being higher contrast, as having a higher spatial frequency, as having a more saturated color? What if different measured units (voxels, neurons, etc) exhibit different types of modulation across the same experimental manipulation? Do all these changes average out, or is the individual contribution of each unit to the representation meaningful? Using an observation-based approach, you can't easily address how such changes are related to the representation of a stimulus: a change in firing rate might lead to a representation that is higher in contrast or that is more saturated, but it might not, and indeed researchers have debated these types of interpretational questions for decades.

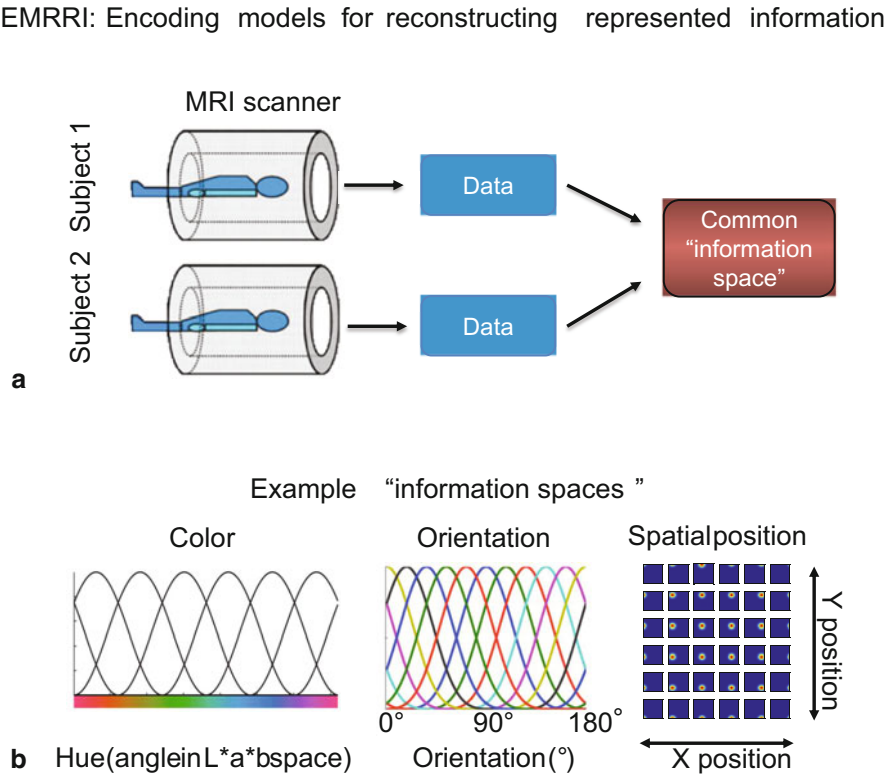


Fig. 12.1 Representation-based approach to cognitive neuroscience. **a** Data gathered from different subjects can be mapped to a common "information space" that is shared across all subjects in an experiment. This mapping is computed for each subject individually. Once data is transformed from the native signal space (e.g., voxel space) into an information space via an inverted encoding model for reconstructing represented information (EMRRI), representations of stimuli can be compared across different task demand conditions. Each data point in information space constitutes a representation of the stimulus viewed by the subject. Furthermore, this procedure eliminates the need to coregister brains from different subjects to one another. All analyses are performed in the common information space, so ROI definitions can be performed at an individual-subject level. **b** Examples of different information spaces which have been used in conjunction with the EMRRI technique (though note that this is not meant to be an exhaustive list). Panels adapted, with permission. (Color: [24]; Orientation: [18]; Spatial position: [23])

However, a representation-based approach can more directly address these questions by extending the observation-based approach. The observed neural responses of all measurement units (neurons, voxels, scalp electrodes, etc) are passed through an explicit model that transforms the high-dimensional pattern of neural activity into a 'stimulus space' (Fig. 12.1). The data, after being transformed into a common stimulus space, now form an explicit *reconstruction of the stimulus* as represented by the group of observed measurement units (neurons, voxels, or electrodes), and the representations associated with different experimental manipulations can be compared side-by-side in a more intuitive and cognitively-relevant way (i.e., similar to comparing two photographs of identical scenes taken under different lighting conditions, or

temperature readings taken from different cities). This approach brings us closer to directly examining how mental representations are formed and transformed across the brain while subjects perform complex cognitive tasks involving visual attention, learning, and short-term memory.

In this chapter, we will review both the more traditional observation-based techniques and the newer representation-based techniques with a focus on the relative strengths and weaknesses of each approach. We first review observation-based analysis approaches that focus on decoding stimulus features [8–13] and that seek to identify independent variables that best predict brain activity [14–16]. Then, we contrast these approaches with complementary representation-based methods that attempt to directly reconstruct stimulus features based on patterns of neural responses [17–28]). In all cases, we place particular emphasis on how these different methods can be used to test formal models of top-down cognitive control.

12.2 Human Neuroimaging Tools: Advantages and Disadvantages

Like any measurement tool, human neuroimaging methods (here, we'll focus primarily on BOLD fMRI and EEG, introduced in earlier chapters) carry both advantages and disadvantages. The most obvious problem with the BOLD signal is that it assesses changes in the magnetic properties of hemoglobin across relatively large cortical areas ($\sim$ 1–2 mm, [29]), and our relatively poor understanding of neuro-vascular coupling means that the link between the BOLD signal and actual changes in neural activity such as spike rates is not entirely clear [30–34]. Moreover, most typical measurements of the BOLD signal are temporally sluggish, with the signal peaking many seconds following the onset of stimulus-evoked neural activity and well after most task-related cognitive operations have finished. Similarly, even though EEG signals are instantaneously related to neural activity and complement the BOLD signal with their millisecond temporal precision, the EEG signal aggregates electrical activity across large populations of neurons, and is further distorted as those signals pass through the scalp and other tissue. Thus, the neural generators that drive EEG signals are difficult to unambiguously identify.

However, despite these weaknesses, BOLD and EEG neuroimaging are being increasingly used for a purpose at which they excel: making inferences about how cognitive events influence population level responses within and across functionally specialized regions of human cortex. There are several reasons this ability should not be underappreciated, as it provides a unique perspective on information processing that is currently unavailable in the domain of complementary techniques such as single-unit neurophysiology or two-photon Calcium imaging in animal model systems. First, perception and behavior are not typically linked to the response properties of single neurons but are instead thought to be linked more closely with the joint activity of millions of neurons that form population codes representing everything from basic sensory features to complex motor plans [35–38]. Recently developed methods allow BOLD and EEG measures to assess these population responses with increasingly

high precision [17–28, 39–50], and in turn these responses are likely to be much more closely coupled with perception and behavior compared to isolated single-unit spike rates. Second, these imaging tools can be used to assess changes in neural activity associated with nuanced and complex cognitive tasks that human subjects can master in a matter of minutes but that non-human primates would be unable to easily perform.

12.3 Univariate and Multivariate Decoding Methods: Labeling Trials Based on Patterns of Brain Responses

BOLD neuroimaging experiments generate large and information-rich datasets, with independent measurements obtained from thousands of voxels ("volumetric pixels", often $>$ 50,000 are measured from each subject's brain) every few seconds (typically 0.25–1 Hz). A main goal of cognitive neuroscience, particularly in investigating the role of top-down factors in mediating the efficiency of sensory processing, is to make inferences about whether a particular cognitive manipulation increases or decreases the amount of information about a stimulus display encoded by neural responses. However, the majority of early fMRI studies focused on analyzing each voxel independently (i.e., making a large number of univariate comparisons) such as: is the activity in each voxel greater during condition 1 or during condition 2? In turn, a spatial cluster threshold is often imposed to correct for multiple comparisons, and groups of contiguous voxels exceeding this threshold are considered to differ significantly with respect to the experimental factor(s) of interest. This type of analysis is an early and clear example of the observation-based approach described in Sect. 1.

Moving a step beyond this early analysis technique, researchers soon began to ask more nuanced questions about how well activation levels in a voxel (or the aggregate activation level averaged across a cluster of voxels in a region of interest, or ROI) could correctly predict which experimental condition evoked a response. This type of analysis is often called "decoding" or "stimulus classification" [8–12]. To perform this analysis, experimenters must be extremely careful to first separate their data into two distinct parts: a set of data that is used to "train" the decoder, and an independent set of data that is used to evaluate—or to "test"—the effectiveness of the decoder; it would be cheating to use the same data to both train and test a decoder, as any decoder, even one that was given pure noise as input, would always be at least somewhat successful [8–12]. This separation of the data into two independent sets is typically referred to as cross-validation.

The subsequent decoding analysis follows in two stages. First, the response in each voxel (or the average response across all voxels in a ROI) is measured in each condition using only data from the training set. Then a decoder can be used to determine if activity measured in a given voxel or ROI on a trial from the test set more closely resembles the mean response evoked in the training set by either condition 1 or by condition 2 (Fig. 12.2). Note that in this simple univariate case, the decoder is just a simple comparison rule that assesses the distance between the response on the test trial and the mean response from each condition in the training set.

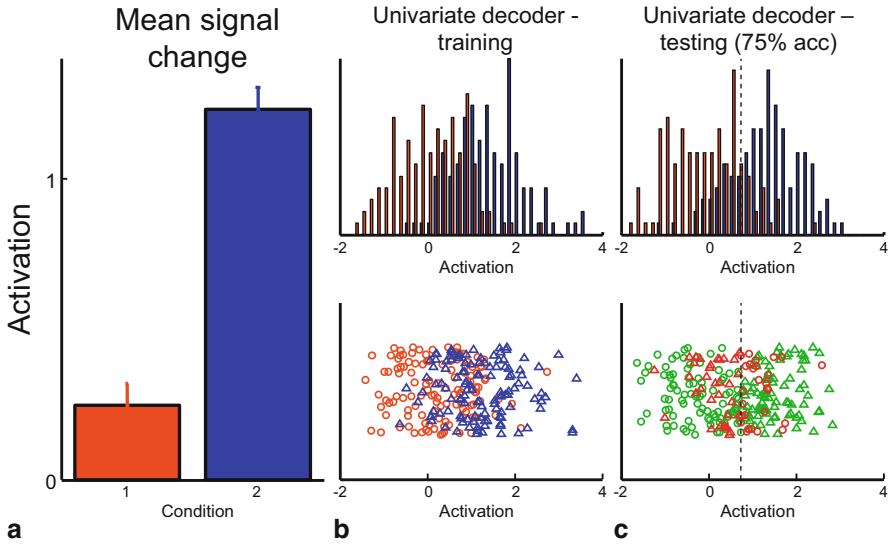


Fig. 12.2 Univariate decoder. For any region (or voxel) in which activation level discriminates between multiple conditions, a simple univariate decoder can be built. **a** Simulated mean activation of a region for which a univariate decoder would be successful. Activation is different between two conditions. **b** Data from *A*, training set. Each trial is plotted as a symbol (histogram above). *Orange circles*: condition 1, *blue triangles*: condition 2. Data are spread over the *y* range for clarity; this is a univariate measure. **c** Novel data, decoded using a classifier trained on the training set (*B*). The decision boundary is plotted as a dashed black line. Trials are color-coded according to decoding accuracy (green is correct, red is incorrect), and the symbol represents the correct condition label. By knowing only mean activation in this example, trials can be sorted with 75 % accuracy

More recently, studies have exploited the full information content contained in the spatially-distributed patterns of fMRI responses across *all* voxels within a ROI to construct *multivariate decoders* [8–12, 15, 17, 39–41, 51]. In principle, these decoders work much the same way as the simple univariate example described above, but by adding more voxels (where each voxel is often referred to as a *variable* or a *dimension* along which the decoder operates), information that is often obscured by averaging across all of the voxels in a ROI can provide a far more powerful means of correctly categorizing a data pattern on a novel trial into the correct experimental condition. For example, a simple two-class decoder can be imagined as a small ROI which consists of, say, 2 voxels (Fig. 12.3; and note that this example generalizes to any arbitrary number of voxels in a ROI—a 2 voxel ROI was chosen solely for ease of exposition). During condition 1, voxel 1 responds strongly, but voxel 2 responds weakly. During condition 2, voxel 2 responds strongly, but voxel 1 responds weakly. If data from this ROI were analyzed using a univariate analysis in which the responses were averaged across all voxels, then all information that discriminates condition 1 from condition 2 would be lost (Fig. 12.3c). In contrast, if the voxels are treated independently, a simple multivariate decoding scheme can be constructed and data from a novel trial could be easily classified as correctly belonging to either condition

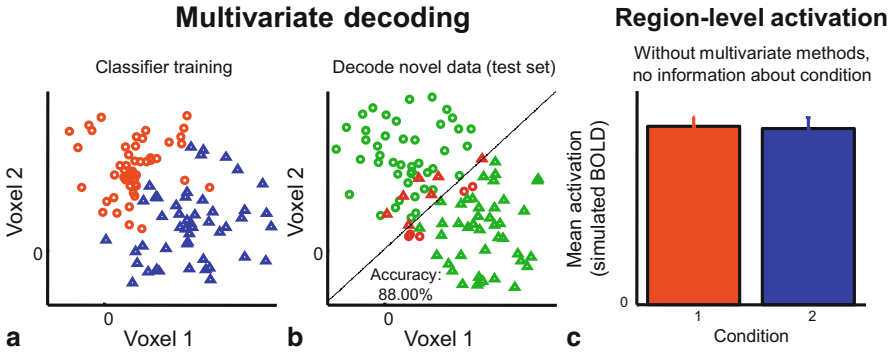


Fig. 12.3 Multivariate decoder. We generated a hypothetical dataset in which activation across a "region" of 2 voxels carries information which can discriminate between 2 conditions, but the mean activation level is constant across the two (compare to Fig. 12.2). We can then make a scatter plot of the activation in each voxel for every trial, and color-code the data points according to their condition **a**. In order to identify which of two conditions (condition 1, *orange circles*, condition 2, *blue triangles*) the activation from a given trial corresponds to, we first train a linear classifier to find a line (because this is a 2-dimensional classifier, for a higher-dimensional, more realistic voxel space, this would be a hyperplane) which best discriminates conditions 1 and 2. Then, we use an independent test set to evaluate how well this decision rule discriminates between the two conditions **b**. Trials in the test set known to be from condition 1 (*circles*) and condition 2 (*triangles*) are color-coded based on whether they are accurately classified (*green* is correct, *red* is incorrect). **c** Without a multivariate analysis, this simple "region" would be assumed to carry no information about which condition a trial belongs to

1 or 2 (Fig. 12.3a, b). Importantly, this logic can be extended to situations in which far more than 2 voxels in a ROI are considered in a multivariate analysis (often 100 or more, depending on the brain region in question—the same logic that differentiates univariate from multivariate analyses applies for any number of voxels). Indeed, if the response pattern (or *vector* of response amplitudes for each voxel) across any arbitrary number of voxels in a ROI carries information that *reliably* discriminates between different experimental conditions, then a 'hyperplane' (a boundary plane in high-dimensional space) can be computed that best separates all response vectors associated with category 1 and all response vectors associated with category 2 from the training set. Then, a response vector across all of the voxels on a test trial can be categorized simply by determining on which side of the hyperplane the novel activation vector falls (see Fig. 12.3, and see [5–11] for further discussion).

In sum, a decoder can be univariate or multivariate. However, in practice, multivariate decoders are now more often used, as they are generally more powerful because they can aggregate and exploit even small differences across all of the single voxels in each response vector [8–12]. Additionally, these methods have recently been applied to electrophysiology datasets, both in human EEG [22, 52] and animal single-unit electrophysiology [53]. Note that decoding analyses are another example of an observation-based approach—all analyses are performed in the measured neural signal space (i.e., by partitioning signal space into different sections corresponding to the different stimulus classes).

12.4 Encoding Models: Predicting Neural Responses Using Explicit Models of Neural Activity

Both of the toy examples presented above (Figs. 12.2 and 12.3) constitute decoding methods in which a condition 'label' is assigned to either the activation of a single voxel or the response vector of activations across many voxels recorded on a given experimental trial. In contrast, recent work has also focused on the complementary goal of identifying the particular stimulus feature(s) that robustly drive the activation of a single voxel or the activation vector across a set of voxels. In other words, understanding how the activation of a voxel or set of voxels *encodes* information about a set of features in a stimulus. Thus, these are termed encoding models [15–17] (Figs. 12.4 and 12.5). Like decoding methods, an encoding model can be estimated on a voxel-by-voxel (univariate) basis (Fig. 12.4), or by combining information across many voxels (multivariate) to characterize voxel-level and region-level stimulus selectivity, respectively. Furthermore, note that the classical experiments in which response characteristics of single neurons are measured in cats and monkeys are also testing encoding models.

Because significant treatment has been given to multivariate decoding methods in the context of classification (identification of a brain state among several alternatives, see [8–12, 15, 17] for a review, and Sect. 3), we here focus on the applications of univariate and multivariate encoding models and their application to identifying the sensitivity profiles of individual voxels to visual [42, 44, 48, 54, 55] and semantic [43, 45, 46] features of the stimulus, and how those sensitivity profiles change as a function of task demands [23, 47, 54–56]—all *observation-based* endeavors. We also will address the novel procedure of "inverting" encoding models to reconstruct region-wide representations of stimulus features given patterns of neural activity [17–28]—early investigations using the *representation-based approach*.

First, we will describe several implementations of univariate encoding models which have enabled careful characterization of novel elements of the early visual system in humans using fMRI. Next, we will discuss insights gleaned about neural coding of visual and semantic information from univariate encoding models. Finally, we will summarize recent efforts by our group and others to use *inverted* encoding models to examine how top-down factors such as attention and working memory modulate visual representations carried by patterns of neural activity in order to adaptively encode relevant sensory information in light of shifting behavioral demands.

12.5 Univariate Encoding Models for Single Stimulus Features

Univariate encoding models characterize how the underlying neural populations within individual voxels are selective for particular sensory features. Typically, the end goal of these methods is to compare best-fit encoding models for voxels which belong to different visual field maps along the cortical surface or within subcortical structures. Comparing models in this manner allows for investigation into how

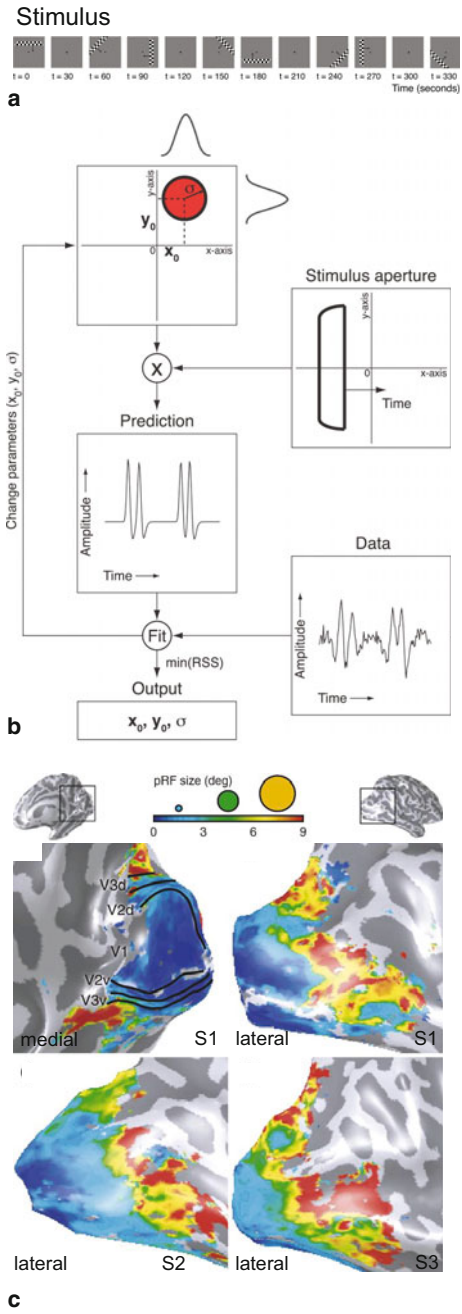


Fig. 12.4 Population receptive field modeling. Each voxel is modeled as responding to a single portion of the visual display. The population receptive field (pRF) modeling procedure operates by changing the location and size of a filter applied to a known stimulus sequence until a best-fit filter is identified for each voxel. This results in a set of estimated filter parameters (x, y, σ) for each voxel, which can then be plotted on the cortical surface. **a** Drifting bar stimulus typically used for pRF estimation. However, any stimulus which samples all positions of the screen can in principle be used for pRF estimation (see [23, 42]). **b** Model fitting procedure. A filter, here shown as a symmetric 2-dimensional Gaussian, is fit so that there is the minimum distance between the predicted response (given filter, stimulus, and a prototype hemodynamic response function) and the measured BOLD signal. This procedure is applied separately for each voxel. **c** pRF size plotted along the cortical surface for 3 participants (from [48]). Similar to measurements of spatial receptive field sizes in macaques, pRF filter estimates increase in size with increasing distance from the fovea. Source: [48] (*C*) and [49] (*A* and *B*)

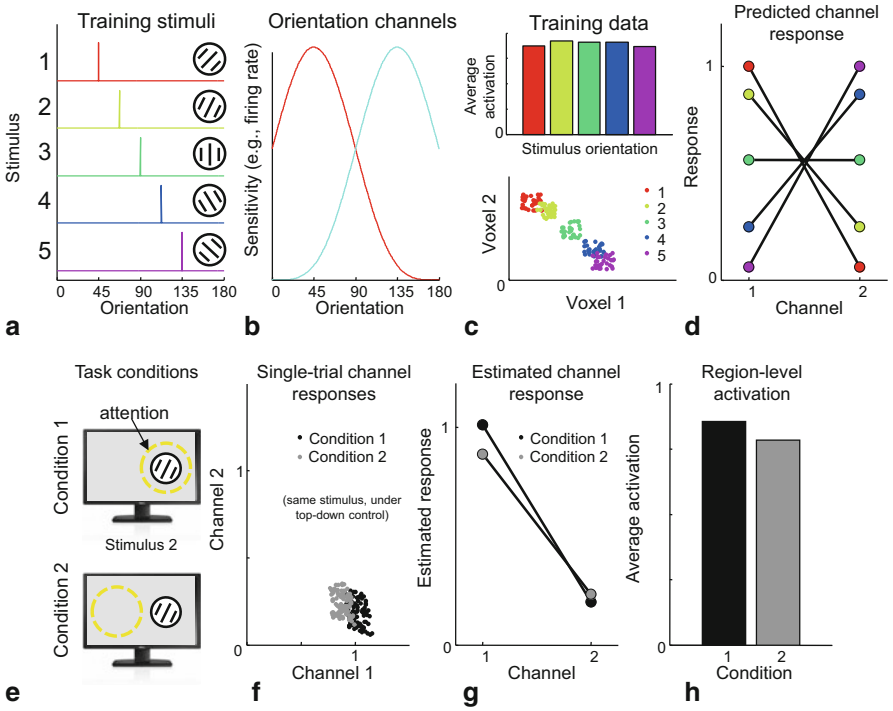


Fig. 12.5 Encoding models for reconstructing represented information (EMRRI). The response in each voxel is modeled using linear regression as a weighted sum of the responses of different hypothesized "information channels", corresponding to hypothesized response properties of neural populations. Once these weights are computed using a "training set" of data (*A-D*), test data collected in voxel (or electrode) space can be mapped onto "channel space" (*E-H*), which corresponds to the responses of putative neural populations. Here, we illustrate a simple 2-channel encoding model for stimulus orientation used to transform our measurement space (simulated signals from a 2-voxel "region") into a more intuitive channel space. Then, we examine how top-down control might change the way an identical stimulus is encoded in channel space. **a** shows the set of stimuli used to train a simple 2-channel encoding model for stimulus orientation **b**. Voxel responses to each of the 5 stimuli are recorded **c** and used in conjuction with the predicted channel responses **d** to estimate the weight for each channel within each voxel. Note that average "region"-level activation is identical across the 5 stimuli, but the multivariate pattern of responses varies (*lower panel*, *C*). Then, a participant might be asked to perform different tasks using these identical stimuli **e**. Signals measured in voxel space are then mapped into channel space **f**, in real data this typically involves mapping from high-dimensional voxel space, which contains hundreds of dimensions, into lower-dimensional channel space and sorted by task condition **g**. Here, we can see that task demands increase activation in channel 1, but slightly decrease activation in channel 2. If we were only to look at the mean signal in this region **h**, we might instead interpret the effects of top-down control as only increasing the mean response amplitude of the entire region with no change in represented information

cognitive variables (such as behavioral relevance) influence low-level sensory responses in the visual system. However, this approach can also be combined with stimulus classification, reconstruction, or other decoding approaches to further determine the encoding fidelity of a region (we return to this below in Sect. 7).

12.5.1 Feature-Selective Response Profiles

An early effort to characterize feature-selective response profiles at the level of the BOLD fMRI signal involved estimating voxel-level "tuning functions". When performing visual neurophysiology experiments on single neurons in animals it is often common practice to characterize the degree to which a recorded neuron is selective to a particular value of a given feature. For example, a monkey might be shown gratings of several orientations while firing rates are recorded from a visual neuron. The response of that neuron is then plotted as a function of the presented orientation, and this response vs. feature value plot is taken as the "tuning function" of that neuron.

In 2009, Serences et al. [54] adapted this procedure for use with fMRI data. The authors presented observers with gratings of different orientations while recording the BOLD signal from early visual cortex. They then plotted the response of each voxel as a function of the presented orientation, and subsequently binned voxels based on which orientation they responded to most strongly. This procedure resulted in "voxel tuning functions" (VTFs). The results of these studies confirmed predictions about how the shape of feature-selective response profiles should change with selective attention. For example, voxels that were 'tuned' to an attended feature were more active than voxels that were tuned far from the attended feature, a finding predicted by work using single-unit physiology [57, 58].

This method was recently adapted for use with a more complex set of visual stimuli: faces [55]. Similar to the procedure implemented by Serences et al. [54], Gratton et al. [55] binned face-responsive voxels (those which responded more to face stimuli than to non-face stimuli) according to the particular face stimulus along a continuous morph dimension which resulted in the highest BOLD response. The authors observed that voxels found in posterior, but not anterior, face-responsive ROIs showed tuning preferences to single faces. Anterior voxels within these ROIs instead seemed to respond differentially along a categorical boundary: whether the face was more masculine or more feminine. Additionally, in a follow-up experiment the authors asked observers to attend to one of two superimposed face stimuli. Posterior, but not anterior, face-responsive voxels selective for the attended face stimulus increased their response relative to voxels selective for the unattended face.

These results collectively demonstrate that voxel-level sensitivity to both simple (orientation) and complex (face) stimulus features can be estimated at the level of the fMRI BOLD signal, and that these voxel-level response profiles can be modulated as a function of top-down cognitive control.

12.5.2 Population Receptive Field Models

Traditional retinotopic mapping procedures identify the location of the visual field that best drives the response of a voxel [59, 60] and use resulting maps superimposed on a computationally-reconstructed cortical surface to delineate boundaries between visual field maps (such as V1, V2, etc [61]). Typically, this is accomplished using a periodic visual stimulus, such as a flickering rotating wedge (to identify polar angle) or a concentrically expanding annulus (to identify eccentricity). The response measured from each voxel is transformed into the frequency domain where the phase and power of the BOLD response at the stimulus frequency can be used to identify the preferred visual field location of a voxel and the strength with which that visual field location drives that voxel (for more information, see [59, 60]). While this approach has been successfully used for nearly two decades to identify continuous visual maps in occipital, parietal and frontal cortex [61, 62], these measurements only provide information about *where* in the visual field a voxel best responds to and do not support the precise characterization of other properties of voxel-level receptive fields, such as their shape or size.

In 2008, Serge Dumoulin and Brian Wandell [48] developed an analysis method that extends these traditional retinotopic mapping procedures by quantifying the spatial receptive field that best describes the visual sensitivity of each voxel—its "population receptive field" (pRF). As this method is an example of an encoding model, this analysis uses an explicit, experimenter-defined model which translates from stimulus space into an "information" space (corresponding to predicted activation, given a stimulus, for each voxel). In its most common implementation, the encoding model used for pRF analysis takes the form of voxel-level spatial receptive fields: responses in each voxel are assumed to be driven strongly when stimuli occupy a "blob"-like area of the screen. Some voxels, such as those towards the periphery, will likely respond to a larger portion of the screen than other voxels, such as those near the fovea, as will voxels in later regions of the visual processing stream [63, 64]. These response characteristics are predicted by well-documented single-unit recording studies that show a similar increase in receptive field size both with increasing eccentricity and for later visual areas.

The specific form of the encoding model most often used to describe the "blob" on the screen is a 2D isotropic (round) Gaussian (Fig. 12.4a, b [48]) or difference of Gaussians [49] (though others have also been used; [23, 65]). These functions can be characterized by a small number of parameters, such as their center position on the screen (x, y coordinates) and their size (standard deviation of the Gaussian, σ). pRF analyses involve using an encoding model (e.g., 2D Gaussian) to predict the response of a voxel given an assumed set of values for each of these parameters and the stimulus presented to the participant, comparing that prediction to the measured response in a voxel, and adjusting the parameters so as to maximize the prediction accuracy of the model (Fig. 12.4a, b). A good "fit" for an encoding model is one which accurately predicts the response of that voxel on runs that were not used to fit the encoding model (another example of cross-validation, see Sect. 3). Because finding

an optimal model using limited data becomes increasingly challenging with a greater number of free model parameters, functions characterized by a smaller number of free parameters ensure voxel responses can be well-fit. Additionally, adding parameters might lead to overfitting rather than to a model that more accurately describes the true sensitivity profile of a voxel (also see [23, 42, 65] for various implementations of a higher-dimensional pRF mapping approach which implements a regularized regression procedure). In the end though, because the pRF technique seeks to catalog visual sensitivity profiles for individual voxels, we consider it an example of an observation-based approach described in Sect. 1.

12.5.3 Novel Results Using pRF Methods

As mentioned above, the pRF model has added a further dimension to standard retinotopic mapping approaches: the size of the visual field which drives a voxel (Fig. 12.4c). This analysis technique has also proven to be useful in identifying visual field maps in regions with maps that are difficult to identify using standard polar angle or eccentricity mapping stimuli. For example, Amano et al. [66] used a standard pRF-mapping approach to characterize 4 extrastriate (outside primary visual cortex) visual hemifield maps: LO-1, LO-2, TO-1 and TO-2. TO-1 and TO-2 are novel regions identified using the more-sensitive pRF approach and likely correspond to the human homolog of macaque motion processing regions MT and MST [66]. Additionally, the pRF method allowed for a potentially more accurate RF size estimate in LO-1 and LO-2, which are the visual field maps corresponding to functionally-defined object-selective lateral occipital complex (LOC; [67]).

While pRF modeling techniques initially proved useful for describing the voxel-level sensitivity profile across the visual field (the spatial receptive field of each voxel), these methods have also been applied to characterize voxel-level tuning to a higher-order feature of a visual display: numerosity [50]. Characterizing voxel-level tuning to numerosity is an especially interesting question because number is an inherently cognitive attribute of the scene that can remain identical despite drastic changes in the stimulus on the screen. For example, the "eight-ness" of a stimulus could be conveyed by an image with 8 different-sized dots, or 8 small drawings of airplanes, or some drawings of trains and others of airplanes [68].

Adapting the pRF technique, Harvey et al. [50] built a pRF-type encoding model for numerosity which described a voxel's sensitivity to number as a function of its numerosity preference and tuning width (similar to how the traditional pRF approach describes a voxel's sensitivity to space as a function of its spatial receptive field center and size). This model was then fit using stimuli controlled across a large number of different features, such as screen area, circumference, and density. After visualizing best-fit models on the cortical surface, the authors identified a region of the intraparietal sulcus which contains a topographic representation of stimulus numerosity, similar to the topographic representation of spatial position in the early visual system.

Looking forward, pRF models are well suited as a novel tool for evaluating visual system organization and the role of top-down factors such as attention in mediating stimulus selectivity. However, some outstanding caveats remain that are important to keep in mind. A chief assumption of most currently published pRF methods is that voxel-level receptive fields are round. This assumption works well in many cases, but one could easily imagine a voxel which contains neural sub-populations with spatial RFs tuned to regions of space that are shaped like an ellipse (see [65]). Or, in a more extreme case in which a voxel spans two disjoint neural populations across a sulcus, the true pRF would be two distinct "blobs" of the visual field. The function used to describe pRFs necessarily restricts the kinds of inferences that can be made about visual response properties. Furthermore, pRF analysis is an example of a technique in which activation pooled over a large number of neurons is used to infer collective RF properties across that population (see [49]). This strategy is very successful for identifying properties of visual field maps, though the inferential power of this technique (and other population-level techniques) becomes limited when it is used to identify *changes* in RF properties with task demands. For example, pRFs computed under different stimulus or task conditions might change size (e.g., when attentional state is manipulated; [23, 56], see Sect. 6). This cannot be used as evidence to suggest that top-down cognitive control via attention acts to change *neural* RF properties, as several different patterns of neural modulation could give rise to this population-level behavior. For instance, neurons with RFs centered at the edge of the pRF (i.e., those near the edge of a voxel's RF) could simply increase or decrease their response, which would result in a larger or smaller estimated pRF despite no change in the size of neural RFs. And finally, while the insights gleaned from careful application of the pRF technique have improved our understanding of how different visual ROIs selectively respond to particular stimulus features, this technique remains an example of the observation-based approach in neuroscience. These methods advance models that predict neural responses to visual stimuli, but how the collective neural responses across entire brains or ROIs together represent visual information cannot be inferred using these methods.

12.6 Using Univariate Encoding Models to Observe Complex Feature Selectivity

The canonical pRF analysis (Sect. 5; [48]) has been remarkably successful at characterizing visual response properties of individual voxels when applied to relatively simple and artificial visual stimuli (Fig. 12.4a). However, electrophysiological recordings from single units have revealed more complex response properties, such as tuning for orientation [69], spatial frequency [70], motion direction [71], motion speed [72], color [73], numerosity [68] and even complex form [74]. Simultaneous with the development of the pRF method, other groups have developed more complex, high-dimensional encoding models to relate activation changes at the

single-voxel level to more specific visual features of images, such as local spatial frequency, orientation, and motion energy information (e.g., [42, 43]).

In one early report, Kay et al. [42] presented 2 participants with briefly flashed natural image stimuli (1705 black and white photographs) while measuring the BOLD signal from early visual areas. The authors sought to describe each image using a high-dimensional feature space, and then, using measured brain activity in response to each image, map changes in feature space across images to changes in BOLD responses for each voxel. The authors used an encoding model which described each stimulus image via the combined response of 10,921 Gabor filters [43], with each filter having a different orientation, spatial location, and spatial frequency tuning. This model captures known single-unit response properties within the early visual system, and thus is a good candidate model for how voxels containing many functionally selective neural populations might respond to such images. When estimating the best-fit encoding model, the authors first estimated the response of each of the Gabor filters that was used to decompose the natural images using only data from a "training" set of images. Note that this step requires no neural data whatsoever. Then, they used observed activation levels within a given voxel and these estimated filter responses to computationally identify a model which best describes the observed activation in response to all the training stimuli as a weighted sum of the responses of the filters. That is, if neural populations corresponding to each filter exist in a voxel, and the encoding model correctly describes their response as a function of the visual stimulus (the filters accurately capture neural information processing), how much of each filter's corresponding neural population must be present within each voxel in order to best account for the measured signal?

Once the best-fit encoding model was selected for a voxel, the feature selectivity of that voxel could be quantified in terms of its spatial receptive field location, size, preferred spatial frequency and preferred orientation. In essence, this method and that used for pRF-based analyses are very similar: the experimenter searches to find the optimal set of visual stimulus features which drive a voxel. Furthermore, this method was shown to recover the same relationship between eccentricity and spatial RF size for each voxel that the pRF method achieved [48]. Though there are differences in the form of the filter models and their estimation procedures, the end goal of each analysis is the same: characterize the sensitivity profile of each voxel along a hypothesized feature space using a known stimulus set and measured activations.

A similar approach was applied by Nishimoto et al. [44] to identify voxel responses to natural movie stimuli which contain motion energy information. Interestingly, using a similar model to that used in Kay et al. [42], the authors identified a previously-unknown principle of visual system organization: voxels which prefer higher motion velocities are those which respond to more peripheral visual field locations.

Purely visual encoding models like these transform information that is in the domain of visual stimulus space (e.g., pixels on the screen) to a domain spanning a set of features that the visual system is believed to operate along (e.g., orientation, spatial frequency, numerosity) in order to characterize the stimulus features which best drive a given voxel. Recent work has also incorporated semantic properties of visual scenes (e.g., alive or not alive) to identify both the semantic "dimensions" along which

voxel sensitivities vary and how different semantic dimensions might be represented across the cortex [43, 45–47]. Traditional univariate analyses of fMRI responses when viewing semantic categories, such as faces or outdoor scenes, reveal rather discrete loci of activation (in the fusiform gyrus for faces and the parahippocampal gyrus for outdoor scenes, forming the "fusiform face area" and "parahippocampal place area", respectively [3–5]). However, a more flexible approach might identify a more nuanced representation of semantic information across much of the cortex. For instance, Naselaris et al. [43] extended these efforts by combining a structural encoding model for visual features like that shown in Kay et al. [42] with a model incorporating semantic labels of visual images. This revealed voxels which responded more strongly when animate categories were shown and responded less strongly when nonanimate image categories were shown. Furthermore, the authors compared encoding accuracy (evaluated using a validation stimulus set) to determine that the voxels which most accurately encode semantic information and the voxels which most accurately encode information about structural visual features are disjoint, with the former clustering in regions of occipital cortex anterior to early visual areas, and the latter clustering in early visual regions V1-V3.

Natural movie stimuli similar to those used by Nishimoto et al. [44] and the semantic labeling approach demonstrated by Naselaris et al. [43] were combined in a recent study in which whole brain data was acquired while participants viewed 2 h of natural visual movies which had been carefully labeled with 1750 hierarchically-organized semantic features [45]. Once a semantic encoding model was estimated for every voxel (which results in a weight for each of 1750 semantic features in each voxel), the correlated variation in semantic weights across all voxels was computed using principal components analysis. This analysis identifies which weights covary together most reliably. For example, if the only important dimension in the brain's "semantic space" were "face-like vs. house-like", voxels which care about faces would have positive weights for face-related labels and voxels which care about houses would have positive weights for house-related labels. In contrast, voxels would all have identical weights for any semantic label which was not related to faces or to houses. Thus, the first principal component of the weights for this hypothetical semantic space would describe variation along a "face vs. house" axis. In other words, all the variability in semantic weights could be described along a single dimension—the first principal component—in this much larger space (the space of all semantic label weights).

In the Huth study [45], the first 4 principal components derived from voxel-level semantic weights described moving vs. non-moving stimuli, social interactions vs. all others, civilization vs. nature, and biological vs. nonbiological stimuli, respectively, and were consistent across participants. Furthermore, the authors determined that this semantic space is smooth across the cortex and is not built only from multiple discrete processing modules (e.g., face areas or house areas).

12.6.1 Comparing Stimulus Sensitivity Across Task Demand Conditions

For several decades, neurophysiologists have recorded responses of single neurons while animals perform different types of tasks with identical stimuli [75–78]. Some of these experiments have additionally characterized how neural RFs change under different task demands ([79] for a review). Recently, investigators have extended these experiments to humans using fMRI in order to examine whether manipulation of task demands change *voxel-level* stimulus sensitivity profiles. We highlight several of these early results below.

In a recent report, Çukur et al. [47] asked observers to attend to faces or attend to vehicles while viewing several hours of natural movie stimuli. They then fit a semantic encoding model (like that used in [45]) to data from runs in which subjects attended to faces, and compared these estimated models to those fit to data when subjects were attending to vehicles. The authors found a remarkable whole-brain shift in voxel-level semantic encoding models towards the attended feature, resulting in a local expansion of semantic space around that feature. Importantly, this modulation of semantic sensitivity occurred across nearly the entire cortical surface, and was not confined or driven by changes in sensitivity within specific processing 'modules' (e.g., "face" or "place" areas).

As a simpler example, in a recent report [23] we compared voxel-level spatial RFs measured across several different attention conditions (attend to the fixation point, or attend to a flickering checkerboard stimulus used to map spatial RFs). When we compared the size of the spatial filters for each voxel measured across these conditions, we found that for several extrastriate visual ROIs (hV4, hMT + and IPS0) spatial RFs for most voxels increased in size with allocation of attention [23]. However, though voxel RF size increased in many voxels with attention, RF size shrunk in others. Later, we return to this issue.

A similar experiment was conducted in which the experimenters manipulated the attentional requirements of a fixation task while pRFs were estimated using a standard mapping stimulus (Fig. 12.4a; [56]). These authors found that, with greater attentional demands at fixation, pRFs for voxels tuned to higher eccentricity portions of the screen expanded across V1, V2 and V3. They interpreted this pRF expansion as a form of neural "tunnel vision" in which distracting stimuli (those used to map pRFs) are suppressed to a greater extent under task conditions in which there are more stringent attentional demands.

All of these encoding methods presented thus far—voxel tuning functions, population receptive fields, motion energy models, and hierarchical semantic models—remain examples of the observation-based approach. All these experiments have characterized changes in neural response or sensitivity properties as a function of stimulus attributes or task demands.

12.7 Understanding the Impact of Task Manipulations on Stimulus Representations Using Multivariate Inverted Encoding Models for Reconstructing Represented Information (EMRRI)

While the analysis techniques described above provide compelling evidence for the *existence* of information carried by a region (e.g., stimulus orientations can be decoded using V1 activation patterns, V1 responses are well characterized by motion energy, and more anterior regions are better described with semantic labels), these methods have recently been adapted to use the responses across all encoding units (e.g., voxels) to constrain estimates of the information content within entire regions of interest. This has allowed for a new paradigm for testing models of cognitive control over representations of information in which the region-wide *representation* of a feature of interest can be *reconstructed* from brain activation patterns (Fig. 12.1). This is an important step, because when forming hypotheses about how behavioral goals might change underlying cognitive states which in turn influence behavior, we are often more interested in understanding how *information* is manipulated, not in how *voxels* or *neurons* are manipulated.

12.7.1 Estimating Information Content using Inverted Encoding Models

In 2009, Gijs Brouwer and David Heeger [24] implemented a novel twist on the more standard implementations of an encoding model which allowed them to use patterns of activation across entire ROIs measured using fMRI to reconstruct region-wide representations of particular features of visual stimuli. In this study, the authors built an encoding model which described feature-selective responses for color hue. In this implementation, rather than modeling information channels which correspond to local spatial [42] or motion energy [44] filters, the authors modeled only responses to colors of different hues but identical luminance. Using this model, they could predict the response of each information channel to any stimulus presented to a participant (similar to the high-dimensional encoding models described above; see Fig. 12.5a, b, c, and d for a graphical depiction of a simple example encoding model for a single stimulus feature), and accordingly estimate the contribution of each information channel to the observed signal in each voxel. At this point, the authors had computed a set of weights which characterize how the measured BOLD response from each voxel is modulated by the activity of each information channel (this is essentially identical to the encoding models described in Sect. 5 and 6). Because the authors were modeling BOLD activation as a linear combination of information channel responses, once the weights of each channel were estimated they were able to "invert" the model to determine, for a given pattern of BOLD activation, how strongly each information channel must have been activated. The result of this model inversion is a matrix which maps from a high-dimensional signal

space (here, BOLD activation in voxels) to a lower-dimensional information space (here, responses in several hue-selective channels). This matrix can be used for any novel pattern of BOLD data to "reconstruct" the information content of that BOLD response pattern. To contrast this approach with that described in the previous section, we will refer to encoding models built with the purpose of reconstructing information content represented across groups of voxels as "encoding models for reconstructing represented information", or EMRRI.

Using this new EMRRI technique, the authors confirmed earlier findings which suggested that human extrastriate visual area V4 carries information about color. Furthermore, activity in this region could be used to reconstruct novel colors in a continuous fashion (in addition to being able to correctly classify which of several colors was viewed). This is especially important for evaluating information content across different task demands. We are often interested not only in whether the information carried within a ROI changes under different top-down cognitive requirements, but *how* that information changes (see Fig. 12.5e, f, g and h for a demonstration of how the EMRRI technique can be used to establish the manner in which represented information changes with different task demands).

In further work, these authors have extended these methods to reconstruct the orientation [25] and motion direction [27] of presented stimuli, as well as to evaluate how attention to color induces categorical structure in the neural color space of extrastriate visual regions hV4 and VO1 [28], and to reconstruct a participant's eye position using activation from visual ROIs [26].

Next, we present some recent work from our lab in which we have applied the EMRRI method to evaluate the quality of information represented in different regions as a function of attentional factors and task demands.

12.7.2 Optimal Feature Representation

How does cognitive control act to best represent the most relevant information for accomplishing a given task? Because the early visual system has well-characterized responses to simple stimuli, such as oriented gratings, and in light of novel analysis tools described above, Scolari et al. [18] used the EMRRI technique to reconstruct the response of stimulus orientation channels under different attentional demand conditions. They used a task in which participants were asked to determine whether a change in orientation was in the clockwise or counter-clockwise direction from a presented target orientation. When this task requires fine discrimination of very small orientation changes, neural populations tuned away from the target orientation ('off-target' neurons) should undergo a large change in firing rate with attention because they better discriminate nearby stimulus features [80–83]. On the other hand, when the task involves a coarse discrimination—that is, when the orientations to be discriminated are far away from the target—the optimal sensory strategy is to enhance populations tuned to the target orientation as these neurons undergo the largest firing rate change in response to the two stimuli.

In order to evaluate how different signal channels are modulated when attentional demands are varied, Scolari et al. [18] used a simple encoding model for stimulus orientation—based on the EMRRI approach of Brouwer and Heeger [25]—to measure responses in different orientation channels as participants were performing either a fine orientation discrimination task or an equally-challenging contrast discrimination task with identical visual stimuli. Thus, the only parameter manipulated by the experimenter was the task performed by the participant, and accordingly any changes in measured orientation channel response reflect top-down attentional control over the representation of sensory information. The results are consistent with the pattern predicted for optimal attentional gain: in early retinotopic visual areas V1-V3, activation was relatively high in the most informative off-target populations during the fine discrimination task compared to the control task requiring attention to stimulus contrast (Fig. 12.6a). Importantly, the implementation of EMRRI technique allowed for a continuous, trial-by-trial estimate of the response in each of the modeled orientation channels. Though a classification experiment may have revealed greater decoding accuracy when the orientation of the stimulus was attended [84], such an experiment would not yield direct insight into *how* the information represented in early visual cortex was adjusted under top-down control [17]. In contrast, the application of an encoding model allowed for the direct assessment of a functional model of attentional modulation of sensory representations.

The EMRRI technique was again implemented by Ho et al. [19] to identify how the relative emphasis of decision speed or accuracy (the speed/accuracy tradeoff) modulates the encoded information content when subjects were asked to make a fine discrimination between two oriented gratings (i.e., is there a 5° offset between stimuli?). By combining a mathematical model in which cognitive parameters underlying perceptual decision making, such as drift rate (speed at which evidence is accumulated from a stimulus) can be estimated using response choices and response times [85] and an inverted encoding model for stimulus orientation, the authors identified a novel mechanism whereby sensory processing is optimally enhanced when participants are asked to emphasize accuracy over speed. This sensory encoding enhancement mirrors that found above [18] in that responses in off-target orientation channels that putatively carry the information most relevant for a decision are enhanced more for correct trials than for incorrect trials (Fig. 12.6b). This revealed that decision mechanisms can selectively pool inputs from the most informative sensory signals [19] to facilitate efficient and accurate decision making.

12.7.3 Sensory Signals Maintained During Working Memory

Visual working memory experiments are useful test-cases for examining top-down influences on perceptual representations even in the absence of visual stimulation (i.e., information stored in the ‘mind’s eye’). Several earlier demonstrations have

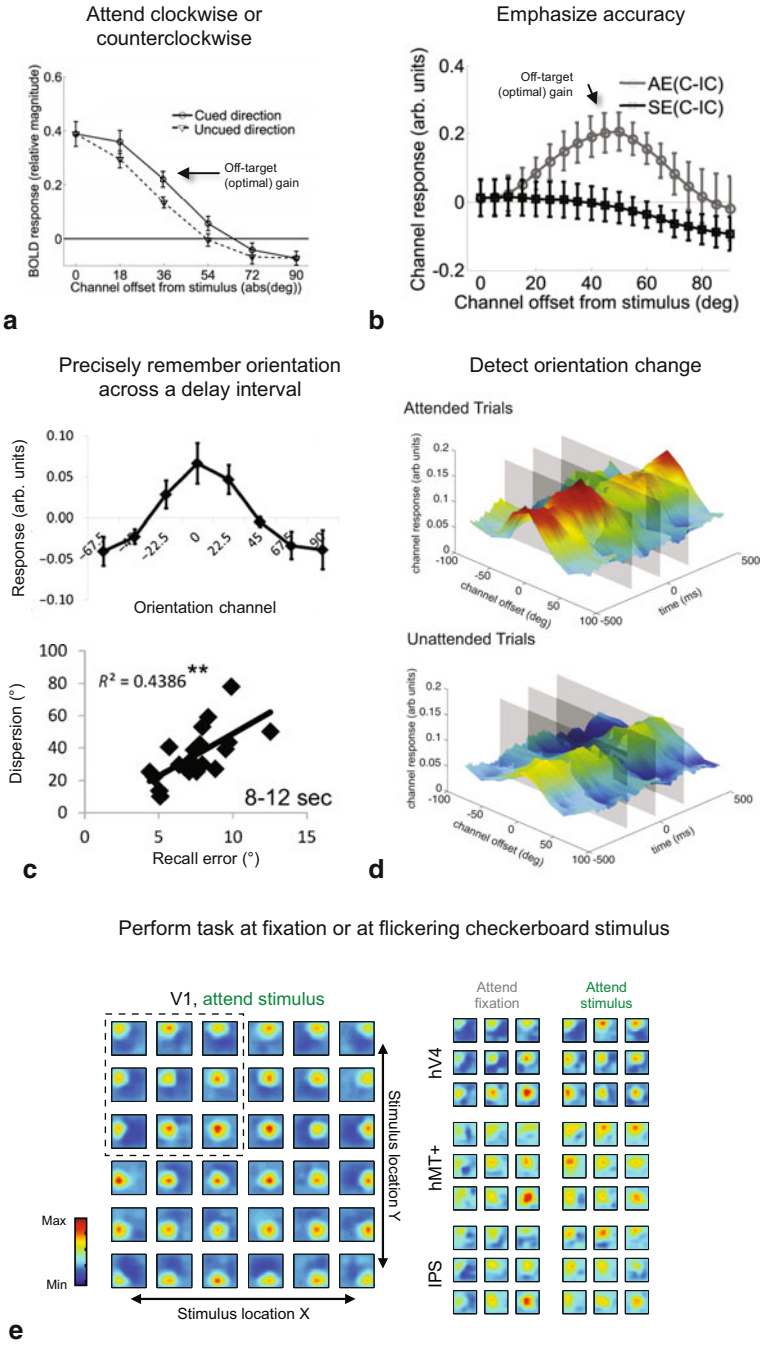


Fig. 12.6 EMRRI enables measurements of top-down control over sensory encoding processes by implementing the EMRRI technique to analyze visual cortex activations while performing attention or working memory tasks with simple visual stimuli, we have revealed several features of top-down

shown that the private contents of visual working memory could be decoded using classifier techniques described above [41, 52, 86–89]. Recently, Ester et al. [21] built on this work using the EMRRI method to reconstruct stimulus orientation using activation measured during a delay interval during which participants remembered the precise orientation of a grating that was no longer present on the screen. They found that the reconstructed orientation response profile measured from early visual cortex (that is, the response of a modeled set of orientation channels reconstructed from the pattern of brain activation) can predict both the orientation of the remembered stimulus and the precision with which each participant maintained a representation of the remembered orientation (Fig. 12.6c). During the delay period, the orientation channel corresponding to the orientation held in working memory had the highest activation, and activation fell off in a graded fashion for more distant orientation channels. Furthermore, for participants with narrow reconstructed orientation response functions during a delay interval (that is, a more precise representation), orientation recall performance was better than that for participants for whom orientation response functions were wider [21]. Moreover, the overall amplitude of the reconstructed response profiles was not related to working memory precision. Thus, without using an inverted encoding model to reconstruct orientation representations, it may have been possible to observe better decoding accuracy in participants with smaller recall errors, but the selective link between the width, but not the amplitude, of orientation-selective response profiles during a delay interval and recall performance could not have been identified.

12.7.4 Attentional Modulations of Spatial Representations

Recently, we have extended the EMRRI technique to allow us to reconstruct spatial information [23]. Instead of using orientation tuning functions to predict channel

←

Fig. 12.6 (continued) control over sensory encoding. **a** When a participant is asked to make a challenging orientation judgment and is cued to the direction of an orientation change, top-down processes enhance the response of information channels tuned *away* from the target orientation (*arrow*), which is an optimal strategy as changes in activation of these channels carry more information about fine orientation changes than do activation changes in channels tuned to the target orientation (Source: [18]). **b** When accuracy is emphasized (*AE*) during a similar task (as opposed to speed, *SE*), channels tuned away from the target orientation are enhanced more on correct trials (*C*) than on incorrect trials (*IC*) than are channels tuned to the target orientation (*arrow*). Differences between Correct and Incorrect (*C-IC*) plotted. (Source: [19]). **c** Participants are asked to precisely remember an orientation over a delay interval. During the delay interval, responses of channels tuned to the remembered orientation are enhanced (*upper panel*). The dispersion of channel response functions (width of *curve* in *upper panel*) during the delay period predicts recall error for individual participants (*lower panel*). [21]. **d** Using steady-state visual evoked potentials (SSVEP) measured with EEG, near-real-time channel response functions can be measured when a participant is attending to an oriented grating (Attended Trials) or an RSVP letter stream presented at fixation (Unattended Trials) (Source: [22]). **e** Reconstructed spatial representations using activation patterns from several visual ROIs reveal that spatial attention to a stimulus enhances its response amplitude, but does not change its represented size. (Source: [23])

responses during the training phase (Fig. 12.5a, b, c and d), we instead used simulated spatial filters (similar to spatial receptive fields). The result of this application of the EMRRI analysis framework is a reconstructed spatial representation carried by a particular set of voxels of the visual scene viewed by an observer.

We used this technique to examine how spatial representations of simple visual stimuli (flickering checkerboard discs presented at different positions around the screen) are modulated by attention. It might be that spatial attention simply enhances the net response to a stimulus within a particular ROI (that is, raising its mean response amplitude). Or, it could act to maintain a "sharper" representation of the visual scene (effectively shrinking this stimulus representation). Or, the amplitude (or "contrast energy") of the attended stimulus could be selectively enhanced independent of other parts of the scene.

To explore these possibilities, we scanned participants while they performed different tasks using identical visual stimuli. Then, we compared reconstructed spatial representations computed using activation patterns from each of several visual ROIs during each task condition. When participants attended to the contrast of a flickering checkerboard disc, the spatial representation carried by hV4, hMT+, and IPS increased in amplitude above baseline, but remained a constant size (Fig. 12.6e). Interestingly, as described in Sect. 6, voxel-level receptive field sizes on average *increased* with attention in these same ROIs. Importantly, though, some voxels showed an increase in RF size with attention, while others showed a decrease, and this *pattern* of voxel-level RF size modulations with attention together resulted in an increase in spatial representation amplitude, but no change in size [23]. When we instead simulated data in which the mean RF size change remains the same, but the pattern of RF size changes across voxels is randomized, then spatial representations increased in size. This result highlights the importance of considering attentional modulations of encoding properties across *all* encoding units, rather than simply considering the direction of the average modulation.

By using an inverted encoding model to reconstruct the spatial representation carried by each ROI, we were able to more directly evaluate the role of attention on modulating the information content of different brain regions, rather than its effects on signal amplitude or classification accuracy. Furthermore, while the retinotopic structure of early visual cortex has allowed experimenters to assign voxels to visual field positions and use activation along the cortical map as a proxy for representation strength at a visual position [90, 91], this requires both a very confident and clear measurement of cortical topography, as well as an assumption of a one-to-one relationship between voxel responses and stimulus representation magnitude at the corresponding visual field position. The EMRRI technique, as described above, does not explicitly require retinotopic topography along the cortical surface—instead, all that's required is for different voxels to have different patterns of visual selectivity, and for their responses to be consistent across an experimental session. If these hold, any cortical pattern of responses to different visual field positions can be used to reconstruct spatial representations. This is especially useful for investigations of parietal and frontal regions which contain a degraded retinotopic representation of the visual field compared to early visual cortex.

12.7.5 EMRRI Analysis with EEG Reveals Temporally Dynamic Stimulus Representations

While the experiments described above have proven important for understanding how cognitive control can act to optimize representations of sensory information, they are inherently limited because they have utilized fMRI data and necessarily lose a great deal of temporal resolution. In contrast, methods such as EEG and MEG can be used to provide high temporal resolution metrics of information processing, but have rarely been used to evaluate feature-selective attentional modulations. Thus, applying inverted encoding model analysis techniques to measures of brain activity with higher temporal resolution, such as EEG or MEG, allows us to place important constraints on the temporal dynamics of top-down influences on neural representations of information.

Garcia et al. [22] measured EEG responses to flickering oriented gratings and a rapid serial visual presentation (RSVP) letter stream presented at fixation while participants either detected a slight change in stimulus orientation (attend orientation) or a letter target which appeared in the RSVP letter stream (attend letters). By rapidly flickering the gratings (21.25 Hz), they were able to isolate an EEG signal called the steady-state visual evoked potential (SSVEP), which is a peak in the frequency spectrum at the stimulus flicker frequency and its harmonics [92]. By flickering the oriented grating and the RSVP letter stream at different frequencies, power at the corresponding frequency could be isolated and used in combination with the EMRRI technique to reconstruct stimulus orientation. First, the authors implemented a standard decoding analysis (Sect. 3)—could the orientation of the stimulus be decoded from the power and phase of measured SSVEP responses at the stimulus flicker frequency? They found improved decoding accuracy for stimulus orientation when the stimulus was attended [93–95]. However, like decoding analyses in fMRI mentioned above, an improvement in decoding accuracy does not characterize how encoded information is manipulated [17].

To assess the representation of information more directly, the authors used an inverted encoding model and a wavelet decomposition of the SSVEP (which assesses changes in power at different frequencies across time) to reconstruct the response of 9 orientation channels at each point in time. This revealed the temporal evolution of top-down attentional modulations in human cortex (Fig. 12.6d). The observed "dynamic tuning functions" increased in amplitude when participants were attending to the stimulus (Fig. 12.6d), and they more strongly represented the stimulus orientation when participants responded correctly than when they responded incorrectly. Interestingly, the pattern of response enhancement with attention seen in these dynamic orientation representations more closely mimics that observed in electrophysiological experiments with macaques, in which responses to all orientations are increased by a similar coefficient ("multiplicative gain"), than the pattern of orientation responses observed in fMRI, in which responses of all orientation channels often increase by a similar amount ("additive shift", [58]). This suggests that SSVEP-derived orientation representations might more faithfully capture the underlying neural response properties than the fMRI BOLD signal.

So far, this novel analysis technique whereby an estimated encoding model is inverted in order to reconstruct the information content of a brain signal (EMRRI) has only been applied to simplistic, artificial visual stimuli such as moving dots [27], oriented gratings [18, 20–22, 25], color [24, 28] and spatial position [23]. However, in principle, the same approach can be extended towards more and more complex encoding models, which could be used to study visual stimuli more similar to those we encounter in everyday life. Though there are mathematical restrictions on the number of encoding dimensions which can be successfully reconstructed given a set number of signal dimensions and observations, within those constraints any well-built encoding model which accurately captures the relevant features for a neural system should be suitable to use for reconstructing information content. Furthermore, we have demonstrated that the EMRRI technique can successfully be applied to both fMRI and EEG datasets, but this is certainly not the full extent of measurement methodologies for which this approach is applicable. Both multielectrode single-unit electrophysiology and 2-photon Calcium imaging in animal model systems result in datasets that could be well-analyzed using the EMRRI technique.

12.8 Conclusions

Multivariate tools for analyzing neuroimaging data are reaching mainstream adoption. Decoding tools have been used to assess how cognitive control changes information encoded in early sensory areas [84, 87]. However, only the recent application of inverted encoding models to fMRI and EEG data has allowed for careful investigation into how representations of information in the human visual system change under differential top-down demands.

Several applications of encoding models have proven useful for understanding the kinds of information carried by different brain regions [42–45]. When simple stimuli are used, such as colored annuli [24, 28] or oriented gratings [18–22, 25], feature selective responses can be measured using fMRI and EEG to directly reconstruct the features viewed, attended to, or remembered by a participant. Furthermore, comparing changes in these feature reconstructions has proven useful for evaluating theories concerning how sensory representations should be optimally modulated during different task conditions [18–20, 23, 28], as well as theories describing how visual short-term memory utilizes sensory codes to maintain information across a delay [21].

Thus, we propose that these low dimensional and relatively simple invertible encoding models for reconstructing represented information (EMRRI) are primarily useful when identifying how low-level sensory representations change as a result of top-down control during tasks manipulating attention, planned eye movements, visual memory, and perceptual learning. In contrast, high-dimensional visual or semantic encoding models reveal novel insights into the way that more complex representations (e.g., semantic space) might warp under different attentional demands. Although these higher-level modulations are inherently more difficult to interpret, they move us beyond the realm of artifice in stimulus design, and closer to the

ultimate goal of understanding how the brain processes complex natural images. Together these approaches provide an exciting glimpse at how task demands shape the encoding, processing, and representation of visual scenes.

Exercises

1. Describe the differences between the "decoding" methods described in Sect. 3 and the region-level "encoding" methods for reconstructing represented information (EMRRI) described in Sect. 7. Sketch a hypothesis that would be best tested using decoding methods, and what you would learn using these tools. Then, do the same for EMRRI.
2. Using methods described in the chapter, design an experiment to test a hypothesis that the human amygdala preferentially identifies fearful features of faces. What kind of approach did you choose? How will you ensure your chosen analysis technique is free from confounds introduced during analysis?
3. Many of the examples of implementations of the EMRRI method described in this chapter (Sect. 6) have concentrated on sensory responses (and the manipulation of those sensory responses by top-down control). How would these methods be adapted to test hypotheses about, say, motor planning or decision-making?

Further Reading

1. A fantastic place to start for the interested student is the textbook "Visual Population Codes: Toward a Common Multivariate Framework for Cell Recording and Functional Imaging", edited by Nikolaus Kriegeskorte and Gabriel Kreiman.
2. For an additional perspective on the strengths and weaknesses of decoding methods and encoding methods, see Naselaris et al. (2011) and Serences and Saproo (2011; both cited below [15, 17]).
3. A classical study implementing decoding approaches is Kamitani and Tong (2005; cited below [39])
4. Some classical studies implementing novel encoding methods are the original pRF methods demonstration by Dumoulin and Wandell (2008), and the high-dimensional "systems identification" approach developed by Kay et al. (2008) and extended by Nishimoto et al. (2011; cited below [42, 44, 48])
5. For more discussion of the development and different implementations of inverted encoding model methods (EMRRI), see Brouwer and Heeger (2009), Scolari et al. (2012), and Sprague and Serences (2013; all cited below, [18, 23, 24]).

Acknowledgments Supported by NSF Graduate Research Fellowship to T.C.S. and R0I MH-092345 and a James S. McDonnell Scholar Award to J.T.S. We thank Alexander Heitman, Sirawaj Itthipuripat, and Mary Smith for useful discussion during the preparation of this manuscript.

References

1. Barlow HB (1972) Single units and sensation: a neuron doctrine for perceptual psychology. Perception 1:371–394
2. Hubel DH, Wiesel T (1959) Receptive fields of single neurones in the cat's striate cortex. J Physiol 148:574–591
3. Epstein RA, Kanwisher N (1998) A cortical representation of the local visual environment. Nature 392:598–601
4. Kanwisher N et al (1997) The fusiform face area: a module in human extrastriate cortex specialized for face perception. J Neurosci 17:4302–4311
5. Kanwisher N (2010) Functional specificity in the human brain: A window into the functional architecture of the mind. Proc Natl Acad Sci 107:11163–11170
6. Tsao DY, Livingstone MS (2008) Mechanisms of face perception. Annu Rev Neurosci 31: 411–437
7. Freiwald WA et al (2009) A face feature space in the macaque temporal lobe. Nat Neurosci 12:1187–1196
8. Tong F, Pratte MS (2012) Decoding patterns of human brain activity. Annu Rev Psychol 63:483–509
9. Norman KA et al (2006) Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn Sci 10:424–430
10. Cox DD, Savoy RL (2003) Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. Neuroimage 19:261–270
11. Kriegeskorte N (2011) Pattern-information analysis: from stimulus decoding to computational-model testing. Neuroimage 56:411–421
12. Haynes J-D, Rees G (2006) Decoding mental states from brain activity in humans. Nat Rev Neurosci 7:523–534
13. LaConte SM (2011) Decoding fMRI brain states in real-time. Neuroimage 56:440–454
14. Wu MC-K et al (2006) Complete functional characterization of sensory neurons by system identification. Annu Rev Neurosci 29:477–505
15. Naselaris T et al (2011) Encoding and decoding in fMRI. Neuroimage 56:400–410
16. Gallant JL et al. (2012) Systems Identification, encoding models and decoding models: a powerful new approach to fMRI research. In: Kriegeskorte N, Kreiman G (eds) Visual population codes, MIT Press, Cambridge, pp 163–188
17. Serences JT, Saproo S (2011) Computational advances towards linking BOLD and behavior. Neuropsychologia 50:435–446
18. Scolari M et al (2012) Optimal deployment of attentional gain during fine discriminations. J Neurosci 32:1–11
19. Ho T et al (2012) The optimality of sensory processing during the speed-accuracy tradeoff. J Neurosci 32:7992–8003
20. Anderson DE et al (2013) Attending multiple items decreases the selectivity of population responses in human primary visual cortex. J Neurosci 33:9273–9282
21. Ester EF et al (2013) A neural measure of precision in visual working memory. J Cogn Neurosci. doi:10.1162/jocn_a_00357
22. Garcia J et al (2013) Near-real-time feature-selective modulations in human cortex. Curr Biol 23:515–522 (Cell Press)
23. Sprague TC, Serences JT (2013) Attention modulates spatial priority maps in the human occipital, parietal and frontal cortices. Nat Neurosci 16:1879–1887
24. Brouwer G, Heeger D (2009) Decoding and reconstructing color from responses in human visual cortex. J Neurosci 29:13992–14003
25. Brouwer G, Heeger D (2011) Cross-orientation suppression in human visual cortex. J Neurophysiol 106:2108–2119
26. Merriam EP et al (2013) Modulation of visual responses by gaze direction in human visual cortex. J Neurosci 33:9879–9889

27. Kok P et al (2013) Prior expectations bias sensory representations in visual cortex. J Neurosci 33:16275–16284
28. Brouwer GJ, Heeger DJ (2013) Categorical clustering of the neural representation of color. J Neurosci 33:15454–15465
29. Hyde JS et al (2001) High-resolution fMRI using multislice partial k-space GR-EPI with cubic voxels. Magn Reson Med 46:114–125
30. Sirotin YB, Das A (2009) Anticipatory haemodynamic signals in sensory cortex not predicted by local neuronal activity. Nature 457:475–479
31. Cardoso MMB et al. (2012) The neuroimaging signal is a linear sum of neurally distinct stimulus- and task-related components. Nat Neurosci 15(9):1298–306
32. Devor A et al (2008) Stimulus-induced changes in blood flow and 2-deoxyglucose uptake dissociate in ipsilateral somatosensory cortex. J Neurosci 28:14347–14357
33. Logothetis NK, Wandell BA (2004) Interpreting the BOLD Signal. Annu Rev Physiol 66: 735–769
34. Heeger DJ et al (2000) Spikes versus BOLD: what does neuroimaging tell us about neuronal activity? Nat Neurosci 3:631–633
35. Pouget A et al (2003) Inference and computation with population codes. Annu Rev Neurosci 26:381–410
36. Kang K et al (2004) Information tuning of populations of neurons in primary visual cortex. J Neurosci 24:3726–3735
37. Seung HS, Sompolinsky H (1993) Simple models for reading neuronal population codes. Proc Natl Acad Sci 90:10749–10753
38. Johnson KO (1980) Sensory discrimination: decision process. J Neurophysiol 43:1771–1792
39. Kamitani Y, Tong F (2005) Decoding the visual and subjective contents of the human brain. Nat Neurosci 8:679–685
40. Kamitani Y, Tong F (2006) Decoding seen and attended motion directions from activity in the human visual cortex. Curr Biol 16:1096–1102
41. Harrison SA, Tong F (2009) Decoding reveals the contents of visual working memory in early visual areas. Nature 458:632–635
42. Kay K et al (2008) Identifying natural images from human brain activity. Nature 452:352–355
43. Naselaris T et al (2009) Bayesian reconstruction of natural images from human brain activity. Neuron 63:902–915
44. Nishimoto S et al (2011) Reconstructing visual experiences from brain activity evoked by natural movies. Curr Biol 21:1641–1646
45. Huth AG et al (2012) A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron 76:1210–1224
46. Çukur T et al (2013) Functional subdomains within human FFA. J Neurosci 33:16748–16766
47. Çukur T et al (2013) Attention during natural vision warps semantic representation across the human brain. Nat Neurosci 16:763–770
48. Dumoulin S, Wandell B (2008) Population receptive field estimates in human visual cortex. Neuroimage 39:647–660
49. Zuiderbaan W et al (2012) Modeling center-surround configurations in population receptive fields using fMRI. J Vis 12:10
50. Harvey BM et al (2013) Topographic representation of numerosity in the human parietal cortex. Science 341:1123–1126
51. Haynes J-D, Rees G (2005) Predicting the orientation of invisible stimuli from activity in human primary visual cortex. Nat Neurosci 8:686–691
52. LaRocque J et al (2013) Decoding attended information in short-term memory: an eeg study. J Cogn Neurosci 25:127–142
53. Meyers E, Kreiman G (2012) Tutorial on pattern classification in cell recording. In: Kriegeskorte N, Kreiman G (eds) Visual population codes, MIT Press, Cambridge, pp 517–538
54. Serences JT et al (2009) Estimating the influence of attention on population codes in human visual cortex using voxel-based tuning functions. Neuroimage 44:223–231

55. Gratton C et al (2013) Attention selectively modifies the representation of individual faces in the human brain. J Neurosci 33:6979–6989
56. De Haas B et al (2014) Perceptual load affects spatial tuning of neuronal populations in human early visual cortex. Curr Biol 24:R66–R67
57. Martinez-Trujillo JC, Treue S (2004) Feature-based attention increases the selectivity of population responses in primate visual cortex. Curr Biol 14:744–751
58. Saproo S, Serences JT (2010) Spatial attention improves the quality of population codes in human visual cortex. J Neurophysiol 104:885–895
59. Engel SA et al (1994) fMRI of human visual cortex. Nature 369:525
60. Sereno MI et al (1995) Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. Science (80-.) 268:889–893
61. Wandell B et al (2007) Visual field maps in human cortex. Neuron 56:366–383
62. Silver MA, Kastner S (2009) Topographic maps in human frontal and parietal cortex. Trends Cogn Sci 13:488–495
63. Gattass R et al (2005) Cortical visual areas in monkeys: location, topography, connections, columns, plasticity and cortical dynamics. Philos Trans R Soc B Biol Sci 360:709–731
64. Freeman J, Simoncelli EP (2011) Metamers of the ventral stream. Nat Neurosci 14:1195–1201
65. Lee S et al (2013) A new method for estimating population receptive field topography in visual cortex. Neuroimage 81:144–157
66. Amano K et al (2009) Visual field maps, population receptive field sizes, and visual field coverage in the human MT + complex. J Neurophysiol 102:2704–2718
67. Malach R et al (1995) Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. Proc Natl Acad Sci 92:8135–8139
68. Nieder A, Dehaene S (2009) Representation of number in the brain. Annu Rev Neurosci 32:185–208
69. Schiller PH et al (1976) Quantitative studies of single-cell properties in monkey striate cortex. II. Orientation specificity and ocular dominance. J Neurophysiol 39:1320–1333
70. Schiller PH et al (1976) Quantitative studies of single-cell properties in monkey striate cortex. III. Spatial frequency. J Neurophysiol 39:1334–1351
71. Albright T (1984) Direction and orientation selectivity of neurons in visual area MT of the macaque. J. Neurophysiol 52(6):1106–1130
72. Rodman H, Albright T (1987) Coding of visual stimulus velocity in area MT of the macaque. Vision Res 27(12):2035–2048
73. Lennie P, Movshon JA (2005) Coding of color and form in the geniculostriate visual pathway. J Opt Soc Am A Opt Image Sci Vis 22:2013–2033
74. Desimone R et al (1984) Stimulus-selective properties of inferior temporal neurons in the macaque. J Neurosci 4:2051–2062
75. Luck SJ et al (1997) Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. J Neurophysiol 77:24–42
76. Moran J, Desimone R (1985) Selective attention gates visual processing in the extrastriate cortex. Science (80-.) 229:782–784
77. Reynolds JH et al (2000) Attention increases sensitivity of V4 neurons. Neuron 26:703–714
78. Treue S, Maunsell JHR (1996) Attentional modulation of visual motion processing in cortical areas MT and MST. Nature 382:539–541
79. Anton-Erxleben K, Carrasco M (2013) Attentional enhancement of spatial resolution: linking behavioural and neurophysiological evidence. Nat Rev Neurosci 14:188–200
80. Navalpakkam V, Itti L (2007) Search goal tunes visual features optimally. Neuron 53:605–617
81. Regan D, Beverley KI (1985) Postadaptation orientation discrimination. J Opt Soc Am A 2:147–155
82. Jazayeri M, Movshon JA (2006) Optimal representation of sensory information by neural populations. Nat Neurosci 9:690–696
83. Butts DA, Goldman MS (2006) Tuning curves, neuronal variability, and sensory coding. PLoS Biol 4:e92

84. Serences JT, Boynton GM (2007) Feature-based attentional modulations in the absence of direct visual stimulation. Neuron 55:301–312
85. Brown SD, Heathcote A (2008) The simplest complete model of choice response time: linear ballistic accumulation. Cogn Psychol 57:153–178
86. Ester EF et al (2009) Spatially global representations in human primary visual cortex during working memory maintenance. J Neurosci 29:15258–15265
87. Serences JT et al (2009) Stimulus-specific delay activity in human primary visual cortex. Psychol Sci 20:207–214
88. Christophel TB et al (2012) Decoding the contents of visual short-term memory from human visual and parietal cortex. J Neurosci 32:12983–12989
89. Emrich SM et al (2013) Distributed patterns of activity in sensory cortex reflect the precision of multiple items maintained in visual short-term memory. J Neurosci 33:6516–6523
90. Tootell RB et al (1998) The retinotopy of visual spatial attention. Neuron 21:1409–1422
91. Kastner S et al (1999) Increased activity in human visual cortex during directed attention in the absence of visual stimulation. Neuron 22:751–761
92. Regan D (1989) Human brain electrophysiology: evoked potentials and evoked magnetic fields in science and medicine, Elsevier, Michigan
93. Regan D, Regan MP (1987) Nonlinearity in human visual responses to two-dimensional patterns, and a limitation of fourier methods. Vision Res 27:2181–2183
94. Duncan KK et al (2010) Identifying spatially overlapping local cortical networks with MEG. Hum Brain Mapp 31:1003–1016
95. Kaneoke Y et al (2009) Visual motion direction is represented in population-level neural response as measured by magnetoencephalography. Neuroscience 160:676–687

Part III
How the Cognitive Neurosciences Inform Cognitive Models

Chapter 13
Distinguishing Between Models of Perceptual Decision Making

Jochen Ditterich

Abstract Mathematical models are a useful tool for gaining insight into mechanisms of decision making. However, like other scientific methods, its application is not without pitfalls. This chapter demonstrates that it can be difficult to distinguish between alternative models and it illustrates that a model-based approach benefits from the availability of a rich dataset that provides sufficient constraints. Ideally, the dataset is not only comprised of behavioral data, but also contains neural data that provide information about the internal processing. The chapter focuses on two examples taken from perceptual decision making. In one case, information about response time distributions is used to reject a model that is otherwise consistent with accuracy data and mean response times. In the other case, only the availability of neural data allows a distinction between two alternative models that are both consistent with the behavioral data.

13.1 Introduction

Mathematical models can be extremely helpful tools for interpreting or making sense of experimental data, but there are also pitfalls associated with their use. For example, a model might be consistent with certain aspects of an experimental dataset, but it might be inconsistent with other aspects. If those latter aspects had not been considered during the modeling process or, even worse, if those data had never been collected in the first place, one would never have realized the inappropriateness of the model. What also tends to happen quite frequently is that multiple models can account roughly equally well for a given dataset. Thus, finding a model that is consistent with a given dataset does not mean that it is the only one that could potentially explain the data. Also, one might have to collect additional data that are able to distinguish between models that otherwise seem indistinguishable.

J. Ditterich (✉)
Center for Neuroscience and Department of Neurobiology, Physiology and Behavior, University of California, 1544 Newton Ct, Davis, CA 95618, USA
e-mail: jditterich@ucdavis.edu

B. U. Forstmann, E.-J. Wagenmakers (eds.), *An Introduction to Model-Based Cognitive Neuroscience,* DOI 10.1007/978-1-4939-2236-9_13

To make things more concrete, in this chapter we will be looking at two examples taken from perceptual decision making. In the first one we will see a model that is able to account for certain aspects of the decision behavior, accuracy and mean response times, but that fails to explain another aspect of the behavioral data, the response time distributions. We will then consider a modified model that can capture all aspects of the behavioral dataset. In the second example we will see two models that can account equally well for another behavioral dataset. Only the availability of neurophysiological data allows rejecting one of them.

13.2 Distinguishing Between Time-Invariant and Time-Variant Mechanisms of Perceptual Decision Making Based on Response Time Distributions

The behavioral data that will be used in this section is taken from an experiment by Roitman and Shadlen [1]. Monkeys were trained to make a decision between leftward or rightward motion in a dynamic random dot display. While fixating a central spot on a computer screen, monkeys were watching a cloud of dots in a circular aperture around the fixation point. On a given trial, a certain fraction of these dots were translated coherently either leftward or rightward, whereas the remaining dots flickered randomly. The percentage of coherently moving dots, the stimulus strength or motion coherence, was varied randomly from trial to trial. The monkeys had to identify the direction of net motion in the visual stimulus and to make a two-alternative forced choice between leftward or rightward. The monkeys were allowed to watch the stimulus as long as desired and provided an answer by making a goal-directed eye movement to one of two choice targets that were presented to the left and to the right of the motion stimulus. On each trial, the monkey's choice and the associated response time (RT), the time between appearance of the motion stimulus and initiation of an eye movement to a choice target, were recorded and a fluid reward was given for a correct choice. The task is illustrated in Fig. 13.1a.

When confronted with this task, both humans and monkeys show a behavioral pattern where both accuracy and mean RT vary systematically with stimulus strength [1, 2]. The symbols in Fig. 13.1b show the monkeys' accuracy and the mean RT associated with correct choices as a function of motion coherence: the more useful information is provided by the sensory stimulus (higher coherence), the more accurate and, on average, faster the decisions are.

How did the monkeys make a decision between left and right? A major advantage of using a task like random-dot motion discrimination is the availability of knowledge about how the sensory evidence is represented in the brain. Britten et al. [3] have recorded from motion-sensitive neurons in the middle temporal area (MT), which have been demonstrated to carry the sensory evidence that is used to make the decision [4, 5]. These neurons are direction-selective and increase their firing rate roughly linearly when motion in the preferred direction is presented and coherence is increased. Likewise, the firing rate decreases roughly linearly when motion in the

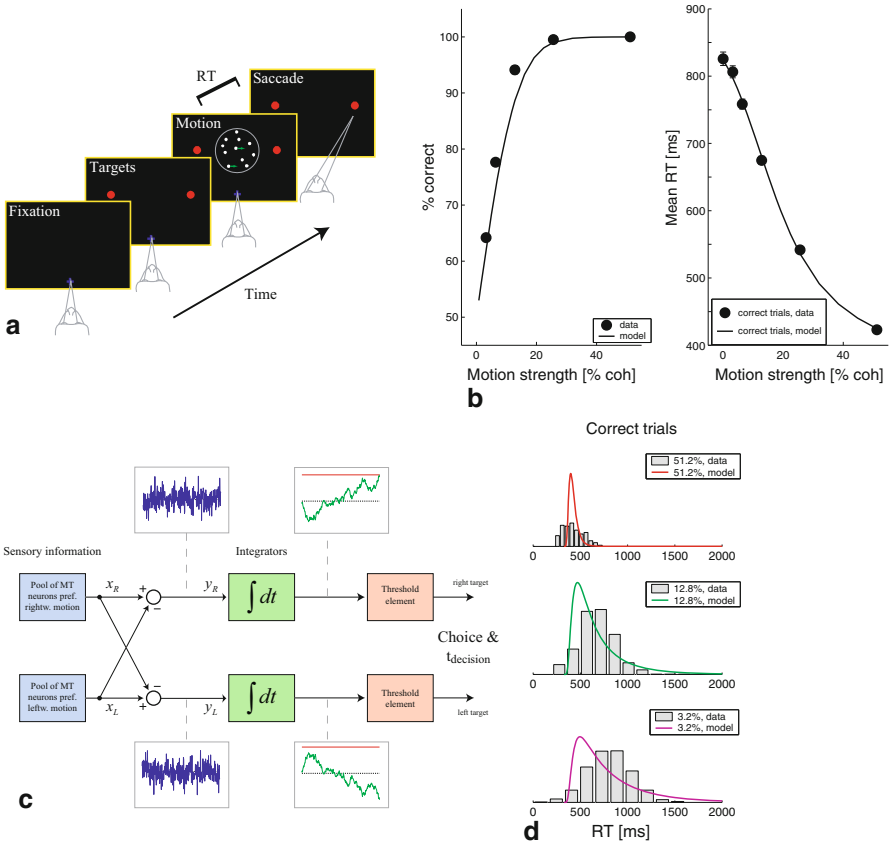


Fig. 13.1 Time-invariant decision mechanism (adapted from [6]). **a** Random-dot motion direction discrimination task. **b** Accuracy and mean RT of correct responses. The symbols represent the data, the lines the best-fitting model of the type shown in **c**. **c** Time-invariant integration-to-threshold decision mechanism. **d** Mismatch between predicted (*colored lines*) and actual RT distributions (*gray histograms*)

opposite direction is presented and coherence is increased. These neurons show a robust response when a non-coherent motion stimulus is presented (pure noise, no net motion), which fluctuates substantially over time. Two task-relevant pools of such neurons, one tuned to rightward motion, the other one tuned to leftward motion, are represented by the blue boxes in Fig. 13.1c. The difference between the activities of these two pools ($y_R = x_R - x_L$ in the figure) would be a particularly informative signal for making the decision: it is zero on average for a pure noise stimulus, positive for a stimulus with net rightward motion, and negative for a stimulus with net leftward motion. Furthermore, the absolute value of the signal is expected to scale linearly with motion coherence. However, the signal would still fluctuate considerably over time (as indicated by the blue traces in the figure). A reasonable way to deal

with these fluctuations, at least in the case of a stationary signal that does not change its properties over time, would be to integrate the signal over time, which improves the signal-to-noise ratio as time progresses. This integration process (for both net sensory evidence signals y_R and y_L) is indicated by the green boxes in Fig. 13.1c.

Making a decision requires solving two problems: which option to choose and when to make the choice. The assumption in the proposed model is that the decision process terminates as soon as one of the accumulated net evidence signals reaches a fixed decision threshold (indicated by the reddish boxes in the figure). Since the decision time cannot directly be measured in the experiment, we further make the assumption that the measured RT is the sum of the decision time and a constant processing time associated with initial processing of the visual information and saccade preparation ("residual time"). Thus, we assume that decision making, visual processing, and motor preparation are independent processes and that trial-by-trial RT variations are dominated by the variability of the decision time. Further model details can be found in [6]. Note that this model is mathematically equivalent to the drift-diffusion model that has been proposed to account for a variety of perceptual decision-making datasets [7]. Since $y_L = -y_R$ and both integrators are perfect, the outputs of the two integrators have the same absolute value, but opposite sign. Comparing each of these outputs to the same fixed decision criterion θ is therefore equivalent to comparing the output of just one integrator, reflecting the current state of the drift-diffusion process, with two decision boundaries that are located at $+\theta$ and $-\theta$.

We can now ask how well such a model can account for the observed decision behavior. We can estimate the model parameters through an optimization process, in this case, for example, by minimizing the sum of squared deviations between the predicted and actual mean RTs. The right panel in Fig. 13.1b shows that with three free model parameters (after arbitrarily fixing the decision threshold at one, the proportionality factor between motion coherence and the mean of y_R, the variance of y_R, and the residual time need to be determined) a perfect match between model and data mean RTs can be achieved. Furthermore, as shown in the left panel, the model can also account for the psychometric function quite well.

A different picture arises when the trial-by-trial variability of RTs is taken into consideration. Figure 13.1d shows a comparison between the actual RT distributions (gray histograms) and the ones predicted by the model (colored lines). These distributions have obviously quite different shapes: the RT distributions in the dataset are almost symmetric, whereas the ones that are predicted by the model are very asymmetric. By following Exercise 1 in Sect. 5, a programming exercise, the reader can easily convince himself that decision time distributions with exponential tails are a general property of sequential sampling mechanisms with a fixed decision criterion (at least when the samples are i.i.d.), regardless of how much temporal integration is involved. The data are therefore obviously inconsistent with such a mechanism. A possible way out of the dilemma is the consideration of time-variant decision mechanisms. In particular, we will be looking at a decision mechanism where the decision criterion changes as a function of the time that has passed since the beginning of the current decision trial: the decision threshold drops as time progresses (as shown

in Fig. 13.2a). As a consequence, long decision times become less frequent and the exponential tails of the distributions disappear. As can be seen from Fig. 13.2b, such a model is able to produce RT distributions that match the shape of the observed distributions. Furthermore, the model is still able to account for accuracy and mean RTs (Fig. 13.2c) and can even explain why error RTs are expected to be longer than RTs associated with correct responses, which is not predicted by the basic drift-diffusion model (but see [8] for a different modification). Further details regarding the model and the fitting procedure can again be found in [6].

Why might a decision mechanism take advantage of a time-variant property like a decision criterion that changes as a function of how much time has already been spent on collecting evidence? It has been pointed out in the literature [9] that a mechanism of the type shown in Fig. 13.1c would be a good approximation of the Sequential Probability Ratio Test (SPRT), an algorithm that is optimal in the sense that it minimizes the mean decision time for a given desired level of accuracy [10]. The key argument is based on the observation that the net sensory evidence (y_R or y_L) is roughly proportional to a log likelihood ratio (see [9] for details). However, the proportionality factor changes with motion strength. If all trials had the same motion coherence, a fixed decision criterion could be set such that a particular accuracy level is achieved and the cumulative decision time would be minimized. The experiment, however, is typically performed with a random mix of motion coherences. In this situation, as has been shown in Fig. 13.1b, subjects' accuracy changes considerably with motion strength, meaning that they keep integrating the difference between x_R and x_L to a criterion and do not adjust the proportionality factor based on the difficulty level of a given trial. I have demonstrated in [11] that, under these circumstances, a time-variant decision mechanism can provide a larger reward rate than an otherwise comparable mechanism that does not change its properties over time. In the meanwhile it has been shown that it is indeed a decision mechanism with a criterion that is lowered as time passes that maximizes reward rate [12]. Intuitively, the more time has already been spent on a decision without having reached threshold, the more likely it is that one is dealing with a low-coherence trial. As has been demonstrated in [13], the closer the mean of the net sensory evidence signal is to zero ("drift rate" in the context of the drift diffusion model), the lower the decision criterion has to be to maximize reward rate (in the regime of small drift rates).

In this example, evaluating the RT distributions was essential for noticing that a model that would otherwise have been compatible with accuracy and mean RTs of the decisions was really inconsistent with the experimental data. Let us now have a look at a second example. In this case, we will see two models that are both consistent with the observed decision behavior (including RT distributions) and it is only through the availability of additional neural data that one of the models can be ruled out.

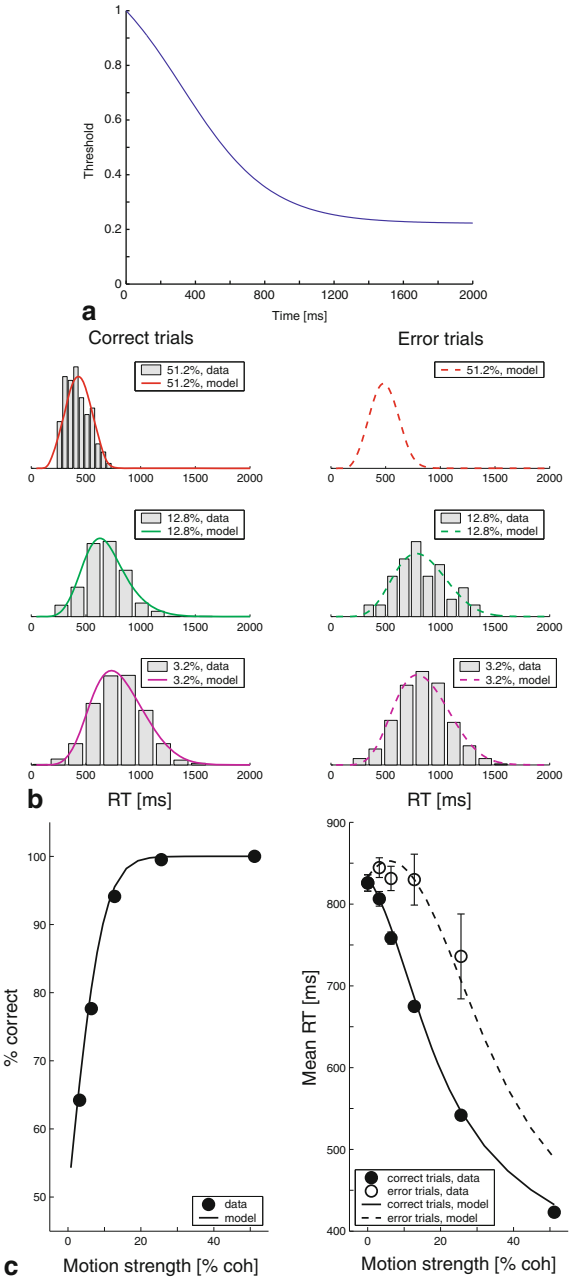


Fig. 13.2 Time-variant decision mechanism (adapted from [6]). **a** Time-variant decision threshold. **b** The time-variant decision mechanism predicts RT distributions (*colored lines*) that match the shape of the measured RT distributions (*gray histograms*). **c** Psychometric and chronometric functions. The symbols represent the data, the lines the best-fitting model with a time-variant decision criterion. The model can also account for the, on average, longer RTs on error trials (open symbols and *dashed line*) compared to correct responses (filled symbols and *solid line*)

13.3 Distinguishing Between Feedforward and Feedback Inhibition Mechanisms Based on Neurophysiological Data

We will again be looking at data from a random-dot motion direction discrimination task. This time, however, the subjects had to make a judgment about a stimulus that had coherent motion in three different directions at the same time. The task was to identify the direction of the strongest motion component in the stimulus [14]. The advantage over the basic version of the task, as discussed above, is that the experimenter has control over how much sensory evidence is provided for and against each of the potential choices. The task is illustrated in Fig. 13.3a. Let us first have a look at human behavioral data and potential models that could explain the observed decision behavior. Later we will be looking at neural data that have been recorded from monkeys performing the same task.

We had already introduced a mechanism for perceptual decisions between two alternatives in Fig. 13.1c. Figure 13.3b shows a generalization of such a mechanism for more than two alternatives. Each choice alternative is associated with an independent integrator (yellow boxes). Whichever integrator reaches a fixed decision criterion first (orange boxes) determines choice and decision time. Pools of sensory neurons (red, green, and blue boxes) provide evidence in favor of a particular choice (feedforward excitation; solid colored arrows) as well as evidence against the other alternatives (feedforward inhibition; dashed colored arrows). Such a mechanism can account for the distribution of choices (Fig. 13.3c), mean RT (Fig. 13.3d), and RT distributions (Fig. 13.3e). Since each stimulus is characterized by a set of three coherence values, the data are plotted such that the x-coordinate reflects the coherence of the strongest motion component and the color codes for the coherences of the other two motion components. The different shapes in Fig. 13.3c code for choices of the target associated with the strongest motion component (circles; correct choices) and errors (squares: intermediate component; diamonds: weakest component). Further details regarding the model and the model fit can be found in [15].

Many decision mechanisms that have been proposed in the literature have an architecture that is quite different from the one shown in Fig. 13.3b: the integrators are not independent, but compete with each other through lateral (or feedback) inhibition [16, 17]. For perceptual decisions between two alternatives, it has been pointed out in the literature that these different types of decision mechanisms (feedforward vs. feedback inhibition) can produce virtually indistinguishable decision behavior [13]. This also turns out to be the case in our 3-choice experiment. Figure 13.4a shows a decision mechanism that is based on feedback inhibition. The integrators compete with each other through feedback inhibition (dashed arrows). The sensory pools (red, green, and blue boxes) provide only excitatory input to one integrator each. As can be seen in Figs. 13.4, such a model can account for the behavioral data just as well as the mechanism that was based on feedforward inhibition and that was shown in Fig. 13.3. Further details of the model can again be found in [15].

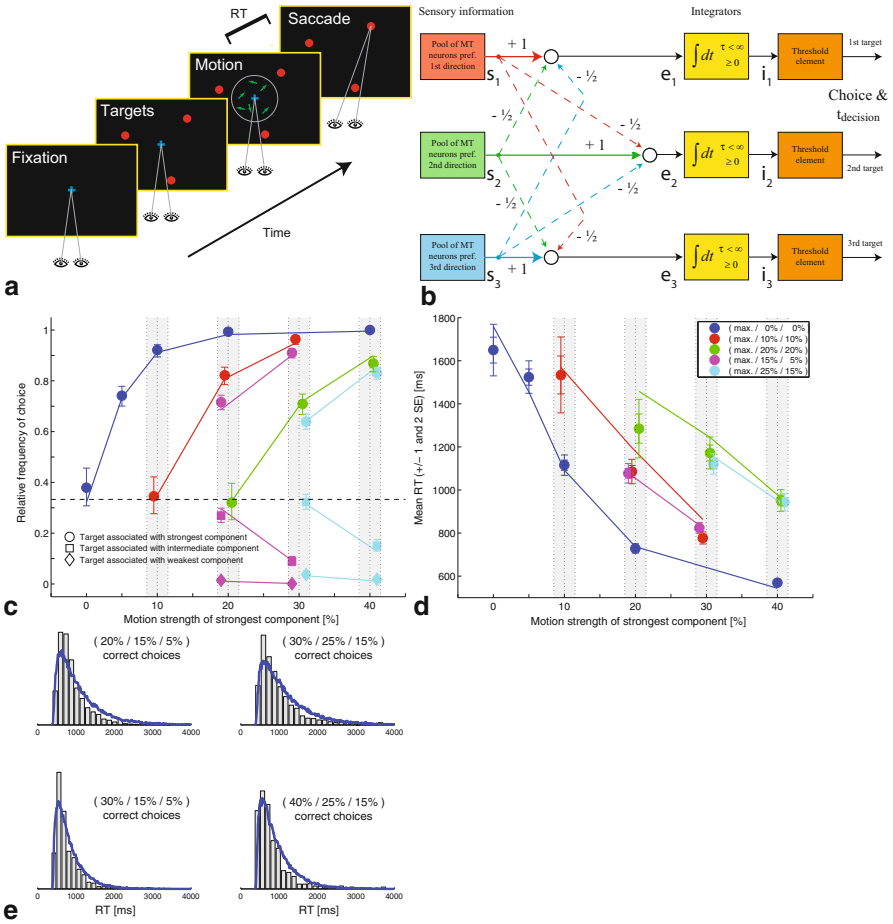


Fig. 13.3 Feedforward inhibition decision mechanism (adapted from [21] and [15]). **a** 3-choice random-dot motion direction discrimination task, using a 3-component stimulus. The subject has to identify the direction of the strongest motion component. **b** Decision mechanism based on feedforward inhibition. **c** Distribution of choices. The symbols represent the data, the lines the predictions of the best-fitting model of the type shown in **b**. The x-coordinate reflects the coherence of the strongest motion component, the other two coherences are coded for by the color. The different shapes are associated with choosing the targets associated with the strongest (*circles*), intermediate (*squares*), or weakest (*diamonds*) motion component. **d** Mean RT. The symbols represent the data, the lines the best-fitting model of the type shown in **b**. **e** RT distributions. The *gray histograms* represent the data, the *blue lines* the predictions of the best-fitting model of the type shown in **b**

Although both mechanisms produce virtually identical decision behavior, their inner workings are quite different. In the case of the feedforward inhibition mechanism, sensory evidence for a particular direction pushes one of the integrators towards its decision threshold and the other integrators away from their decision thresholds throughout the decision. In the case of the feedback inhibition mechanism, sensory

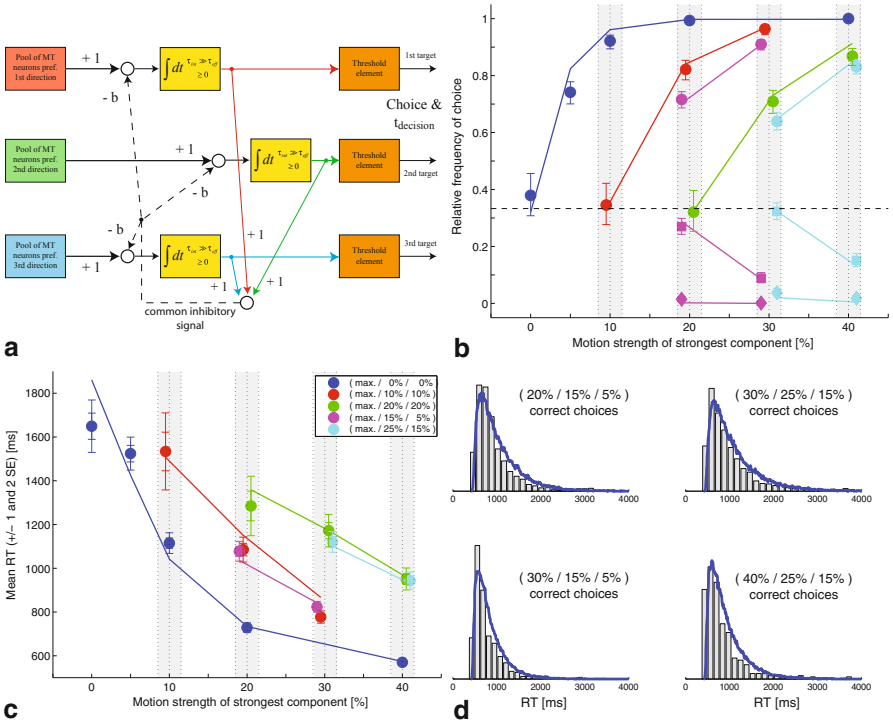


Fig. 13.4 Feedback inhibition decision mechanism (adapted from [15]). **a** Decision mechanism based on feedback inhibition. **b** Distribution of choices. The symbols represent the data, the lines the predictions of the best-fitting model of the type shown in **a**. **c** Mean RT. The symbols represent the data, the lines the best-fitting model of the type shown in **a**. **d** RT distributions. The *gray histograms* represent the data, the *blue lines* the predictions of the best-fitting model of the type shown in **a**

evidence pushes only a particular integrator towards its decision threshold, and it is the buildup of activity in a particular integrator that suppresses activity in the competing integrators. Our goal was to reveal the inner workings of the decision mechanism by training monkeys to perform the same decision task and by recording from neurons in the brain that have previously been shown to carry decision-related activity. When monkeys perform a random-dot motion direction discrimination task and report the choice with a goal-directed eye movement, the activity of neurons in the lateral intraparietal area (LIP) in parietal cortex reflects the ongoing decision [1, 18, 19] and has even been shown to have a causal effect on the outcome of the decision [20]. The response of this type of neurons in our 3-choice task is shown in Fig. 13.5a. The plot shows the average firing rate as a function of time, aligned with the onset time of the motion stimulus (dashed line). The recordings are performed by placing one of the choice targets inside the response field (RF) of the recorded

neuron. The coherence of the motion component that moves towards this target defines the strength of the sensory evidence for choosing this target (pro evidence) and the average coherence of the other two motion components defines the strength of the evidence against choosing this target (anti evidence). For the purpose of this plot, all stimuli were grouped into four different categories according to whether the pro evidence was high or low and whether the anti evidence was high or low (different colors in Fig. 13.5a). As can be seen, approx. 200 ms after motion stimulus onset, the decision-related LIP neurons show a ramping response whose slope depends on the sensory evidence. Figure 13.5b further illustrates that the slope in the shaded area in Fig. 13.5a is roughly a linear function of the difference between the pro and the anti evidence. Further details regarding the experiment and the data analysis can be found in [21].

The particularly interesting analysis of the neural data for the purpose of telling the difference between decision mechanisms that are based on feedforward or feedback inhibition is shown in the remaining panels of Fig. 13.5. We isolated the effect of the pro evidence on the neural response by performing a trial selection such that we had two groups of trials with a large difference in pro evidence, but identical distributions of anti evidence. (Details of the method can be found in [21].) The result is shown in Fig. 13.5c with the red trace representing trials with high pro evidence and the blue trace representing trials with low pro evidence. Figure 13.5e shows an even cleaner version of the difference between the two traces by first subtracting the common signal before averaging across recorded neurons. The same principle was used to isolate the effect of the anti evidence on the neural response and the result is shown in Fig. 13.5d and 13.5f. In this case, the red trace represents trials with high anti evidence and the blue trace trials with low anti evidence. Overall, it can be seen that the pro evidence has an excitatory effect and that the anti evidence has an inhibitory effect. The triangles and shaded areas in Figure 13.5e and 13.5f indicate the continuous time periods during which the pro and anti evidence had a significant effect on the neural response. Interestingly, the inhibitory effect of the anti evidence manifests itself substantially earlier (approx. 200 ms into the trial; Fig. 13.5f) than the excitatory effect of the pro evidence (approx. 280 ms into the trial; Fig. 13.5e). This observation has important consequences for constraining the structure of the underlying decision mechanism. The observed inhibition cannot solely rely on a feedback mechanism, in which case it would have to follow a change in neural activity that is brought about by the excitatory effect of the pro evidence. In the recorded pool of decision-related neurons, the inhibitory effect of the anti evidence actually precedes the excitatory effect of the pro evidence. This clearly indicates the presence of feedforward inhibition and rules out a mechanism of the type shown in Fig. 13.4a that only relies on feedback inhibition. However, we cannot rule out the possibility that the actual decision mechanism takes advantage of a mixture of feedforward and feedback inhibition mechanisms and might therefore be a hybrid of the classes of mechanisms shown in Fig. 13.3b and Fig. 13.4a. We have some experimental evidence that this might indeed be the case, but a more in-depth analysis is required before final conclusions can be drawn.

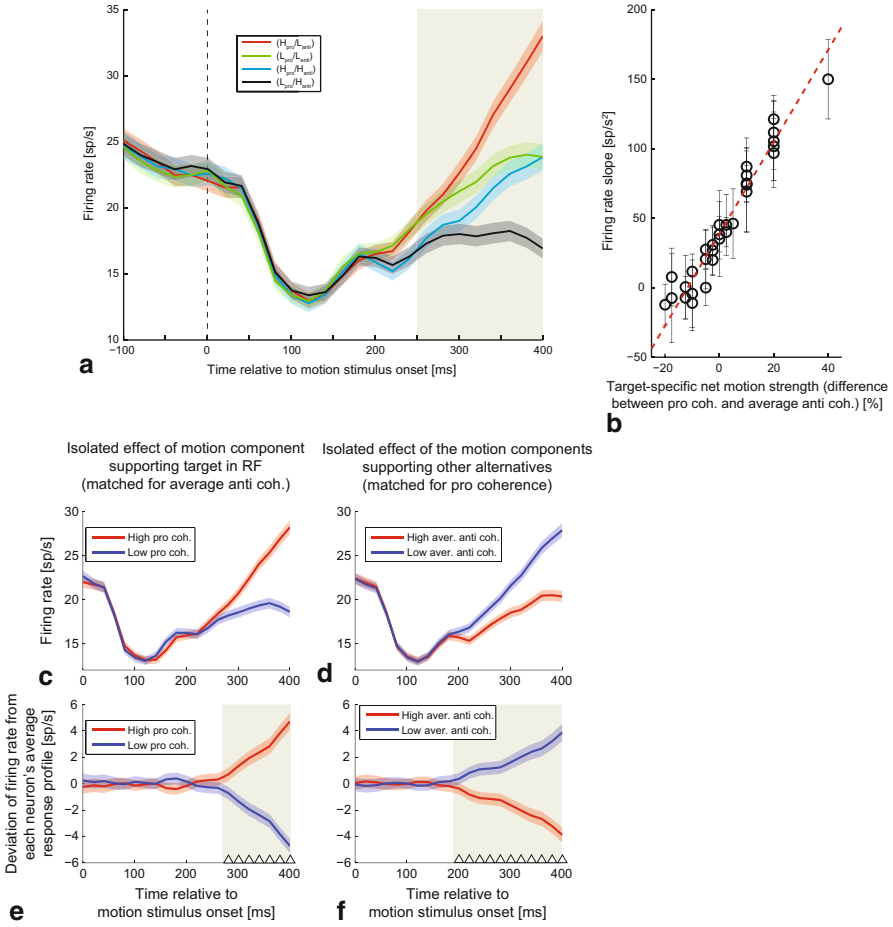


Fig. 13.5 Neural activity in parietal cortex (area LIP) during the 3-choice task (adapted from [21]). **a** Average firing rate, time-locked to motion stimulus onset, as a function of the strength of both pro and anti sensory evidence for the choice associated with the target inside the recorded cell's RF. The slope in the shaded area is steepest for trials with high pro evidence and low anti evidence and shallowest for trials with low pro evidence and high anti evidence. **b** The firing rate slope in the shaded area in a is roughly a linear function of net motion strength (the difference between the coherence of the motion component that supports choosing the target in the RF and the average coherence of the other two motion components). **c** Isolated effect of the strength of pro evidence on neural activity. More than 250 ms into the stimulus, the firing rate is higher for strong pro evidence (*red*) compared to weak pro evidence (*blue*). **d** Isolated effect of the strength of anti evidence on neural activity. The firing rate is higher for weak anti evidence (*blue*) compared to strong anti evidence (*red*). **e** Deviation from average response profile due to strength of pro evidence. A significant effect is observed 280 ms into the trial. **f** Deviation from average response profile due to strength of anti evidence. A significant effect is observed 200 ms into the trial

13.4 Conclusion

We have seen two examples where it was difficult or even impossible to distinguish between different decision mechanisms based on particular datasets. In one case, a model seemed consistent with choice accuracy and mean RTs and only taking the shape of RT distributions into account allowed rejecting the model. In the other case, two models were able to explain choice accuracy, mean RTs, and even RT distributions and only the availability of neural data allowed rejecting one of them. More generally, rejecting a model based on available data is relatively straightforward when the model cannot account for the observed data. It is much more difficult to make the claim that a model accurately describes the actual mechanism when it is able to account for observed data. Oftentimes it would be possible to find alternative models that could account for the observed data just as well. Thus, the more constraining data one has access to, ideally data from different domains like, for example, behavior and physiology, the easier it is to discriminate between alternative models.

Exercises

The following exercises are simulation exercises to make the reader familiar with properties of decision mechanisms based on sequential sampling. Readers with some programming experience are strongly encouraged to write their own programs for solving the exercises, but we also provide scripts for MATLAB and the freely available alternative Scilab. These scripts can be downloaded from http://www.peractionlab.org/supmat/2.

1. In this first exercise we want to develop a feeling for how decision times are distributed when decisions are made based on comparing sequentially sampled data to decision criteria, both when integrating the incoming evidence over time and without doing so. Let's assume that we observe an incoming stream of random numbers (our "sensory" evidence) and we want to decide whether these numbers have been drawn from a distribution with a positive or a negative mean (the two alternatives to choose from). Let's say that the numbers are either drawn from a normal distribution with a mean of $+1$ and a standard deviation of 3 or from a normal distribution with a mean of -1 and the same standard deviation. These distributions are highly overlapping, which makes it a difficult decision problem. Let's look at temporal integration first, which makes it a discrete approximation of a drift-diffusion process. We keep adding the incoming numbers until the sum is either larger than a positive decision criterion A, in which case we decide in favor of a positive mean, or until the sum is smaller than $-$A, in which case we decide in favor of a negative mean. Since it is a symmetric problem, we only have to simulate sampling from one of the distributions (for example, the one with a mean of $+1$). Thus, crossing the $+$A threshold corresponds to correct decisions and crossing the $-$A threshold corresponds to errors.

2. Let's set A to 25 and simulate 20,000 decision trials. For each trial, record how many random numbers had to be summed until one of the decision thresholds was crossed, which corresponds to the decision time on that trial, assuming that the evidence samples arrive at a constant rate. Plot the distribution of decision times. What does the distribution look like? You should notice the asymmetric shape with a long exponential tail.
3. Let's now do the same for a decision mechanism that does not take advantage of temporal integration. Each individual sample is tested whether it exceeds a positive threshold B, in which case we decide in favor of a positive mean, or whether it is smaller than $-$ B, in which case we decide in favor of a negative mean. Let's set B to 7 and simulate again 20,000 decision trials. What does the distribution look like? What do both distributions have in common? What is the difference between the shapes of the two distributions?
4. So far we have only looked at the distributions of decision times. Let's get accuracy into play to see what the advantage of the decision mechanism with temporal integration is. Return to your first simulation and, in addition to keeping track of the decision times, keep also track of how many decisions are correct. How does the decision threshold have to be chosen (somewhere between 5 and 20) to obtain an accuracy of approximately 90 %? How many samples have to be evaluated on average to achieve this accuracy?
5. Let us now return to the simulation of the decision mechanism without temporal integration. How does the decision threshold have to be chosen (somewhere between 5 and 10) to obtain an accuracy of approximately 90 %? How many samples have to be evaluated on average in this case to achieve such accuracy? How does this compare to the result for the mechanism with temporal integration? What conclusion would you draw from this observation?

Further Reading

Deco G, Rolls ET, Albantakis L, Romo R (2013) Brain mechanisms for perceptual and reward-related decision-making. Prog Neurobiol 103:194–213

Tsetsos K, Gao J, McClelland JL, Usher M (2012) Using Time-Varying Evidence to Test Models of Decision Dynamics: Bounded Diffusion vs. the Leaky Competing Accumulator Model. Frontiers in Neuroscience 6:79

Tsetsos K, Usher M, McClelland JL (2011) Testing multi-alternative decision models with non-stationary evidence. Frontiers in Neuroscience 5:63

References

1. Roitman JD, Shadlen MN (2002) Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. J Neurosci 22(21):9475–9489
2. Palmer J, Huk AC, Shadlen MN (2005) The effect of stimulus strength on the speed and accuracy of a perceptual decision. J Vis 5(5):376–404. doi:10:1167/5.5.1

3. Britten KH, Shadlen MN, Newsome WT, Movshon JA (1993) Responses of neurons in macaque MT to stochastic motion signals. Vis Neurosci 10(6):1157–1169
4. Ditterich J, Mazurek ME, Shadlen MN (2003) Microstimulation of visual cortex affects the speed of perceptual decisions. Nat Neurosci 6(8):891–898. doi:10.1038/nn1094
5. Salzman CD, Murasugi CM, Britten KH, Newsome WT (1992) Microstimulation in visual area MT: effects on direction discrimination performance. J Neurosci 12(6):2331–2355
6. Ditterich J (2006) Stochastic models of decisions about motion direction: behavior and physiology. Neural Netw 19(8):981–1012. doi:10.1016/j.neunet.2006.05.042
7. Ratcliff R, Smith PL (2004) A comparison of sequential sampling models for two-choice reaction time. Psychol Rev 111(2):333–367. doi:10.1037/0033–295X.111.2.333
8. Ratcliff R, Rouder JN (1998) Modeling response times for two-choice decisions. Psychol Sci 9(5):347–356. doi:10.1111/1467–9280.00067
9. Gold JI, Shadlen MN (2001) Neural computations that underlie decisions about sensory stimuli. Trends Cogn Sci 5(1):10–16
10. Wald A (1945) Sequential tests of statistical hypotheses. Ann Math Stat 16(2):117–186. doi:10.1214/aoms/1177731118
11. Ditterich J (2006) Evidence for time-variant decision making. Eur J Neurosci 24(12):3628–3641. doi:10.1111/j.1460–9568.2006.05221.x
12. Drugowitsch J, Moreno-Bote R, Churchland AK, Shadlen MN, Pouget A (2012) The cost of accumulating evidence in perceptual decision making. J Neurosci 32(11):3612–3628. doi:10.1523/JNEUROSCI.4010–11.2012
13. Bogacz R, Brown E, Moehlis J, Holmes P, Cohen JD (2006) The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. Psychol Rev 113(4):700–765. doi:10.1037/0033–295X.113.4.700
14. Niwa M, Ditterich J (2008) Perceptual decisions between multiple directions of visual motion. J Neurosci 28(17):4435–4445. doi:10.1523/JNEUROSCI.5564–07.2008
15. Ditterich J (2010) A comparison between mechanisms of multi-alternative perceptual decision making: ability to explain human behavior, predictions for neurophysiology, and relationship with decision theory. Front Neurosci 4:184. doi:10.3389/fnins.2010.00184
16. Usher M, McClelland JL (2001) The time course of perceptual choice: the leaky, competing accumulator model. Psychol Rev 108(3):550–592
17. Wang XJ (2002) Probabilistic decision making by slow reverberation in cortical circuits. Neuron 36(5):955–968
18. Bollimunta A, Totten D, Ditterich J (2012) Neural dynamics of choice: single-trial analysis of decision-related activity in parietal cortex. J Neurosci 32(37):12684–12701. doi:10.1523/JNEUROSCI.5752–11.2012
19. Shadlen MN, Newsome WT (2001) Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. J Neurophysiol 86(4):1916–1936
20. Hanks TD, Ditterich J, Shadlen MN (2006) Microstimulation of macaque area LIP affects decision-making in a motion discrimination task. Nat Neurosci 9(5):682–689. doi:10.1038/nn1683
21. Bollimunta A, Ditterich J (2012) Local computation of decision-relevant net sensory evidence in parietal cortex. Cereb Cortex 22(4):903–917. doi:10.1093/cercor/bhr165

Chapter 14
Optimal Decision Making in the Cortico-Basal-Ganglia Circuit

Rafal Bogacz

Abstract This chapter presents a model assuming that during decision making the cortico-basal-ganglia circuit computes probabilities that considered alternatives are correct, according to Bayes' theorem. The model suggests how the equation of Bayes' theorem is mapped onto the functional anatomy of a circuit involving the cortex, basal ganglia and thalamus. The chapter also describes the relationship of the model to other models of decision making and experimental data.

14.1 Introduction

The basal ganglia are a set of subcortical nuclei that are critically important for action selection [1]. The connectivity and response properties of different neuronal populations in the basal ganglia have been characterized in many studies [2, 3]. These rich experimental data allowed development of computational models describing how different functions of basal ganglia are implemented in their circuitry [4–9]. For example, the basal ganglia are important for learning the expected rewards associated with selecting different actions in different contexts, and Chapter 8 by Frank described how this reinforcement learning is achieved in the neural substrate. This chapter presents a model suggesting that during action selection on the basis of noisy sensory information, the cortico-basal-ganglia circuit approximates a statistically optimal decision making procedure [4, 10].

Before describing the model, let us start with an example of a very simple decision task involving noisy sensory information. Consider a rat that has to press a left or a right lever on the basis of an auditory stimulus. The auditory stimulus consists of a sequence of short intervals (e.g. 100 ms) during which a low or high tone is presented. On trials on which pressing the left lever is rewarded, the low tone has 70 % chance of occurring in each interval, while the high tone has only 30 % probability of occurring. Conversely, on trials with the right lever being rewarded, the high and low tones have

R. Bogacz (✉)
Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford OX3 9DU, UK
e-mail: rafal.bogacz@ndcn.ox.ac.uk

B. U. Forstmann, E.-J. Wagenmakers (eds.), *An Introduction to Model-Based Cognitive Neuroscience,* DOI 10.1007/978-1-4939-2236-9_14

70 and 30 % probabilities, respectively. Please note that in this task, to maximize its reward, the rat needs to listen to the stimulus, accumulate information from successive beeps, and only make a choice once it reaches a certain level of confidence. The model presented in this Chapter proposes how the computations required to conduct such decision process are implemented in the cortico-basal-ganglia circuit.

14.2 Model

While presenting the model in this section we will follow the three levels of analysis proposed by Marr [11]. He suggested that when developing a computational model for any system within a brain, it is useful to progress through three levels of its description:

- Computational level—describing **what** computations the system performs, and why such computation is appropriate for achieving a particular function.
- Algorithmic and representation level—describing **how** the computation defined above is performed, i.e. how the inputs and outputs are represented, and what algorithm transforms the input representation into the output representation.
- Implementation level—describing how the algorithm described above is physically **implemented** in neural hardware.

The following three subsections provide the description of the model on the above three levels of analysis.

14.2.1 Computational Level

The description of the computation the model performs is very simple. Let us denote the actions available in a given context by A_i, thus in the example in the Introduction, the rat has two potentially rewarded actions A_1 and A_2 corresponding to pressing the left and the right lever respectively. The model proposes that during action selection the circuit is computing for each actions A_i the probability of this action being appropriate. Let us denote this probability by $P(A_i)$. Thus in our example, after each beep the probabilities of the two actions are updated according to the stimulus. Finally, the model suggests that whenever for any action its probability exceeds a threshold of confidence, this action is initiated.

Let us now discuss in what sense the above decision procedure is optimal. The above procedure is known as the Multihypothesis Sequential Probability Ratio Test (MSPRT) [12]. It is a generalization to multiple alternatives of the Sequential Probability Ratio Test (SPRT) developed earlier for the choice between two alternatives [13]. To understand the optimality properties of these procedures we need to first note that any decision procedure on the basis of sequentially sampled noisy information exhibits the speed-accuracy tradeoff. Namely, if the decisions are made quickly,

e.g. by lowering the threshold of confidence required to trigger choice, they are less accurate. Conversely, if a higher accuracy is required, the decisions need to take longer. The SPRT is optimal in a sense that for any required level of accuracy, it achieves the fastest average decision time [14]. For the MSPRT the same optimality property has been proven analytically in a limit of accuracy close to 100 % [15], and simulations indicate that for lower accuracy levels, the MSPRT makes choices faster or equally fast than other decision procedures [16]. The optimality property of the SPRT/MSPRT is ecologically relevant, because making fast decisions allows increasing the rate of receiving rewards. In particular, it has been shown that in a wide range of tasks the reward rate is maximized when the animal employs a procedure that minimizes decision time for a given accuracy, and the threshold parameter is set appropriately [17].

14.2.2 Algorithmic and Representation Level

In this section we first describe how the probabilities of actions are computed on the basis of sensory input, and then how the probabilities are represented in the model. Let us assume for simplicity that time is divided into discrete intervals. Let us denote the sensory input provided in the current time step by S. The sensory input can influence animal's estimates of probability of each action A_i, because from past experience, the animal could have learnt how often stimulus S occurred on trials when action A_i was appropriate. Let us denote this rate of occurrence by $P(S|A_i)$. Thus for example, in the task described in the Introduction, if we denote pressing the left and right levers by A_1 and A_2, while hearing the low and high tones by S_1 and S_2 respectively, the animal could have learnt from previous trials that $P(S_1|A_1) = 0.7$, $P(S_1|A_2) = 0.3$, etc. The computation required to update probabilities of actions on the basis of sensory input is described by Bayes' theorem.

$$P(A_i|S) = \frac{P(A_i)P(S|A_i)}{P(S)} \tag{14.1}$$

Bayes' theorem has been discussed in Chap. 9 by Mars and O'Reilly. Bayes' theorem simply says that in order to compute the updated or posterior probability of action $P(A_i|S)$, one needs to multiply the previous or prior probability $P(A_i)$ by the probability of the sensory input S appearing on trials on which action A_i is correct, denoted $P(S|A_i)$, which could have been learnt from experience. Additionally, to ensure that the posterior probabilities add up to 1, this product is divided by a normalization term $P(S)$ equal to the sum of corresponding products across all N actions:

$$P(S) = \sum_{i=1}^{N} P(A_i)P(S|A_i) \tag{14.2}$$

Figure 14.1a represents the equation of Bayes' theorem as a network of computational elements. For simplicity the diagram is shown just for two actions. The computed

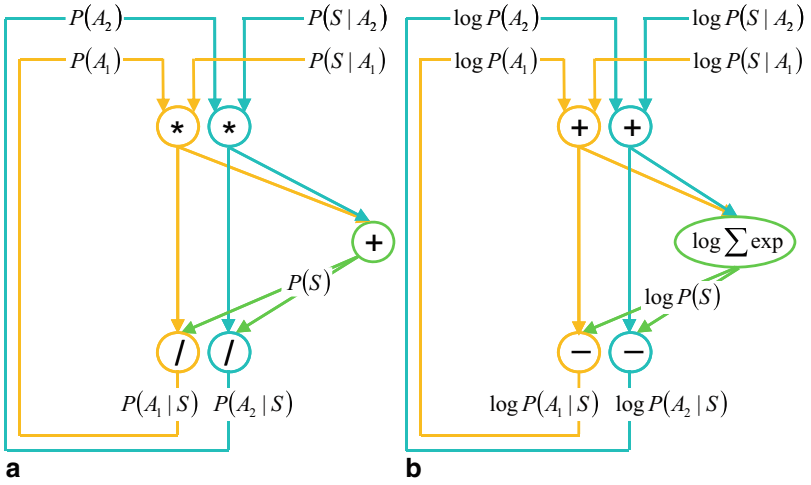


Fig. 14.1 Representation of the computation of the probabilities of actions (panel a) and their logarithms (panel b). *Circles* denote mathematical operations and *arrows* denote flow of information. *Yellow* and *blue* pathways show computations of the probabilities for two sample actions, while *green* pathways compute the normalization

posterior probabilities become the basis of the computation for the next step, hence they need to be fed back with a time delay. To simplify notation, we will assign the posterior probabilities $P(A_i|S)$ as the prior probabilities $P(A_i)$ for the next iteration. For readers who have not used Bayes' theorem before, we recommend doing now Exercise 1 listed at the end of the Chapter, to gain an intuition for how Bayes' theorem updates probabilities of actions on the basis of a sequence of stimuli in the task described in the Introduction.

The way the probabilities are represented in the model affects how easy it is for neurons to implement the above update of probabilities. Equation 1 includes multiplication and division, that are not natural operations for neurons, but this problem can be solved by taking logarithm. Recall that the logarithm has the following properties: $\log a \cdot b = \log a + \log b$, and $\log a/b = \log a - \log b$. Hence taking the logarithm of both sides of Eq. 1 we get:

$$\log P(A_i|S) = \log P(A_i) + \log P(S|A_i) - \log P(S) \tag{14.3}$$

Thus if the neurons have firing rates proportional to the logarithms of probabilities, the update according to Bayes' theorem can be performed just using addition and subtraction. The computation of the logarithm of the normalization term becomes only slightly more complex, as it needs to include nonlinear transformations:

$$\log P(S) = \log \sum_{i=1}^{N} \exp\left(\log P(A_i) + \log P(S|A_i)\right) \tag{14.4}$$

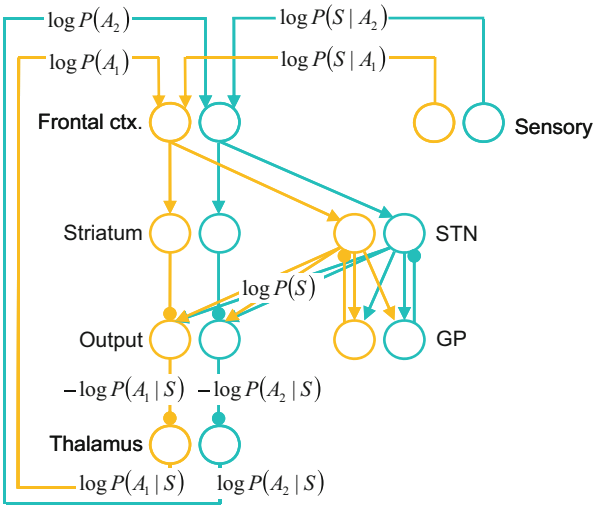


Fig. 14.2 Mapping of Bayes' theorem on a subset of the anatomy of the cortico-basal-ganglia-thalamic circuit. *Yellow* and *blue circles* denote neural populations selective for two sample actions, and the labels next to the *circles* indicate their locations. *Arrows* denote excitatory connections, while *lines* ended with *circles* denote inhibitory connections

Figure 14.1b illustrates the update of logarithms of probabilities as a network of computational elements.

14.2.3 Implementation Level

Figure 14.2 illustrates how the computations described in Eq. 3 or Fig. 14.1b could be mapped on the subset of the known connectivity of the cortico-basal-ganglia-thalamic circuit [10].

The logarithms of sensory inputs given actions, $\log P(S|A_i)$, could be represented in the firing rates of sensory neurons. Let us illustrate it in the task considered in the Introduction. If a low tone is presented, then $\log P(S|A_1)$ is higher than $\log P(S|A_2)$, and the neurons in auditory cortex selective for low tone have higher firing rate than the neurons selective for high tone (and vice-versa for high tone). Thus $\log P(S|A_1)$ and $\log P(S|A_2)$ are proportional to firing rates of auditory neurons selective for low and high tone respectively. Hence terms $\log P(S|A_1)$ and $\log P(S|A_2)$ could be provided by appropriately weighted inputs for the sensory neurons (such appropriate weightings could be learnt using reinforcement learning, e.g. [18]). It has been also shown that $\log P(S|A_i)$ could be represented in the firing rates of sensory neurons for more complex perceptual tasks, e.g. the motion discrimination task [19].

In the model illustrated in Fig. 14.2, the neurons in the frontal cortex add the input from sensory neurons to the logarithm of the prior probability which is provided by a feedback from the thalamus, thus they perform the addition in Eq. 3. The logarithm of the normalization term is computed in the model in a circuit of subthalamic nucleus (STN) and external segment of globus pallidus (GP), and this computation is described in detail later. The output nuclei receive excitation from STN (which in

the model is proportional to $\log P(S)$) and inhibition from the cortex via the striatum (which in the model is proportional to $\log P(A_i) + \log P(S|A_i)$), and subtract these two inputs, thus according to Eq. 3, their activity is proportional to $-\log P(A_i|S)$. The output nuclei send inhibition to the thalamus, so the activity in the thalamus is proportional to the logarithm of the posterior probability, i.e. $\log P(A_i|S)$. Finally, the logarithm of the posterior probability is sent back from the thalamus to the frontal cortex as it becomes the basis of the computation (or prior $\log P(A_i)$) in the next time step.

We now need to consider a biological constraint that the firing rates cannot be negative, as logarithms of probabilities are negative (because probabilities are by definition smaller than 1). This problem can be solved by assuming that the firing rates are proportional to logarithms of probabilities increased by a constant c. Equations below describe computations performed by each of the nuclei:

$$SEN_i = \log P(S|A_i) + c \tag{14.5}$$

$$CTX_i = \begin{cases} \log P(A_i) + c + SEN_i & \text{at the first interval} \\ TH_i(t-1) + SEN_i & \text{at subsequent intervals} \end{cases} \tag{14.6}$$

$$STN = \log \sum_{i=1}^{N} \exp CTX_i \tag{14.7}$$

$$OUT_i = -CTX_i + STN \tag{14.8}$$

$$TH_i = c - OUT_i \tag{14.9}$$

In the above equations SEN_i, CTX_i, OUT_i and TH_i are firing rates of populations of sensory, frontal cortical, output and thalamic neurons selective for alternative i. At the start of the trial the cortical neurons are initialized to the logarithms of initial prior probabilities of actions, and subsequently they receive feedback equal to thalamic activity in the previous time step $TH_i(t-1)$. Note in Fig. 14.2 that in the model each neural population in the output nuclei receives input from all populations in the STN in agreement with experimental data suggesting that STN projections are more diffuse [20, 21]. Thus the term STN in Eq. 8 actually denotes the sum of activities across all STN populations:

$$STN = \sum_{j=1}^{N} STN_j \tag{14.10}$$

We will describe how the activity described by Eq. 7 could arise in the STN later, but first let us demonstrate that the model described by Eqs. 5–9 correctly updates probabilities. At the first interval the cortical activity is equal to (from Eqs. 5–6):

$$CTX_i = \log P(A_i) + \log P(S|A_i) + 2c \tag{14.11}$$

Thus cortical neurons encode the sum of log probabilities indicated by labels in Fig. 14.2 increased by a constant that is the same for all alternatives. The model has a property that if the same constant is added to the activity of all cortical populations, the activity in the output nuclei does not change. In particular, if to simplify notation we denote $\log P(A_i) + \log P(S|A_i)$ by X_i (so that $CTX_i = X_i + 2c$), then the feedback provided by the STN is (from Eq. 7):

$$STN = \log \sum_{i=1}^{N} \exp(X_i + 2c) = \log \left(\exp(2c) \sum_{i=1}^{N} \exp X_i \right)$$
$$= \log \exp(2c) + \log \sum_{i=1}^{N} \exp X_i = \log \sum_{i=1}^{N} \exp X_i + 2c \tag{14.12}$$

Constants $2c$ then cancel while computing the activity in the output nuclei:

$$OUT_i = -X_i - 2c - \log \sum_{i=1}^{N} \exp X_i + 2c = -\log P(A_i|S) \tag{14.13}$$

We have shown that at the end of the first interval the model computes the posterior probabilities of actions. Since they are then fed back to cortical integrators as prior for the next interval it can be shown using analogous calculations that the network computes correctly the posterior probability in every subsequent interval.

Let us now consider how the STN-GP circuit could compute the expression in Eq. 7 required for normalizing represented probabilities. Bogacz and Gurney [4] have shown that the STN-GP circuit with the architecture shown in Fig. 14.2 would produce the above activity of STN, if the neural populations in STN and GP had the following relationships between their inputs and their firing rates:

$$STN_i = \exp(CTX_i - GP_i) \tag{14.14}$$

$$GP_i = STN - \log STN \tag{14.15}$$

In the above equations, STN_i and GP_i denote the firing rates of STN and GP neurons selective for action A_i. The STN neurons receive excitation from cortex and inhibition from GP, so their total input is $CTX_i - GP_i$, thus Eq. 14 implies that the STN neurons in the model have an exponential relationship between their input and firing rate. The GP neurons receive input from STN, but this input is coming in Fig. 14.2 from STN neurons selective for all actions which we denote by STN without a subscript (see Eq. 10). Equation 15 implies that the GP provides inhibition proportional to $STN - \log STN$.

Before considering a mathematical proof, let us provide an intuition for how the STN-GP circuit computes Eq. 7. Starting from the right end of Eq. 7, the cortical activity CTX_i is provided to the STN in the model by the hyperdirect pathway from cortical populations (see Fig. 14.2). The exponentiation is performed by the STN neurons (cf. Eq. 14). The summation is achieved due to the diffuse projections from

the STN—in the model each neural population in the output nuclei receives input from all populations in the STN hence the neurons in the output nuclei can sum the activity of STN populations. The only non-intuitive element of the computation of Eq. 7 is the logarithm—it comes from the interactions between STN and GP. Showing that the model of STN-GP circuit described by Eqs. 14 and 15 produces the activity level given in Eq. 7 is actually very easy, and we encourage readers to try it now—this is Exercise 2 at the end of the chapter where we also provide some hints how to do it.

14.3 Relationship to Other Models

In this section we discuss the relationship of the model described above, to which we will now refer as the MSPRT model, to various other models of decision making. The MSPRT model is closely related to the diffusion model, which has been shown to describe reaction times in a wide range of tasks (see Chap. 7 by Forstmann and Wagenmakers). The diffusion model performs the same computations as the SPRT procedure [17, 22]. Since for two alternatives, MSPRT reduces to SPRT, the MSPRT model produces the same behaviour as the diffusion model, or more precisely a simple version of the model without variabilities in drift and starting point called the EZ-diffusion [23].

The MSPRT model differs from previous work on Bayesian models of neural decision making [24, 25] in that the previous work did not consider the normalization in the Bayes' theorem. Indeed, the normalization term does not need to be computed if one just wants to find which action is most likely given the data available (note in Eq. 1, that the normalization is the same for all i, so if one wishes to compare which posterior probability is the highest, it is sufficient to just compare the numerators). However, in many natural scenarios, the animals and humans are free to choose whether to make an action or continue observing sensory information to gain more confidence. As mentioned in Sect. 2.1, in such situations choosing an action when the probability of this action being appropriate reaches a threshold allows maximization of reward rate. To implement such a decision procedure the exact (rather than relative) values of posterior probabilities need to be computed, so the normalization needs to also be performed.

The mapping of Bayes' theorem on the anatomy of the cortico-basal-ganglia circuit described in Sect 2.3 is one of the published mappings [10] and it is the easiest one to explain, but 3 other mappings have also been published [4, 26, 27]. They differ in small details, but they all assume that somewhere in the circuit the posterior probabilities of actions are computed, and that the STN-GP circuit computes the logarithm of the normalization. The mapping of Lepora and Gurney [27] differs from the one described in this chapter that it allows the cortical neurons selective for a particular action to have the activity proportional to the logarithm of the ratio of the likelihood of sensory input given this action and the likelihood of the input given other actions. This property brings the model in line with experimental data

suggesting that neurons in lateral intraparietel area have activity proportional to the log-likelihood-ratio during decision making [28, 29].

This Chapter described the MSPRT model in the context of selection of actions in highly practiced tasks in which the mapping between stimulus and response has been consolidated in the cortex. But as we mentioned in the Introduction, the basal ganglia are also involved in learning which actions are worth selecting in given circumstances, and elegant models describing reinforcement learning in the basal ganglia have been developed (Chapter 8 by Frank). We feel that the reinforcement learning and MSPRT models are complementary and describe different aspects of basal ganglia function. Bogacz and Larsen [10] have also modelled tasks in which the expected rewards for making different actions have to be learnt from feedback, and they showed that the model can also approximate MSPRT once these expected rewards have been learnt in cortico-striatal synapses as described by the reinforcement learning models.

Finally, let us discuss the relationship of the model with previous theories describing the role of the STN in action selection. In the MSPRT model, the STN fulfills the function assigned to it by these theories. In particular, the STN is involved in inhibition of non-selected actions [3], as when the probability of one action increases, STN must ensure that the represented probabilities of other actions decrease. Also, the STN is involved in postponing action execution in the face of conflicting information [8], as when two actions receive equally high input, then after normalization their probability will be 50 % and none of them will exceed a sufficiently high threshold of confidence, until the conflict is resolved. However the model extends the description of STN function by postulating it ensures that represented probabilities add up to one throughout the decision process.

14.4 Relationship to Experimental Data

Due to the close relationship of the MSPRT model to the diffusion model and earlier theories of the STN, the MSPRT model is consistent with a wide range of data supporting these theories. Additionally the MSPRT model makes several precise predictions which are unique to this model. In this Section, we first review the relationship to some of these predictions to existing data, and then list further predictions.

The MSPRT model predicts that the firing rate of the STN neurons should be proportional to the exponent of their input (Eq. 14). In the literature there are reports of seven neurons for which the firing rate as a function of injected current was studied in detail [30, 31]. Figure 14.3 shows that the firing rate of these neurons is indeed very closely described by an exponential function up to 135 Hz, which is approximately the range of firing of the STN neurons in vivo.

Since in the MSPRT model, the STN is a part of the circuit updating probabilities of action on the basis of sensory input, the model predicts that disrupting information processing in the STN with deep brain stimulation (DBS) should impair this integration of probabilistic information. Indeed, Coulthard et al. [18] have shown that

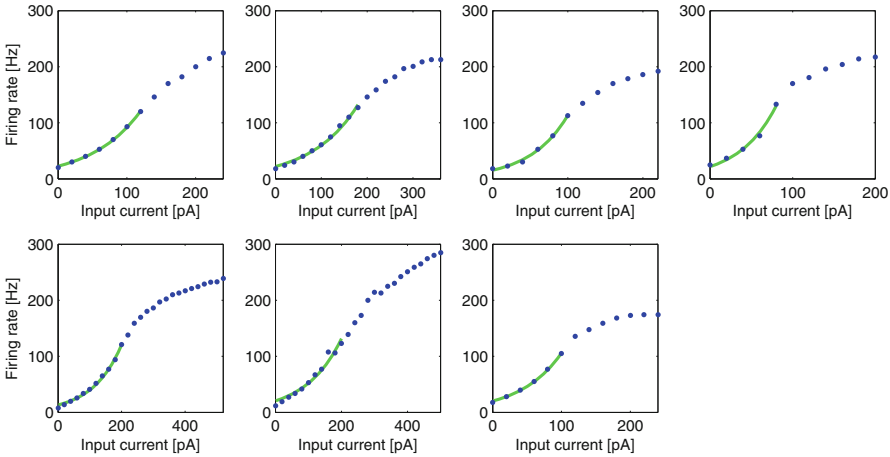


Fig. 14.3 Firing rates f of STN and GP neurons as a function of input current I. *Blue dots* in *top* row of panels re-plot data on the firing rate of STN neurons presented in Hallworth et al. (2003) in Fig. 4b, 4f, 12d, 13d respectively (control condition). *Bottom* row of panels re-plot the data from STN presented in Wilson et al. (2004) in Fig. 1c, 2c, 2f respectively (control condition). *Lines* show best fit of the function $f = a \exp(b\, I)$ to firing rates below 135 Hz

the patients with Parkinson's disease were poorer in action selection on the basis of information in multiple sequentially presented stimuli when the DBS was turned on than when the DBS was turned off.

The MSPRT model makes several other predictions. First, it predicts that the feedback provided by GP neurons is a function of STN activity described by Eq. 15. Second, it predicts that DBS should impair the normalization of probabilities of actions represented by patients. Third, it predicts that the activity of STN during action selection should be proportional to the logarithm of the normalization term. Testing these predictions will reveal to what extent the model presented in this Chapter describes the computations in cortico-basal-ganglia circuit.

Acknowledgement This work was supported by EPSRC grant EP/I032622/1.

Exercises

1. Consider a rat extensively trained in the task described in the Introduction. Assume that the rat correctly learned the probabilities of different tones occurring on trials with different actions being rewarded, and at the beginning of a trial both actions seem equally likely to be correct. Compute how the estimated probabilities of actions $P(A_i)$ change after the rat hears two beeps with low tone. If you do not know how to start this computation, we recommend looking at the first part of

the solution describing how the probabilities are updated after the first beep, and then trying to compute the updated probabilities after the second beep.

2. Show that the model produces the activity of STN required for normalization of probabilities. In particular, show that Eqs. 10, 14 and 15 imply Eq. 7. Hint: Substitute Eq. 15 into Eq. 14 and then sum across all alternatives.

Further Reading

Ditterich JA (2010) Comparison between mechanisms of multi-alternative perceptual decision making: ability to explain human behavior, predictions for neurophysiology, and relationship with decision theory. Front Neurosci 4:184

Frank MJ, Samanta J, Moustafa AA, Sherman SJ (2007) Hold your horses: impulsivity, deep brain stimulation, and medication in parkinsonism. Science 318(5854):1309–1312

Lepora NF, Gurney KN. (2012) The basal ganglia optimize decision making over general perceptual hypotheses. Neural Comput 24(11):2924–2945

Zaghloul KA, Weidemann CT, Lega BC, Jaggi JL, Baltuch GH, Kahana MJ (2012) Neuronal activity in the human subthalamic nucleus encodes decision conflict during action selection. J Neurosci 32:2453–2460

References

1. Redgrave P, Prescott TJ, Gurney K (1999) The basal ganglia: a vertebrate solution to the selection problem? Neuroscience 89(4):1009–1023
2. Alexander GE, Crutcher MD (1990) Functional architecture of basal ganglia circuits: neural substrates of parallel processing. Trends Neurosci 13(7):266–271
3. Mink JW (1996) The basal ganglia: focused selection and inhibition of competing motor programs. Prog Neurobiol 50(4):381–425
4. Bogacz R, Gurney K (2007) The basal ganglia and cortex implement optimal decision making between alternative actions. Neural Comput 19:442–477
5. Doya K (2000) Complementary roles of basal ganglia and cerebellum in learning and motor control. Curr Opin Neurobiol 10(6):732–739
6. Frank MJ, Seeberger LC, O'Reilly RC (2004) By carrot or by stick: cognitive reinforcement learning in Parkinsonism. Science 306(5703):1940–1943
7. Gurney K, Prescott TJ, Redgrave PA (2001) Computational model of action selection in the basal ganglia. I. A new functional anatomy. Biol Cybern 84(6):401–410
8. Frank MJ (2006) Hold your horses: a dynamic computational role for the subthalamic nucleus in decision making. Neural Netw 19(8):1120–1136
9. Frank MJ, Samanta J, Moustafa AA, Sherman SJ (2007) Hold your horses: impulsivity, deep brain stimulation, and medication in parkinsonism. Science 318(5854):1309–1312
10. Bogacz R, Larsen T (2011) Integration of reinforcement learning and optimal decision-making theories of the basal ganglia. Neural Comput 23(4):817–851
11. Marr D (1982) Vision: a computational investigation into the human representation and processing of visual information. W.H. Freeman and Company, San Francisco
12. Baum CW, Veeravalli VV (1994) A sequential procedure for multihypothesis testing. IEEE Trans Inf Theory 40:1996–2007

13. Wald A (1947) Sequential analysis. Wiley, New York
14. Wald A, Wolfowitz J (1948) Optimum character of the sequential probability ratio test. Ann Math Stat 19:326–339
15. Dragalin VP, Tertakovsky AG, Veeravalli VV (1999) Multihypothesis sequential probability ratio tests—part I: asymptotic optimality. IEEE Trans Inf Theory 45:2448–2461
16. McMillen T, Holmes P (2006) The dynamics of choice among multiple alternatives. J Math Psychol 50:30–57
17. Bogacz R, Brown E, Moehlis J, Holmes P, Cohen JD (2006) The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced choice tasks. Psychol Rev 113:700–765
18. Coulthard EJ, Bogacz R, Javed S, Mooney LK, Murphy G, Keeley S et al. (2012) Distinct roles of dopamine and subthalamic nucleus in learning and probabilistic decision making. Brain 135(Pt 12):3721–3734
19. Zhang J, Bogacz R (2010) Optimal decision making on the basis of evidence represented in spike trains. Neural Computation 22:1113–1148
20. Parent A, Hazrati LN (1993) Anatomical aspects of information processing in primate basal ganglia. Trends Neurosci 16(3):111–116
21. Parent A, Hazrati LN (1995) Functional anatomy of the basal ganglia. I. The cortico-basal ganglia-thalamo-cortical loop. Brain Res Brain Res Rev 20(1):91–127
22. Laming DRJ (1968) Information theory of choice reaction time. Wiley, New York
23. Wagenmakers EJ, van der Maas HL, Grasman RP (2007) An EZ-diffusion model for response time and accuracy. Psychon Bull Rev 14(1):3–22
24. Beck JM, Ma WJ, Kiani R, Hanks T, Churchland AK, Roitman J et al (2008) Probabilistic population codes for bayesian decision making. Neuron 60(6):1142–1152
25. Ma WJ, Beck JM, Latham PE, Pouget A (2006) Bayesian inference with probabilistic population codes. Nat Neurosci 9(11):1432–1438
26. Ditterich JA (2010) Comparison between mechanisms of multi-Alternative perceptual decision making: ability to explain human behavior, predictions for neurophysiology, and relationship with decision theory. Front Neurosci 4:184
27. Lepora NF, Gurney KN (2012) The basal ganglia optimize decision making over general perceptual hypotheses. Neural Comput 24(11):2924–2945
28. Gold JI, Shadlen MN (2002) Banburismus and the brain: decoding the relationship between sensory stimuli, decisions, and reward. Neuron 36(2):299–308
29. Yang T, Shadlen MN (2007) Probabilistic reasoning by neurons. Nature 447(7148):1075–1080
30. Hallworth NE, Wilson CJ, Bevan MD (2003) Apamin-sensitive small conductance calcium-activated potassium channels, through their selective coupling to voltage-gated calcium channels, are critical determinants of the precision, pace, and pattern of action potential generation in rat subthalamic nucleus neurons in vitro. J Neurosci 23(20):7525–7542
31. Wilson CJ, Weyrick A, Terman D, Hallworth NE, Bevan MD (2004) A model of reverse spike frequency adaptation and repetitive firing of subthalamic nucleus neurons. J Neurophysiol 91(5):1963–1980

Chapter 15
Inhibitory Control in Mind and Brain: The Mathematics and Neurophysiology of the Underlying Computation

Gordon D. Logan, Jeffrey D. Schall and Thomas J. Palmeri

Abstract We develop desiderata for a computational theory of response inhibition that links mathematical psychology with neuroscience. The theory must be explicit mathematically and computationally, and grounded in behavior and neurophysiology. The theory must provide quantitative accounts of complexities of behavior in response inhibition tasks and must predict the neural activity that underlies performance. We evaluate three current theories of response inhibition in the stop signal paradigm using these desiderata, and we find that one theory fulfills the desiderata better than the others.

15.1 Introduction

> Yawning, Goldilocks walked into the bedroom and saw three beds. "This one's too big," she said. "This one's too small. But this one's just right." She crawled under the covers, fell fast asleep, and dreamed of unimagined wonders.

We are lucky to live in an era in which the dreams we dared to dream are coming true. Mathematical psychology and neuroscience are merging, and the merger is yielding amazing insights into the mind and brain that were unimaginable 20-years-ago. Mathematical psychology has provided us with precise, explicit descriptions of mental processes that are linked tightly to behavior, making strong predictions about behavior that stand up to rigorous empirical tests. Accurate prediction of response time (*RT*) distributions for correct and error responses is now commonplace, and it is the standard by which models are judged. Neuroscience has opened the black box and shown us how the neural processes underlying behavior interact and unfold in real time. Analysis of spike trains from single neurons, local field potentials from groups of neurons, and electroencephalographic activity at the dura, skull, and scalp have revealed the time-course of information processing. Studies of anatomy, lesions, and brain imaging have shown us the networks of neurons that process information. In recent years, we have seen a proliferation of theories that merge the insights from

G. D. Logan (✉) · J. D. Schall · T. J. Palmeri
Department of Psychology, Vanderbilt University, Nashville, TN 37203, USA
e-mail: gordon.logan@vanderbilt.edu

B. U. Forstmann, E.-J. Wagenmakers (eds.), *An Introduction to Model-Based Cognitive Neuroscience,* DOI 10.1007/978-1-4939-2236-9_15

mathematical psychology and neuroscience, identifying the computational mechanisms in mathematical models with individual neurons and systems of neurons that implement the computation, and testing the identification rigorously by fitting both behavioral and neural data. In all these models, the fundamental insight that made the dream come true is the idea that mind and brain are the computers that produce behavior, and the computation is one and the same.

15.2 Imagining the Dream

We dreamed of a theory that applies that fundamental insight to response inhibition, especially in the *stop-signal* or *countermanding* task [13]. We dreamed of a theory that was formulated explicitly in mathematics or computer simulation, grounded in behavior, computation, and neurophysiology. The theory should accurately predict important behavioral phenomena with models that are connected to the extensive theory of stochastic accumulation to a threshold. The theory should specify linking propositions that connect the mathematical description to neurons, groups of neurons, or brain regions [22, 23]. The linking propositions identify the points of contact between theory and neural data, and specify the aspects of the data that are relevant to the theory. In the stop-signal task, a theory of response inhibition must provide a quantitative account of the probability of inhibiting a response and explain how it varies with the time available to stop (*stop-signal delay*, or *SSD*). The theory must provide a quantitative account of RT distributions for error and correct responses. In the stop-signal task, this means accounting for the relation between failures to inhibit (*signal-respond* or *non-cancelled* trials) and successful responses to the go task (*no-stop-signal* or *cancelled* trials), and accounting for changes in the signal-respond RT distribution with SSD.

Our dream theory provides a list of desiderata that we have used to guide our own modeling: The theory must account for behavior, neurophysiology, and computation, it must be explicit mathematically or computationally, and it must fit the data better than plausible alternatives. In this chapter, we use these desiderata to evaluate current theories of response inhibition in the stop-signal task. The theories are formulated at three different levels of analysis. The highest level addresses networks of brain regions that participate in response inhibition, specifying the interactions within and between regions. The middle level addresses firing rates in systems of neurons that participate in response inhibition, specifying excitatory and inhibitory connections. The lowest level addresses spiking neurons, specifying the connections between spike trains and the underlying biochemistry. Like Goldilocks, we will conclude that one of these levels is too big, one is too small, and one is just right. But we are getting ahead of ourselves. Let us begin by describing behavior in the stop-signal task and the independent race model that accounts for it.

Waking just enough to notice the world around us, we realize there are other dreamers and other dreams. In the other dreams, Goldilocks might prefer a bigger or smaller level of theorizing, fulfilling desiderata that emphasize large networks or biochemistry. Rolling over, we snuggle back into our own dream for the rest of this chapter.

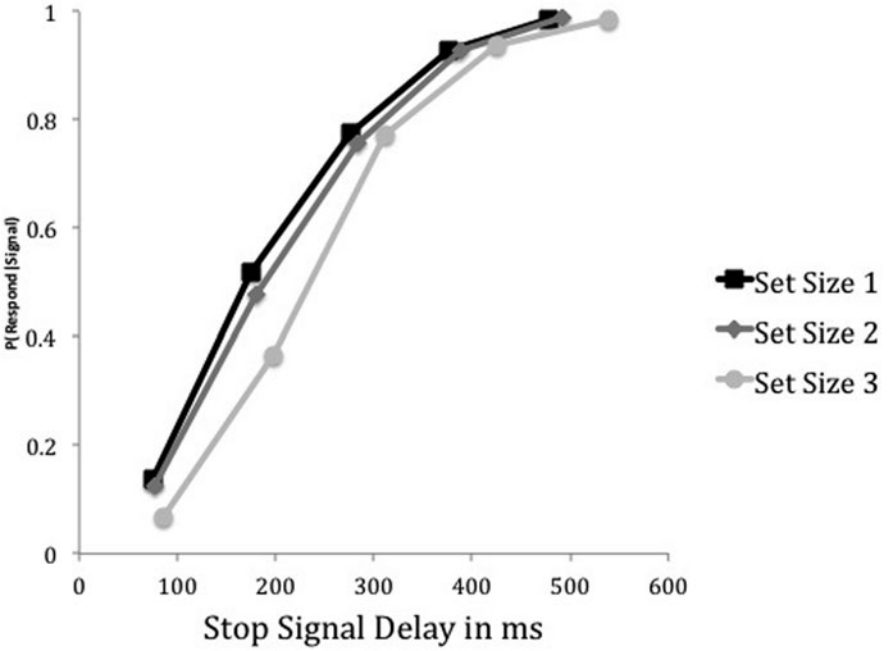


Fig. 15.1 Inhibition Function from a memory-search experiment in which the number of items in the memory set was varied. The probability of responding given a stop signal increases as stop-signal delay (*SSD*) increases and decreases as response time (RT) in the go task increases (RT1 < RT2 < RT3)

15.2.1 Response Inhibition in the Stop-Signal Paradigm

The ability to inhibit our responses voluntarily is a paradigm case of cognitive control. It shows we have "the freedom to do otherwise," which is a hallmark of free will. It reveals itself in many behavioral paradigms, but it is revealed most clearly, simply, and directly in the stop-signal paradigm (for reviews, see [12, 13, 26]). In this paradigm, subjects perform a "go" task, in which they make a speeded response to an imperative stimulus. On some trials, a "stop signal" is presented that tells subjects to inhibit their response to the go signal. Whether or not they are able to is the main datum of interest. Many studies show that the ability to inhibit responses is probabilistic, and the probability of inhibition depends primarily on SSD (see Fig. 15.1). Stop-signal delay controls the amount of time available to detect the stop signal and countermand the go response before the go response is executed; response inhibition is more likely when more time is available. Signal-respond RT is also an important datum. It is usually faster than RT on trials with no stop signal, as if it comes from the faster tail of the go RT distribution (see Fig. 15.2).

These effects have been observed in several species, including rats, monkeys, and humans, in several subject populations, including children, adolescents, young adults, and the elderly. These effects have been observed in several psychiatric

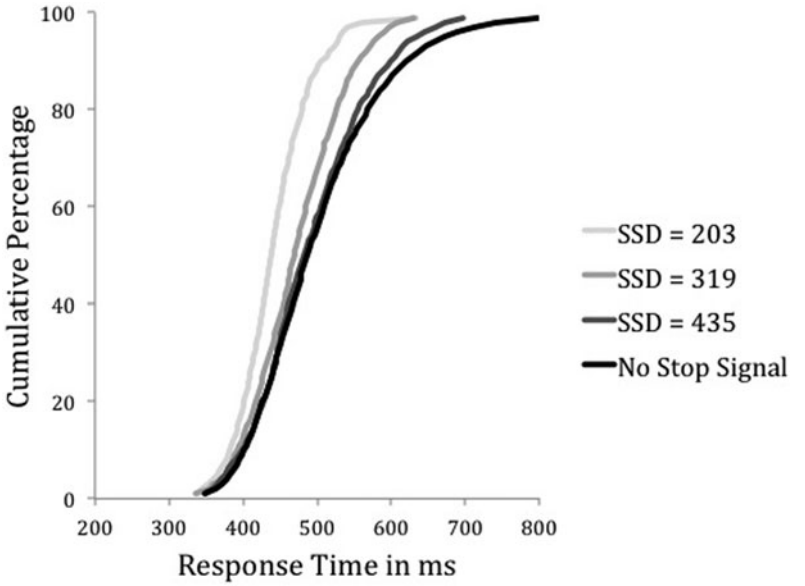


Fig. 15.2 Distributions of response time on no-stop-signal trials and on signal-respond trials with stop signal delay (SSD) equal to 231, 364, and 496 ms. Signal-respond distributions are faster than no-stop-signal distributions. They begin with a common minimum and end with a shorter maximum

disorders, including attention deficit hyperactivity disorder and schizophrenia, and in several neurological disorders, including stroke and Parkinson's disease. They have been observed in different stimulus and response modalities, in different tasks, in different experimental conditions, and with different strategies. The patterns are the same qualitatively, but they differ quantitatively, and the quantitative differences reveal important changes or deficits in cognitive control.

15.2.2 Independent Race Model

Two facts led Logan and Cowan [13] to propose the independent race model of stop signal performance: (1) The probability of response inhibition depends on the time available to detect the stop signal before the go response is executed, and (2) signal-respond RTs are faster than RTs on no-stop-signal trials. These facts suggested that response inhibition depends on the outcome of a race between a go process, initiated by the go stimulus, and a stop process, initiated by the stop signal. If the stop process finishes before the go process, the response is inhibited, producing a signal-inhibit trial. If the go process finishes before the stop process, the response

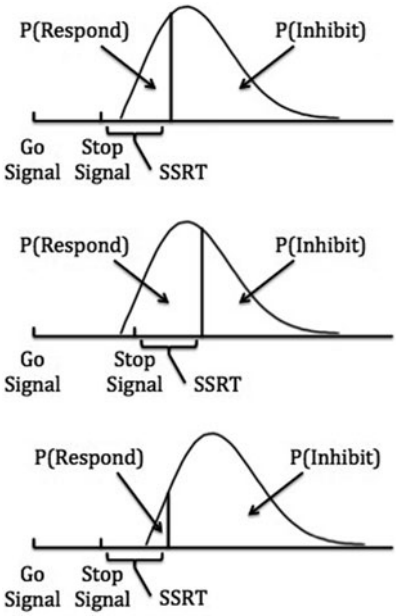


Fig. 15.3 Predictions of the independent race model, assuming *SSRT* is constant. Onset of Go Signal followed by onset of Stop Signal after a stop-signal delay. *Vertical line* across the distribution represents the finishing time of the stop process. Probability of responding is area to *left* of line; probability of inhibiting is area to *right* of line. Top panel: standard condition. *Middle* panel: Stop-signal delay increases, so probability of responding increases. Bottom panel: Go response time increases, so probability of responding decreases

is not inhibited, producing a signal-respond trial. The model assumes that the finishing times for the stop and go processes are independent random variables, and demonstrates that the fundamental results in the stop-signal paradigm follow from these assumptions (see Fig. 15.3).

The independent race model provides a measure of the latency of the stop process, called *stop-signal reaction time* (*SSRT*). This is an important contribution because the stop process is not directly observable. If the stop process finishes before the go process, there is no response whose latency can be measured. If the stop process finishes after the go process, we know SSRT must have been longer than signal-respond RT, but we do not know how much longer. The independent race model provides several converging methods for estimating SSRT from the observed data. These measures of SSRT have been important in documenting differences in the ability to inhibit responses across lifespan development, between clinical and control groups, and between neurological patients and controls. They have also been important in understanding the neurophysiology of response inhibition. Neural processes that cause response inhibition must modulate before SSRT; neural processes that are consequences of response inhibition modulate after SSRT.

Since it was formulated in 1984, the independent race model has been used in virtually every stop-signal experiment. It provides important measures of cognitive

control, like SSRT, and it provides a benchmark against which other models can be evaluated. Its prevalence results from its generality: It is formulated in terms of generic finishing time distributions for the stop and go processes. It makes no commitment to the underlying computational or neural processes that generate these finishing times. It expresses relationships that must hold for any and all distributions, regardless of the process that generates them. This is important because the independent race model provides an important check for the models we consider here that address the computations performed by the underlying neural processes: these models must predict the empirical relationships predicted by the independent race model.

The independent race model is like a dream: it captures the essence but not the details. It formulates the constraints that any model of response inhibition must follow, but it does not provide the structure that seems necessary to explain recent developments in stop-signal research. For example, many studies have shown that go RT is slower when stop trials occur more frequently, as if the go process changes to balance the competing demands of stopping and going. Many other studies have shown that go RT is slower on trials following stop signals than on trials before them, suggesting that a stop trial results in some kind of strategic adjustment to the go process. To explain how these adjustments occur, we need a more detailed model of the go process that tells us which parts can support this strategic adjustment. The independent race model provides no model of the underlying process. It can describe these effects, but it cannot explain them.

15.3 Feeding the Dream

Developments in mathematical psychology and neuroscience around the turn of the twenty-first century set the stage for the development of models that link mind and brain. Mathematical psychologists developed a variety of *stochastic accumulator models* that explained RT distributions for correct and error responses as resulting from processes that accumulate information until a threshold for responding is reached. Many studies evaluated the strengths and weaknesses of random walk, diffusion, race, and leaky competitive accumulator models, using increasingly sophisticated methods for assessing goodness of fit and increasingly stringent comparative tests of one model against another e.g. [20]. Models must fit large amounts of data with a small number of free parameters, and they must fit better than plausible alternatives when model complexity is taken into account. Researchers either compare one model architecture against another or compare different models in the same architecture to determine which parameters are necessary and sufficient to account for the data. These models and the approach they took to modeling inspired more specific models of response inhibition with greater explanatory power.

At the same time, neuroscientists were training animals to perform the stop-signal task and recording from their brains as they performed it. Hanes and Schall [7] showed that monkeys performed a saccadic version of the stop signal task much like humans. The probability they would inhibit their eye movements depended on SSD and their signal-respond RTs were faster than their no-stop-signal RTs. Hanes, Patterson and

Schall [8] recorded from frontal eye fields in monkeys performing the saccadic stop signal task, isolating neurons involved in gaze shifting and gaze holding that represent a larger circuit of such neurons that extends from cortex through basal ganglia and superior colliculus to brainstem. They found that these neurons modulated on stop-signal trials, modulating just before SSRT when the monkey stopped successfully. Paré and Hanes [15] reported similar results in superior colliculus. Meanwhile, studies of humans with lesions in frontal cortex revealed deficits in stop-signal inhibition, and functional magnetic resonance imaging (*fMRI*) on healthy young adults suggested the involvement of a circuit including frontal cortex, basal ganglia, and subthalamic nucleus [1]. These rich neural data sets demand computational explanations that are more detailed than the description the independent race model provides.

15.4 Dreaming the Dream

In recent years, many theories of response inhibition have been developed. We focus on three models that account for behavior, computation, and neurophysiology in the stop-signal task. One focuses on brain regions, one focuses on processes that generate spikes and spike trains, and one focuses on firing rates in single neurons. Like the Goldilocks in our dream, we conclude that one is too big, one is too small, and one is just right. Of course, other Goldilocks' in other dreams may reach different conclusions.

15.4.1 Single Neurons: The Interactive Race Model

Boucher et al. [5] formulated an interactive race model to address a paradox they encountered in linking models to neurons: How can a model that assumes independent stop and go processes explain behavior that is supported by interacting circuits of mutually inhibitory gaze-holding and gaze-shifting neurons? They addressed this question by instantiating the stop and go processes as mutually inhibitory leaky competitive accumulators ([25]; see Fig. 15.4). The go accumulator begins after an afferent delay, D_{go}, accumulating activation until it reaches a threshold, whereupon a response occurs. The stop accumulator begins after an afferent delay, D_{stop}, inhibiting the go response in proportion to its activation. If the stop accumulator becomes active soon enough (if SSD + D_{stop} < go RT), it prevents the go accumulator from reaching threshold and the response is inhibited. If the stop process becomes active too late (if SSD + D_{stop} > go RT), the go accumulator reaches threshold and the response is not inhibited.

Boucher et al. [5] specified the stochastic differential equations that govern the stop and go accumulators and used them to drive computer simulations. They fit the simulations to behavioral data from two monkeys, who also provided neural data from the same test sessions, manipulating the mean and standard deviation of go and stop accumulation rates and the mutual inhibition from stop to go and go to stop to

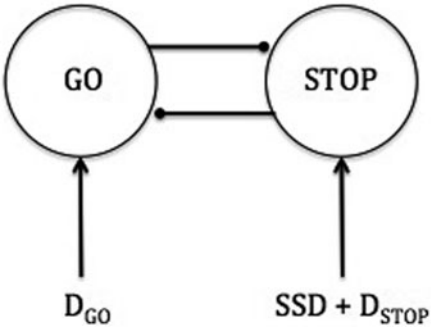


Fig. 15.4 Interactive race model. Arrows represent excitatory connections; dots represent inhibitory connections. The GO unit receives input after an afferent delay (D_{GO}) and the STOP unit receives input after stop-signal delay (*SSD*) plus an afferent delay (D_{STOP}). GO and STOP units inhibit each other. Inhibition from STOP to GO is much greater than inhibition from GO to STOP. A go response occurs if GO activation reaches threshold. The go response is inhibited if inhibition from the STOP unit prevents it from reaching threshold

optimize goodness of fit. The model fit the data well, providing accurate quantitative accounts of the inhibition function, no-stop-signal RTs, and signal-respond RTs at several SSDs (see [5], Fig. 6). Thus, the model fulfills the behavioral side of the desiderata of our dreams.

Boucher et al. then simulated the growth and modulation of activation of the go and stop accumulators, using the parameters that produced the best fits to the behavioral data, and matched the simulated patterns of activation to measured patterns of activity in gaze-holding and gaze-shifting neurons that were recorded while monkeys performed the stop signal task. To assess the match between simulated and recorded activity, Boucher et al. had to decide which aspect of the recorded activity to assess. The pattern of activation for an individual neuron has many idiosyncrasies, but all patterns show some general characteristics. In order to fit the "signal" and not the "noise," Boucher et al. focused on distributions of *cancel times*, which are the times at which neural activity modulates on trials on which subjects stop successfully, relative to SSRT. They assessed this in the simulated data in the same way they assessed it in neural data, by determining the point at which activation on successful stop trials first differed significantly from activation on latency-matched no-stop-signal trials. In the neural data, this point ranges from 50 ms before to 50 ms after SSRT, with a mean 5–10 ms before SSRT. The model predicted distributions with the same range (see [5], Fig. 7). Note these are genuine predictions. They were generated with a fixed set of parameters that provided the best fit to the behavioral data, without any further adjustment to optimize the fit to neural data. Thus, the model fulfills the neural side of the desiderata of our dreams.

The final desideratum is comparative model fitting. Boucher et al. [5] compared the interactive race model with a version of the independent race model in which the stop and go process were modeled as leaky accumulators with no competition. After

their respective afferent delays (D_{stop} and D_{go}) they accumulate activation until one of them reaches a threshold. If the stop process finishes first, the response is inhibited; if the go process finishes first, the response is executed. Boucher et al. found that the independent race model fit the behavioral data as well as the interactive race model, suggesting mimicry. Normally, parsimony would favor the simpler independent race model over the more complex interactive race model. However, Boucher et al. argued that the interactive race model accounted for the neural data, predicting modulation of go activation on stop-signal trials and predicting cancel time distributions accurately, while the independent race model did not. They argued that this favored the interactive race model. Thus, the interactive race model fulfills all of the desiderata of our dreams: it is computationally explicit, it explains the underlying processes computationally and neurally, it provides accurate quantitative accounts of behavioral and neural data, and it won in competitive tests against a plausible alternative. If Goldilocks were a mathematical psychologist, we believe she would find our model just right.

What about the paradox? The interactive race model assumes an interaction between gaze-holding and gaze-shifting units, like the interaction between gaze-holding and gaze-shifting neurons that underlies eye movements. How can it account for data that are described just as well by the independent race model? The answer lies in the values of the best-fitting parameters: In order to fit the behavioral data, D_{stop} had to be long—almost as long as SSRT—and inhibition from the stop process on the go process had to be much stronger than the inhibition from the go process on the stop process. Thus, the stop process and the go process were independent for most of their durations, and response inhibition resulted from late and potent inhibition just before a go response occurred.

15.4.2 Spikes and Spike Trains: The Spiking Neuron Model

Lo et al. [10] implemented the Boucher et al. [5] interactive race model in Lo and Wang's [9] spiking cortico-basal ganglia circuit model of RT (see Fig. 15.5). The model assumes hundreds of units representing populations of movement neurons, fixation neurons, and inhibitory interneurons, and a control unit that turns the fixation neurons on and off. Each population produces Poisson spike trains that depend on the ratio of parameters representing NMDA and AMPA inputs. The model addresses fixation activity at the beginning of a trial and the transition from fixation to movement as well as the rise in movement activation to threshold. The model produces the transition from fixation to movement, and ultimately RT, by turning off the control unit that excites fixation units, thereby releasing tonic inhibition on the movement units and allowing their activity to rise to threshold.

Lo et al. [10] fit data from one of the two monkeys Boucher et al. [5] modeled. They fixed the number of units and many of the parameters across all conditions and manipulated three parameters to maximize goodness of fit: The mean and standard deviation of a Gaussian distribution for the time at which the control unit turned off,

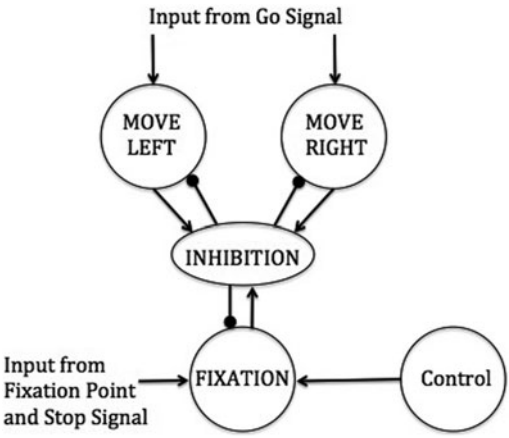


Fig. 15.5 Spiking neuron model. Arrows represent excitatory connections; dots represent inhibitory connections. The MOVE units are excited by input from the go signal. The *FIXATION* unit is excited by input from the fixation point and the stop signal and by control input. *MOVE* and *FIXATION* units activate an *INHIBITION* unit that inhibits them all. The Control unit tonically excites the *FIXATION* unit. A go response occurs when the Control unit releases excitation on the fixation unit. The go response is inhibited if the stop signal excites the *FIXATION* unit before a *MOVE* unit makes its response

and the time at which the stop signal turned the fixation units back on (analogous to D_{stop} in [5]). Their fits to RT distributions for no-stop-signal and signal-respond trials and their fits to inhibition functions were about as good as the fits Boucher et al. [5] obtained. Like Boucher et al., Lo et al. found that D_{stop} had to be relatively long to produce appropriate signal-respond RT distributions; inhibition of stop on go had to be late and potent. The Lo et al. model also predicted modulation of movement and fixation neurons and cancel time distributions qualitatively as well as Boucher et al. [5], although these predictions were not assessed quantitatively.

Lo et al. [10] modeled the effects of changes in baseline activation in fixation and movement units on the probability of successful inhibition. Successful inhibition was less likely when movement units were more active during the baseline period and more likely when fixation units were more active. They tested these predictions by reanalyzing data from the Hanes et al. [8] and Paré and Hanes [15] countermanding studies, and found lower baseline firing rates in movement neurons prior to successful inhibition.

What about our dream? The Lo et al. [10] model fulfills our behavioral desideratum, fitting the behavioral data as well as the Boucher et al. [5] model. The Lo et al. model fulfills our neural desideratum as well, describing stop and go units as spiking neurons and linking the computation to the biochemistry that generates spikes. However, the model does not fulfill our computational desideratum very well. RT depends on turning off a control unit that tonically excites fixation units, which releases inhibition on movement units and allows their activity to rise to threshold. The variability in RT depends primarily on the variability in the time at which the control signal is turned off, which is determined arbitrarily by a Gaussian distribution whose mean and standard deviation were free parameters that were adjusted to optimize

goodness of fit (113 and 95 ms, respectively). The control unit is like a homunculus outside the model that intervenes at the right time to produce the right effect. It is not grounded in the physiology, like movement and fixation units. There are no linking propositions [22, 23] that tie it to neurons or neural structures analogous to the linking propositions that tie movement and fixation units to gaze-shifting and gaze-holding neurons. We prefer models like the Boucher et al. [5] model, in which variability in RT is produced by variable growth in stochastic accumulation [18, 19] over the Lo et al. model, in which variability in RT is produced by an arbitrary control unit.

The Lo et al. [10] model partially fulfills our desideratum of comparative model fitting. Lo et al. compared their fits to Boucher et al.'s [5] fits of the interactive race model and the stochastic-rise-to-threshold version of the independent race model and found that their model fit about as well. They discovered the importance of differences in baseline activity in movement and fixation units in predicting the probability of successful inhibition, but that is not likely to be a unique prediction of their model. Differences in baseline activation could be implemented in the Boucher et al. [5] model, and would likely produce similar results. Thus, the Lo et al. model does not distinguish itself from plausible alternatives in comparative model fits, as our dream model would.

Lo et al. [10] modeled the underlying physiology at a finer grain than Boucher et al. [5], modeling spikes and spike trains rather than firing rates. However, this required many parameters (AMPA and NMDA ratios for each interaction between units) in addition to the three parameters that were varied to optimize goodness of fit. These parameters were fixed for the fitting, but they were tweaked to produce firing rates in the desired range for movement and fixation cells before they were fixed. From the perspective of mathematical psychology, where fitting large amounts of data with a small number of parameters is desirable, this is not a virtue. If Goldilocks were a mathematical psychologist, she would find the focus of this model (on spikes and spike trains) too small.

15.4.3 Brain Regions: The Frontal Cortex-Basal Ganglia Model

Wiecki and Frank [28] formulated a model of inhibitory control that extends Frank's [6] model of basal ganglia to include cortical structures. The new model describes interactions between units in frontal cortex (dorsolateral prefrontal cortex, right inferior frontal gyrus, frontal eye fields), basal ganglia (striatum, globus pallidus external segment, substantia nigra pars compacta, substantia nigra pars reticulata, subthalamic nucleus), and superior colliculus (see Fig. 15.6). It addresses the stop-signal task and an anti-saccade task in which a peripheral target is presented and subjects must inhibit their natural tendency to look directly at it and shift their gaze to a position opposite to it. The model explains stop-signal performance by assuming that the stop signal activates right inferior frontal gyrus, which activates subthalamic nucleus, which activates substantia nigra pars reticulata, which then inhibits superior colliculus. If the superior colliculus is inhibited before its activation reaches threshold, the response is inhibited, producing a signal-inhibit trial. If superior colliculus reaches threshold before it is inhibited, the response is executed, producing a signal-respond trial.

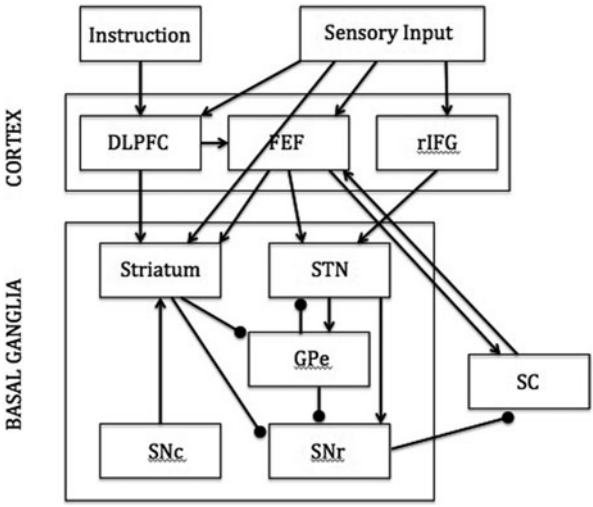


Fig. 15.6 Cortico-basal ganglia model. Arrows represent excitatory connections; dots represent inhibitory connections. *DLPFC* dorsolateral prefrontal cortex; *FEF* frontal eye fields; *rIFG* right inferior frontal gyrus; *STN* subthalamic nucleus; *GPe* globus palladus external segment; *SNc* substantia nigra pars compacta; *SNr* substantia nigra pars reticulate; *SC* superior colliculus. Response inhibition occurs when rIFG activates STN, which activates SNr, which inhibits SC

Wiecki and Frank [28] simulated performance on the stop-signal task but did not fit their model to the data. They simulated inhibition functions and RT distributions on no-stop-signal and signal-respond trials but did not compare the simulated functions quantitatively to observed data. The only observed data they reported were RT distributions taken from one of the monkeys studied by Boucher et al. [5] and Lo et al. [10], and their simulations overestimate the variability in the observed distributions (see their Fig. 12.). They reported simulated activation for units in striatum, substantia nigra pars reticulata, subthalamic nucleus, and dorsolateral prefrontal cortex, but did not compare the changes in activation with observed neural data.

What about our dream? The model fulfills our computational desideratum, explaining the mathematics and computations that occur within and between units, but it does not fulfill the other desiderata as well as we would like. The lack of quantitative fits falls short of fulfilling our behavioral desideratum. Every model of the stop-signal task predicts inhibition functions and RT distributions for no-stop-signal and signal-respond trials, so the model's predictions of the shapes of these functions are far from unique. Moreover, other models predict these functions quantitatively, and the models rise and fall on the accuracy of their quantitative fits. In fact, the independent race model [13] predicts these effects without specifying any of the underlying computations, so it is not clear that the machinery in the Wiecki and Frank [28] model is doing any of the work. We view this as a shortcoming of their model.

The Wiecki and Frank [28] model promises to fulfill our neural desideratum but also falls short. It is clear that stop-signal performance depends on the integrated

action of many brain structures, and the model includes the relevant structures. However, the linking propositions that connect model units to brain structures and interactions between model units to interactions between brain structures are not evaluated very rigorously. The model provides a framework in which these desiderata could be fulfilled, but does not go as far as we would like toward fulfilling them. Quantitative comparisons of critical features of the data (e.g., cancel times) would be steps in the right direction.

The Wiecki and Frank [28] model of the stop task does not fulfill our desideratum of competitive model testing very well. It demonstrates that the model could work, but it does not pit the model against plausible alternatives. Wiecki and Frank evaluate the effects of lesioning model structures and manipulating motivation, comparing different versions of their model, but the evaluation is qualitative, not quantitative. They also apply the model to related tasks, like the antisaccade task, again evaluating the fit qualitatively.

From the perspective of mathematical psychology, this model does not fare well. There are many parameters and essentially no data points. If Goldilocks were a mathematical psychologist, she would find to focus of this model (on brain regions and not on quantitative data) too big. If Goldilocks were a computational neuroscientist with Wiecki and Franks' perspective, she would find this theory just right.

15.5 Waking Up

It is the dawn of a brand new day. We dreamed our dream and still want more. The integration of mathematical psychology and neuroscience has only just begun. We still dream of a grand model that integrates it all, from spikes to brains, and fits a large amount of data with a small number of parameters. In our view, the frontal cortex-basal ganglia model may be too big, the spiking neuron model may be too small, and the interactive race model may be just right, but we dream of a model that integrates all three. The models have moved us significantly forward, but much remains to be done. In the remaining pages, we sketch out our future dreams and some cold, hard realities that we must face.

15.5.1 Choice

Perhaps the most pressing problem is to deal with choice, both in the go task and the stop task. Boucher et al. [5] considered only one accumulator for the go task. Lo et al. [10] and Wiecki and Frank [28] proposed two accumulators, one for each possible go response, but did not model activity in the competing accumulator. This may be appropriate for saccadic stop-signal tasks, where choice errors are exceedingly rare [7], but it is not appropriate for manual stop-signal tasks, which dominate the literature [26]. The probability and latency of choice errors need to be modeled. The alternative responses must be modeled as stochastic accumulators, and their

interaction with the stochastic accumulator for the correct response must be specified. Race models, feed-forward inhibition models, and lateral inhibition models are viable alternatives [18–20]. Choice tasks provide the opportunity to manipulate several factors that affect the go process concurrently, and these factors may influence different parameters of the go process selectively. Selective influence provides important leverage in modeling: Some parameters should stay constant across conditions while others vary, and this adds important constraints in fitting data. We are currently working on developing models that implement choice in the go task. We recently extended the independent race model to deal with choice in the go task and found some evidence for selective influence [14].

Choice is also possible in the stop process. Several investigators have studied varieties of "selective inhibition," in which some responses but not others must be stopped when a stop signal occurs [2], or all responses must be stopped when a stop signal occurs but not when another similar "ignore" signal occurs [4]. Selective stopping may pose a significant challenge for modeling. Bissett and Logan [4] found that selective stopping to one stimulus but not another often produces violations of the independence assumptions of the race model. This is important because all of the models we have discussed, from Logan and Cowan [13] to Wiecki and Frank [28], assume that the stop process and go process are independent for much of their duration. Independence makes modeling simpler. Non-independent stop and go processes are much harder to characterize. We are beginning to work on models of selective stopping.

15.5.2 Mechanisms of Response Inhibition

The models we discussed consider only one mechanism for inhibiting responses: inhibiting the growth of activation in go accumulators. Other mechanisms have been proposed in the literature and must be distinguished from this one [3, 13]. Salinas and Stanford [21] note that the main computational requirement for a mechanism of response inhibition is to halt or reverse the growth of activation in the go accumulator. They propose a generic model that halts and reverses the growth but do not commit to the underlying mechanism. In their view, it need not be inhibition. In our view, which mechanism underlies inhibition is an empirical question, which we are invested in answering.

Logan and Cowan [13]; also see [3] proposed a *blocked input* mechanism for countermanding responses. They suggest that go responses are driven by input from perceptual systems, and go responses can be countermanded by blocking the input to the motor system. The input can be blocked in several ways. One possibility is deleting the goal to act. In production system models, action depends on two conditions: a goal and an appropriate stimulus. The action can be countermanded by removing the goal, by removing the stimulus, or by removing both. Another way to countermand responses is to suppress the input from perceptual systems. In stochastic accumulator models, this involves setting the drift rate to zero (or less). A

third possibility is to break the connection between perceptual and motor systems. The mapping of go stimuli onto go responses is often arbitrary (e.g., "press the left key if an X appears") and must be maintained somewhere in the cognitive system [11]. Disabling the mapping rules would prevent the growth of activation in the motor system. In our model of visual search [18, 19], the connection between perceptual and motor activity is controlled by a gate that prevents noise from accumulating in stochastic accumulators. Responses could be countermanded by raising the gate to a much higher level.

Boucher et al. [5] evaluated a blocked input model, in which the drift rate for the go process was set to zero after the stop accumulator reached threshold. They found it did not fit the data as well as their interactive race model. We have re-evaluated the same model and several variants, and we do not replicate Boucher et al.'s findings. In our simulations, the blocked input model fits the data as well as or better than the interactive race model. We are currently working hard on this issue.

15.5.3 Model Mimicry

The models we discussed make very similar predictions for behavior and physiology. Quantitative fits to behavior—inhibition functions and RT distributions for no-stop-signal and signal-respond trials—were equivalent for the Boucher et al. [5] interactive race model, the Boucher et al. stochastic accumulator version of the independent race model, and the Lo et al. [10] spiking neuron model. Even the Wiecki and Frank [28] frontal cortex-basal ganglia model produced the same qualitative trends. Perhaps considering other data sets and more complex experimental designs can break this mimicry. All of the models were fit to a single data set from one monkey from Hanes et al. [8], (Boucher et al. also fit data from another monkey). In most applications of mathematical psychology, this would not be sufficient. However, the goal of these models is to predict behavior and neurophysiology simultaneously, and that requires fitting data sets in which behavioral and neural measures were gathered in the same session in the same subject. So far, the only data that meet this criterion are from Hanes et al. [8] and Paré and Hanes [15]. We are currently working toward gathering behavioral and neural data from monkeys performing a stop-signal task in which we manipulate choice difficulty in the go task.

Neural measures exhibit mimicry too. The Boucher et al. [5] interactive race model and the Lo et al. [10] spiking neuron model predict similar modulation of activity in movement and fixation neurons and predict similar distributions of cancel times. Our current investigations of blocked input models mimic these predictions. Moreover, the stochastic accumulator version of the independent race model that Boucher et al. investigated could predict similar modulation and cancel time distributions if the measures were defined a little differently. The modulation of go activation could be defined as the maximum value of the go accumulator on signal-respond trials. With that definition, the independent race model would modulate much like the interactive race model. It would stop before it reached threshold, and the level of activation

would be lower the shorter the SSD, like the observed data. Similarly, cancel time distributions could be generated for the independent race model by comparing the maximum activation on signal-respond trials to the activation on latency-matched no-stop-signal trials. These cancel times would fall in the range of observed cancel times (i.e., SSRT $\pm$ 50 ms).

The mimicry in the neural measures may be broken by examining different neural measures and examining activation and modulation quantitatively (e.g., [18, 19]). For example, Pouget et al. [16] compared baseline, onset of growth, growth rate, and threshold measures in neurons on trials that followed stop signal and no-stop-signal trials to determine the cause of slowing after a stop signal. They found the onset of growth changed, but none of the other measures did. Similar measures could be taken for stop-signal and no-stop-signal trials to determine the cause of stopping. The neural measures could be compared with measures taken from simulations of the models to determine which model provides the best account of the physiology. However, model simulations suggest that such measures may not always agree well with the values of the parameters that generated them, especially if there is noise in data and the model predictions, and there always is. Measured onsets do not always correspond to non-decision times, measured rates of growth do not always correspond to drift rates, and measured thresholds do not always correspond to model thresholds [17].

15.5.4 Fitting Behavior and Physiology Simultaneously

Our strategy has been to fit models to behavioral data and then use the best-fitting parameter values to generate predictions for neural measures. The virtue of this strategy is that the predictions are genuine predictions. No further adjustment of the parameters is required or allowed to generate the predictions. We find it impressive that the predictions are so close to the observed data. However, this strategy requires an arbitrary and artificial distinction between behavioral and neural data. One is used to fit the model and the other is used to test its predictions about the dynamics of the units embodying the model. We are currently searching for methods that allow us to fit behavioral and neural data simultaneously, giving each equal weight in assessing goodness of fit (e.g., [24]). Those methods promise a true integration of mathematical psychology and neuroscience—a dream worth waking up to.

Exercises

1. In what sense does the stop signal paradigm measure response inhibition?
2. The independent race model addresses finishing time distributions without specifying the processes that generate the finishing time distributions. How is this an advantage and how is this a disadvantage.

3. The interactive race model does not describe neural activity at the beginning of the trial when the eyes are fixated. During this period, fixation cells are active and their firing rate is stable. After the target appears, activity in fixation-related neurons drops and activity in movement-related neurons increases. Do you think that including this fixation activity at the beginning of a trial in the modeling will change the models' predictions? How?
4. The spiking neuron model assumes a control process that removes inhibition on the go process to generate a response. What problems do you see with this assumption?
5. The cortico-basal ganglia model has not been tested with rigorous fits to data. Do you think it would fit well if such fits were attempted?

Further Reading

Logan and Cowan [13] is a seminal paper in stop-signal modeling. Everyone who works with the stop signal task should be familiar with this model and the approach.

Logan [12] is a "user friendly" introduction to the stop signal task that may be more accessible to novice readers than Logan and Cowan [13].

Boucher et al. [5] is the first model to bring computational modeling and neurophysiology together in the stop-signal paradigm and is worth reading for its place in history.

Verbruggen and Logan [26] provide a useful but brief review of recent research on the stop signal task. Verbruggen and Logan [27] provide a review of recent modeling work on the stop-signal paradigm.

Anything by Kurt Vonnegut Jr. or Robertson Davies.

Acknowledgement This work was supported by NIH grant R01EY021833

References

1. Aron AR, Duston S, Eagle DM, Logan GD, Stinear CM, Stuphorn V (2007) Converging evidence for a fronto-basal-ganglia system for inhibitory control of action and cognition. J Neurosci 27:11860–11864
2. Aron AR, Verbruggen F (2008) Dissociating a selective from a global mechanism for stopping. Psychol Sci 19:1146–1153
3. Band GP, van Boxtel GJ (1999) Inhibitory motor control in stop paradigms: revew and reinterpretation of neural mechanisms. Acta Psychol 101:179–211
4. Bissett PG, Logan GD (2013) Selective stopping? Maybe not. J Exp Psychol Gen (in press)
5. Boucher L, Palmeri TJ, Logan GD, Schall JD (2007) Inhibitory control in mind and brain: an interactive race model of countermanding saccades. Psychol Rev 114:376–397
6. Frank MJ (2006) Hold your horses: a dynamic computational role for the subthalamic nucleus in decision making. Neural Netw 19:1120–1136
7. Hanes DP, Schall JD (1995) Countermanding saccades in macaque. Vis Neurosci 12:929–937

8. Hanes DP, Patterson WF, Schall JD (1998) Role of frontal eye field in countermanding saccades: visual, movement and fixation activity. J Neurophysiol 79:817–834
9. Lo CC. Wang XJ (2006) Cortico-basal ganglia circuit mechanism for a decision threshold in reaction time tasks. Nature Neurosci 9:956–963
10. Lo CC, Boucher L, Paré M, Schall JD, Wang XJ (2009) Proactive inhibitory control and attractor dynamics in countermanding action: a spiking neural circuit model. J Neurosci. 29:9059–9071
11. Logan GD (1979) On the use of a concurrent memory load to measure attention and automaticity. J Exp Psychol Hum Percept Perform 5:189–207
12. Logan GD (1994) On the ability to inhibit thought and action: a users' guide to the stop signal paradigm. In: Dagenbach D, Carr TH (eds) Inhibitory processes in attention, memory, and language. Academic, San Diego, pp 189–239
13. Logan GD. Cowan WB (1984) On the ability to inhibit thought and action: a theory of an act of control. Psychol Rev 91:295–327
14. Logan GD, Van Zandt T, Verbruggen F, Wagenmakers EJ (2014) On the ability to inhibit thought and action: general and special theories of an act of control. Psychol Rev 121:66–95
15. Paré M, Hanes DP (2003) Controlled movement processing: superior colliculus activity associated with countermanded saccades. J Neurosci 23:6480–6489
16. Pouget P, Logan GD, Palmeri TJ, Boucher L, Paré M, Schall JD (2011) Neural basis of adaptive response time adjustment during saccade countermanding. J Neurosci 31:12604–12612
17. Purcell BA (2013) Neural mechanisms of perceptual decision making. Doctoral Dissertation, Vanderbilt University
18. Purcell BA, Heitz RP, Cohen JY, Schall JD, Logan GD, Palmeri TJ (2010) Neurally constrained modeling of perceptual decision making. Psychol Rev 117:1113–1143
19. Purcell BA, Schall JD, Logan GD, Palmeri TJ (2012) From salience to saccades: multiple-alternative gated stochastic accumulator model of visual search. J Neurosci 32:3433–3446
20. Ratcliff R, Smith PL (2004) A comparison of sequential sampling models for two-choice reaction time. Psychol Rev 111:333–367
21. Salinas E, Stanford TS (2013) The countermanding task revisited: fast stimulus detection is a key determinant of psychophysical performance. J Neurosci 33:5668tb–5685
22. Schall JD (2004) On building a bridge between brain and behavior. Annu Rev Psychol 55:23–50
23. Teller DY (1984) Linking propositions. Vis Res. 24:1233–1246
24. Turner BM, Forstmann BU, Wagenmakers EJ, Brown SD, Sederberg PB, Steyvers M (2013) A Bayesian framework for simultaneously modeling neural and behavioral data. Neuroimage 72:193–206
25. Usher M, McClelland JL (2001) The time course of perceptual choice: the leaky, competing accumulator model. Psychol Rev 108:550–592
26. Verbruggen F, Logan GD (2008) Response inhibition in the stop-signal paradigm. Trends Cogn Sci 12:418–424
27. Verbruggen F, Logan GD (2009) Models of response inhibition in the stop-signal and stop-change paradigms. Neurosci Biobehav Rev 33:647–661
28. Wiecki TV, Frank MJ (2013) A computational model of inhibitory control in frontal cortex and basal ganglia. Psychol Rev 120:329–355

Chapter 16
Reciprocal Interactions of Computational Modeling and Empirical Investigation

William H. Alexander and Joshua W. Brown

Abstract Models in general, and computational neural models in particular, are useful to the extent they fulfill three aims, which roughly constitute a life cycle of a model. First, at birth, models must account for existing phenomena, and with mechanisms that are no more complicated than necessary. Second, at maturity, models must make strong, falsifiable predictions that can guide future experiments. Third, all models are by definition incomplete, simplified representations of the mechanisms in question, so they should provide a basis of inspiration to guide the next generation of model development, as new data challenge and force the field to move beyond the existing models. Thus the final part of the model life cycle is a dialectic of model properties and empirical challenge. In this phase, new experimental data test and refine the model, leading either to a revised model or perhaps the birth of a new model. In what follows, we provide an outline of how this life cycle has played out in a particular series of models of the dorsal anterior cingulate cortex (ACC).

16.1 Introduction

A popular, though probably apocryphal, characterization of the geocentric model of the solar system is that, before it was replaced by the heliocentric model, it required "epicycles on epicycles on epicycles" in order to describe the movement of the planets and stars through the sky. Initial attempts explained the path of these heavenly bodies as revolving about the earth at a fixed distance along celestial spheres. While this simple model was sufficient to describe a majority of the data available, it was observed that certain planets exhibited retrograde motion, appearing to reverse the

J. W. Brown (✉)
Department of Psychological and Brain Sciences,
Indiana University, Bloomington, USA
e-mail: jwmbrown@indiana.edu

W. H. Alexander
Department of Experimental Psychology,
Ghent University, Henri Dunantlaan 2, B-9000 Gent, Belgium
e-mail: william.alexander@ugent.be

B. U. Forstmann, E.-J. Wagenmakers (eds.), *An Introduction to Model-Based Cognitive Neuroscience,* DOI 10.1007/978-1-4939-2236-9_16

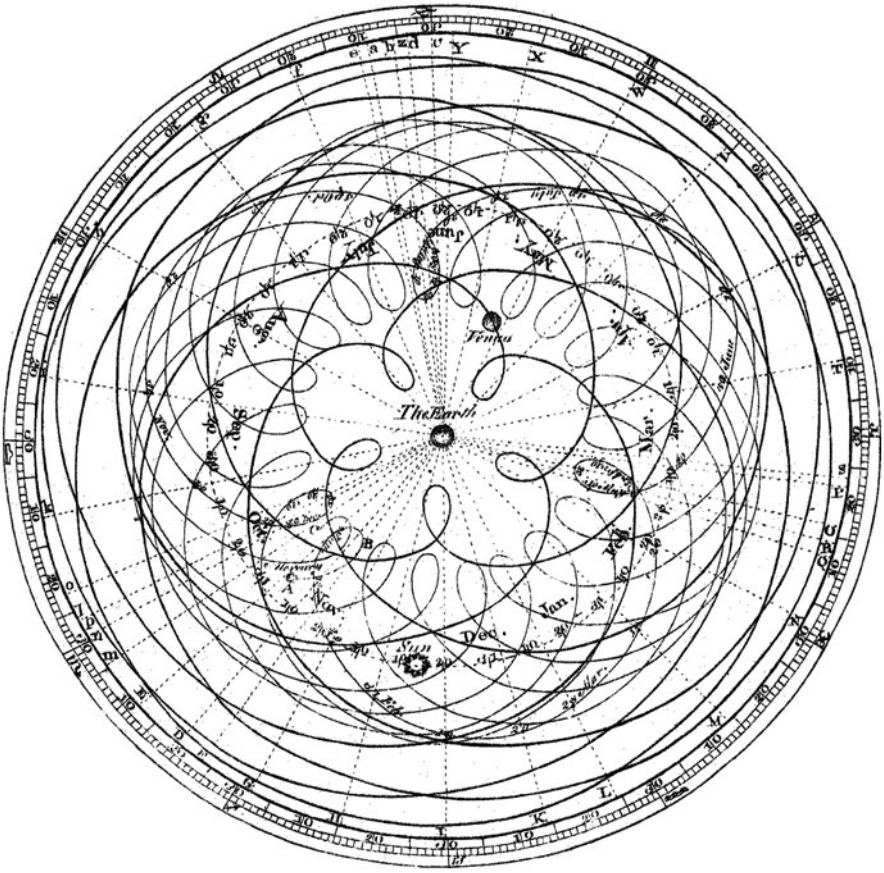


Fig. 16.1 The Ptolemaic model of the solar system. (Public domain)

direction of their path at certain points. In order to account for these changes, Ptolemy introduced epicycles into the geocentric model: in addition to following a circular path around the earth, planets also followed a second revolution around that path. Although epicycles could explain the apparent retrograde motion, the system was imperfect. In order to explain further anomalies between the Ptolemaic system and observations of the planets, the story goes, it was necessary to include ever more epicycles into the model, multiplying the complexity of the system for only modest gains in explanatory ability. Ultimately, as Kuhn argued, scientific progress came as the Copernican heliocentric model replaced the Ptolemaic model in a revolutionary (rather than evolutionary) way [1] (Fig. 16.1).

In a comparable manner to the Ptolemaic model of the solar system, the advent of sophisticated brain imaging techniques has advanced empirical knowledge at a pace that has outrun model building. Rather than epicycles, however, neuroimaging studies appear to assign ever more functions and modules to areas of the brain that show

increased BOLD activity ($p < 0.05$, corrected or not) for one condition over another. Early work in cognitive neuroscience sought to identify areas of the brain supporting cognitive processes whose existence had been inferred through psychological experimentation [2], and exuberant studies in scholarly journals proclaimed on a weekly basis that the area of the brain underlying a highly specific cognitive function had been identified. As research progressed, and the number of independent functional modules in the brain proliferated, a kind of weary cynicism set in, leading some to regard the new-fangled research methods as a modern form of phrenology. Although accusations regarding its status as a pseudoscience may be premature, a significant challenge to the field of cognitive neuroscience is presented by the seemingly endless supply of new data that, while saying much, explain little. The number of distinct effects observed under various experimental paradigms has made the brain appear to be a very crowded place indeed, seeming to include regions coding for everything up to and including the proverbial kitchen sink.

One region of the brain in particular, the anterior cingulate cortex (ACC), has become associated with this kitchen-sink effect [3]. ACC, by virtue of its high interconnectivity, is promiscuously active in almost any task that involves some level of engagement and action. It has been noted that the rate at which cingulate activity is reported in fMRI studies has increased exponentially since the 1990s, and it is projected that by the end of the Twenty-first century neuroscience will achieve the "cingularity", the point at which there are more scholarly works investigating cingulate activity than there are cells in the cingulate itself [4]. Although the Gage, Parikh and Marzullo article is emphatically tongue in cheek, the authors are not far off the mark when they note that, given the diversity of functions that have been attributed to the cingulate, it appears that this region of the brain does everything. Functions attributed to ACC include detection and processing of error [5, 6], resolving behavioral conflict [7, 8], detecting and predicting reward [9, 10], anticipating and indicating painful stimuli [11, 12], signaling negative affect [13], deploying attention [14], learning the value of actions [15, 16], and a host of others.

Investigation of the function of ACC has been of particular interest in the area of cognitive control. In typical cognitive control tasks, subjects are required to inhibit a prepotent, stimulus-driven response in favor of a less-automatic response. A classic example is the Stroop task [17], in which subjects are presented with color words that are displayed in various font colors. The subject is instructed to indicate the color in which the word is written, and to ignore the denotative meaning of the word itself. For trials in which the meaning of the word and the font color both indicate the same response ('congruent'), the task is trivially easy. However, on trials in which the meaning of the word and the font color differ ('incongruent'), successful performance of the task requires the subject to make only one of two cued responses. The processes by which an individual interprets stimuli in order to select and execute a response are collectively referred to as cognitive control.

16.2 The Conflict Model

One highly influential interpretation of ACC activity is the conflict monitoring model [7, 8]. In this interpretation, task stimuli which cue multiple, mutually incompatible responses induce a state of conflict that needs to be resolved in order to successfully generate appropriate responses. ACC activity indexes conflict as the summed multiplicative interaction of cued responses

$$\text{Conflict} = \Sigma W_{ij} a_i a_j \tag{16.1}$$

where a_i and a_j are neural activities representing mutually incompatible response cues, and W_{ij} represents the degree of mutual incompatibility between them. In the case of the Stroop task, when only one response is cued on congruent trials, conflict, and by extension ACC activity, remain low. For incongruent trials, two competing responses are cued, resulting in increased ACC activity.

Although the conflict model is notable for its ability to account for a range of effects observed in ACC from EEG and fMRI studies, the model does not address how the arbitrary stimuli used in many cognitive control tasks might come to be associated with conflicting behavioral responses. Conflict in the Stroop task, for example, only exists due to the demands placed upon the subject to respond only with the color of the font a word is printed in rather than with the color denoted by the word itself. If the subject were instead instructed that it would be acceptable to make a response indicating either the word identity OR the font color, it is difficult to imagine how this would lead to a state of behavioral conflict. In short, conflict is not a property that inheres in external stimuli, but one that is constructed through the interaction of stimuli and the context in which they are experienced.

Situations in which a particular stimulus can cue entirely different behavioral responses are prevalent in day-to-day life. Take the example of encountering a stop sign as one is traveling along a road. Depending on the particular mode of transport, the responses one needs to generate in order to comply with the denotative meaning of the sign may vary a great deal. If one were traveling by bicycle, for instance, the appropriate response would be to apply brakes by operating levers located on both hand grips. If one were traveling by motorcycle, however, operating the controls on the handlebars would result either in increased speed or disengaging the clutch, neither of which would be optimal for coming to a stop.

It is easy to imagine that a proficient bicyclist who is learning how to ride a motorcycle may experience a high degree of conflict as she adjusts to the differences between the two modes of transport. Initial attempts to stop while on a motorcycle might require increased vigilance and attention (i.e., cognitive control) to ensure that the appropriate response is made. As the individual gains experience, however, there is less need to exercise control when switching from a bicycle to a motorcycle, implying that information regarding conflicting responses can be learned. In the case of our novice motorcyclist, the prepotent response when confronted with a stop sign, trained through many years of bicycling, is to apply hand brakes. When riding a motorcycle, as outlined above, this response leads to a sub-optimal result

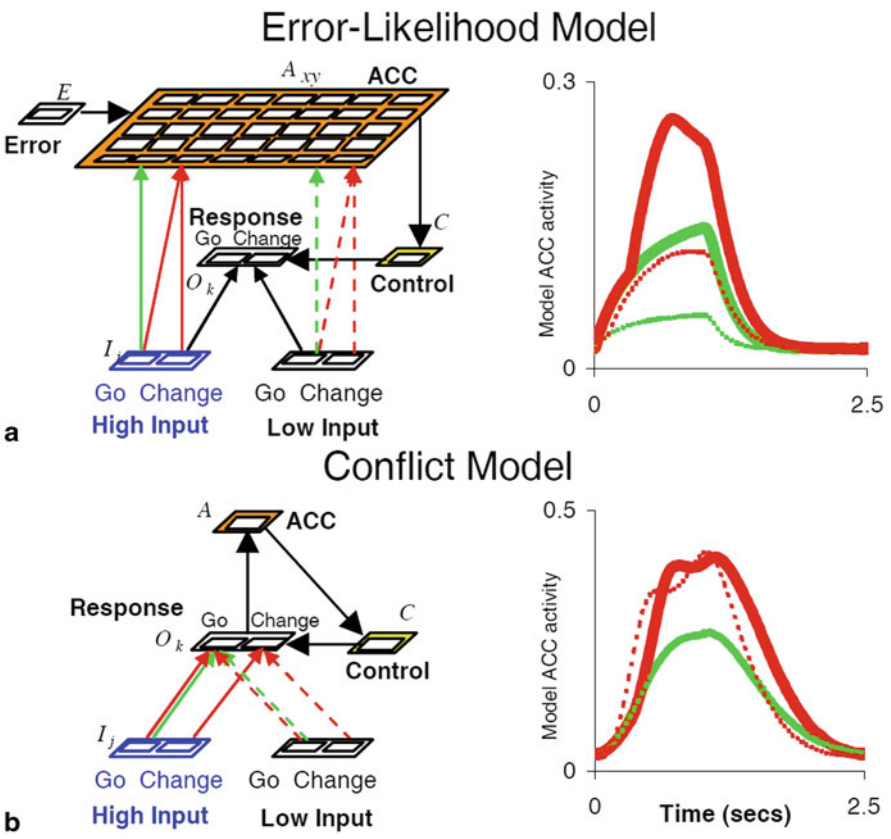


Fig. 16.2 The error likelihood model, as contasted with the conflict model in a change signal task. *Red curve*: conflict. *Green curve*: no-conflict. *Solid line*: high error likelihood. *Dashed line*: low error likelihood. (Adapted by permission of the AAAS from [18])

("not stopping"), and thus constitutes a behavioral error. The intuition, then, is that high-conflict situations, such as choosing between using hand or foot controls on a motorcycle, are those that are associated with an increased likelihood of error, rather than conflict *per se*, because a particular combination of movements may be mutually incompatible in one context but not in another.

16.3 Stage 1: Birth—The Error Likelihood Model

The error likelihood model of ACC [18] (Fig. 16.2) proposed a mechanism by which information regarding behavioral error may contribute to ACC activity. The principal component of the model consists of a self-organizing map (SOM), representing ACC, that receives excitatory projections from representations of task-related stimuli.

Initially, adjustable weights representing the excitatory influence of stimuli on ACC are low. Over the course of experience with a task, the model learns to associate single units in the SOM with the presentation of a particular stimulus. This association is learned by the hypothetical mechanism of dopaminergic (DA) disinhibition of ACC. Tonic activity of midbrain DA neurons may actively inhibit neurons in ACC; transient depressions in baseline DA activity, associated with negative reward prediction errors [19], disinhibit ACC and contribute to Hebbian-type learning between active stimulus representations and stochastically active single units in the SOM. Over the course of training, units within ACC are activated by stimulus representations in proportion to the frequency with which each stimulus is associated with error.

Although the intuition underlying the error likelihood model, that stimulus-dependent behavioral conflict may be learned, is largely consistent with the conflict monitoring theory of ACC, simulations of the two models revealed divergent predictions regarding ACC activity for contexts in which behavioral conflict is absent (Fig. 16.2) yet the likelihood of behavioral error differs between conditions. This discrepancy between the two models informed the design of a new behavioral task designed to investigate whether ACC activity reflected differences in anticipation of error. In the change signal task, subjects are asked to respond according to the direction indicated by an arrow presented to them. For most trials ("Go" trials), the cue remains valid and the subject makes the initially cued response. On a subset of trials ("Change" trials), however, the presentation of the arrow is followed by the presentation of a second arrow, indicating that the subject should cancel the initial response and instead make the alternate response. The timing between the presentation of the first and second arrows can be manipulated to enforce specific error rates, and the color of the arrows serves as an implicit cue indicating whether there is a high or low likelihood of committing an error.

Both the error likelihood and conflict models predict increased ACC activity for Change trials in which behavioral conflict is present. However, for Go trials, in which conflict is absent, the error likelihood model predicts increased ACC activity for conditions in which the likelihood of an error is higher, as indicated by the color of the arrow. These predictions were tested using fMRI, and revealed that, consistent with the error likelihood model, ACC activity was significantly greater for trials with a high likelihood of error, despite the absence of conflicting cues [18]. This suggested that the conflict model was incomplete, as it could not account for the error likelihood effect, and that the error likelihood model may provide a more complete account of the data. We note however that the conflict model has had (and continues to have) a very useful life—it provides a simple, elegant account of existing phenomena, it provided testable predictions, and it provided inspiration for subsequent model development, in this case the error likelihood model. Similarly, the error likelihood model was inspired by an earlier model in which dopaminergic reward omission signals drive activity in the ACC [19].

16.4 Stage 2: Maturity—Empirical Tests of the Error Likelihood Model

While the error likelihood model is able to account for observed error likelihood effects within ACC, the hallmark of a mature model lies in its ability to guide further empirical investigation. Additional simulations of the error likelihood model revealed additional predictions regarding ACC activity that directly informed a new set of experiments. We reasoned that if ACC activity reflects the prediction of an error rather than conflict, then perhaps two different response cues would lead to greater ACC activity regardless of whether or not they led to a state of conflict. We tested this prediction and found evidence in favor of it [20]. Still, while the error likelihood model accounts for a number effects observed within ACC, additional empirical evidence suggests that the error likelihood model itself is incomplete. The first challenge was an apparent failure to replicate the error likelihood effect empirically. A subsequent paper tested for error likelihood effects with fMRI and ERP and reported a null result [21]. This challenged us to ask whether individuals differ in their sensitivity to likely errors, and whether those who are more sensitive to error likelihood at the neural level may likewise tend to be more careful to avoid errors and risky behavior in general. We tested this empirically and found it to be the case [22], and this also led to a refined error likelihood model in which the learning rate from errors could vary across the population [23].

A second challenge to the error likelihood model consists of surprise effects, e.g., the finding that unexpected errors lead to greater ACC activity than less surprising errors [24]. The error likelihood model simply could not account for this effect, because the error likelihood signals constituted a prediction that did not depend on the actual outcome, but the surprise effect depends heavily on the nature of the outcome, e.g., an error. Furthermore, we and others found that the well-known error effects in ACC can actually invert themselves, so that when errors are more likely than correct outcomes, then the correct outcomes yield greater ACC activity than errors [25–27]. Still more challenges confronted the error likelihood model. More recent papers argued that error likelihood effects did not exist or were subsumed by simple correlations between ACC activity and response time [28, 29]. We also found that unexpected delays in feedback could lead to ACC activity, which suggested that the timing of feedback was as important as its valence [30].

Meanwhile, monkey neurophysiology studies of ACC provided their own challenge to both the conflict and error likelihood models. In an earlier study, we failed to find evidence of conflict monitoring cells in ACC, despite having recorded over 450 cells [31]. Instead, ACC cells generally seemed to reflect the value of actions, integrate recent reward history [32], indicate the value of explorative vs. exploitative behavior [33], and signal the prediction [9, 34] and detection [31, 35] of reward. These findings taken together present a significant challenge to the effort to develop a unifying theory regarding ACC activity across species, and, indeed, have been taken as evidence that no such unification is possible [36]. Nevertheless, recent studies

[37] have sought to test this idea by recording BOLD activity in monkeys performing tasks that have been observed to produce conflict-type effects in humans. These studies show increases in BOLD response in similar regions of monkey and human brains commonly associated with cognitive control, including ACC, providing evidence that ACC in monkey and human are functionally similar. This suggests that a reconciliation of human and monkey ACC results is possible.

16.5 Stage 3: Dialectic Methods of Model Development

Confronted with mounting evidence that the error likelihood model is, at best, an incomplete account of ACC function, our goal was to develop a new computational account that was *less incomplete*. In doing so, we sought to resolve two dialectic tensions, that of the theoretical-empirical dialectic, and the empirical-empirical dialectic. In the first, we questioned whether the theory underlying the error likelihood model was disproved by contradictory evidence, or whether minimal extensions could expand the theory without becoming a process of adding ever more epicycles. In the second, we question how empirical evidence in apparent contradiction with itself might reflect a common underlying process.

16.5.1 The Theoretical-Empirical Dialectic

In the face of empirical findings that challenge one's model, what is a modeler to do? One of the biggest challenges we face as modelers, aside from the challenges of developing models, is the parental affection that we feel towards the models we develop. Our modeling efforts have been guided in this regard by a pair of classic works that should be required reading for all scientists, and especially modelers. The first of these papers argues that we must beware of parental affection for a theory—rather than trying to extend the reign of an aged monarch of a theory, or even favoring a working hypothesis, we should entertain multiple competing hypotheses and sincerely ask which is the best account of the phenomenon, even if this means abandoning our own theoretical progeny [38]. This may seem an obvious point, but parental affection, and an associated desire to guard one's reputation as a modeler, is likewise a strong and pervasive instinct. The second paper argues that the best experiment is not one that is designed to prove a theory or model, but rather one that is designed to *dis*-prove a theory. As Platt puts it, our experiments should answer "the Question", which is this: What model or models do your experiments rule out [39]? In this way, as the space of plausible models is reduced, a field may converge on a more faithful model of the phenomena in question. All of this requires that parental affection for one's own theory must be set aside. Thereafter, the modeler is free to embrace the dialectic of Hegel, to wrestle with both the theory and the empirical

results, and find a synthesis that moves the field forward toward a new generation of model.

16.5.2 The Empirical-Empirical Dialectic

A key process in model building is finding ways to reconcile and integrate apparently contradictory findings. It is especially conducive to progress if relevant data can be considered from a range of species (human, monkey) and modalities (behavior, fMRI, ERP, single unit neurophysiology, etc.). In the case of the ACC, if the function is indeed similar across species, how does one reconcile, for example, the apparent involvement of single neurons in ACC in reward prediction and processing with the apparent involvement of the region as a whole with principally error prediction and processing? We approach this overarching question by addressing two related questions. First, why are single neurons whose activity reflects reward-related processes observed many times more frequently than neurons whose activity reflects error and error prediction? Recall that in the error likelihood model, simulated ACC activity prior to the presentation of task-related feedback is proportional to the frequency with which errors are observed for a given stimulus context. Individual units in the SOM representing ACC in the error likelihood model, analogous to single neurons within ACC, are selectively associated with task stimuli that predict error; at the limit, as the number of trials goes to infinity, the entire SOM is partitioned such that the number of units responding selectively to a particular stimulus will be proportional to the frequency with which errors are observed for that stimulus relative to other task stimuli. If ACC neurons were in fact learning associations between stimuli and subsequent error alone, one would expect that single-unit neurophysiology studies would find ubiquitous error-related neurons.

One possible explanation for the predominance of reward-related neurons observed in single-unit studies, in the framework of the error likelihood model, is that *single units within an SOM learn to associate task related stimuli not only with error, but also with correct outcomes.* In the context of the literature on cognitive control, both with monkeys and humans, experimental conditions are generally arranged such that trials on which errors are observed generally represent a minority of the total number of trials in an experiment. While this explanation may account for the prevalence of reward-related neurons within ACC, it poses an additional challenge to the error likelihood model. The error likelihood model learns to associate task cues with subsequent error through the putative mechanism of dopaminergic disinhibition. In this scheme, ascending projections from midbrain DA neurons tonically inhibit activity within ACC. The occurrence of an error produces a transient decrease in the baseline firing rate of DA neurons, resulting in disinhibition of ACC, with an attendant increase in activity in neurons that receive excitatory input from representations of task-related stimuli, resulting in Hebbian-type learning of conjointly active units in ACC and units representing task-cues. Since the activity of DA is generally

depressed only when an outcome is worse than expected (e.g., withholding of a predicted reward), this mechanism cannot explain how neurons in ACC may come to represent reward.

Regardless of the specific mechanisms contributing to the distribution of reward and error-related single neurons within ACC, a second question concerns the time course of neural activity observed within the region. Specifically, is reward and error related activity in ACC predictive in nature, or is the region involved principally in evaluation of actual outcomes? In this regard, evidence from single-unit neurophysiology and fMRI/EEG studies alike is mixed. By far the most robust effect observed in ACC from fMRI and EEG data is that of error. ACC was initially implicated in the processing of behavioral error based upon observations of a negative-going ERP component specifically on trials in which an error occurred, the well-known error-related negativity (ERN) [5, 6] an effect which has since been replicated numerous times across different recording techniques. Additionally, single neurons in monkey have been observed whose activity is specific to the delivery of reward or the occurrence of an error. Together, these findings suggest that ACC activity is primarily evaluative in nature, and depends on feedback from the environment or a subject's own behavior to elicit a response. On the other hand, substantial evidence suggests that ACC activity frequently anticipates or precedes feedback of this nature. Neurons in monkey ACC have been observed whose activity increases as an expected reward draws temporally closer, indicating a role for the area in prediction [34, 40]. Similarly, findings such as the error likelihood effect, wherein differences in activity are observed following a cue that is associated with future outcomes, provide evidence for ACC's involvement in prediction.

Previous theories of ACC have variously assigned primarily evaluative or predictive functions to the region. Evaluative theories of ACC have suggested it signals the discrepancy between actual and intended responses [41], discrepancies in value [19], corrective actions following behavioral error [42], or the multiplicative interaction of multiple cued responses in the conflict model [8, 43]. Although these theories can be described as evaluative in the sense that the computations necessary to perform them operate on information that is present within the system, rather than information that is anticipated but not present, it is not clear that they distinguish between anticipatory and reactive activity in ACC. In the case of Scheffers and Coles [41] and Steinhauser et al. [42], ACC activity is reactive in nature, depending on the comparison between responses or detection of sequences that have already occurred. In contrast, the conflict model suggests ACC activity may be observed prior to the actual generation of a response as two incompatible responses compete. Similarly, the RL-ERN theory [44] suggests that ACC activity may precede the actual generation of a response due to value discrepancies induced by changes in response unit activity in the model. For all of these models, however, the primary computation performed by ACC explicitly evaluates present information; anticipatory or predictive activity is merely implicit, reflecting continuous, ongoing evaluation. In contrast, the error likelihood model is based on explicit prediction; activity prior to feedback reflects the likelihood of an event (behavioral error) that has yet to be experienced, while increased activity in the model following incorrect performance implicitly indicates error as the disinhibition

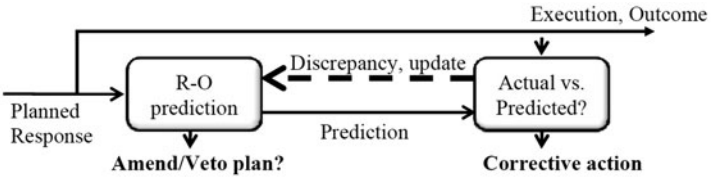


Fig. 16.3 The PRO model, showing distinct response-outcome (R-O) prediction unit (*left*), and evaluation unit (*right*). (Reprinted by permission of the Cognitive Science Society, from [45])

of predictive units. More generally, we find that theories regarding ACC function can be categorized as being either explicitly evaluative/implicitly predictive, or explicitly predictive/implicitly evaluative. A notable lack, therefore, in theorizing about ACC, is the possibility that *ACC activity reflects both explicit predictive and evaluative functions*.

16.6 The PRO Model—A Synthesis

16.6.1 PRO Model Stage 1: Birth

In the preceding section, we identified two questions that informed our thinking regarding the function of ACC in the context of cognitive control and decision making. First, how might individual neurons in ACC become associated principally with the anticipation and detection of reward? And second, is activity in ACC consistent with either explicit or implicit prediction and evaluation? In both cases, we identified potential solutions to these questions, leading to a concise hypothesis regarding possible ACC function that forms the basis of the *predicted response-outcome* (PRO) model (Fig. 16.3), that ACC learns explicit predictions regarding future outcomes, regardless of their affective valence, and signals unexpected deviations from predicted events [45, 46].

Similar to previous models of ACC, the PRO model builds on standard models of reinforcement learning, and extends these models in two ways. First, standard RL is concerned with learning an optimal behavioral policy given an underlying value function, with the goal of selecting at each opportunity the behavior that leads to the highest long-term value. In typical formulations, value is represented as a scalar quantity, reflecting the average reward (positive or negative) that can be obtained from a given state. In contrast, the PRO model learns separate predictions of likely conjunctions of responses and outcomes, regardless of whether the outcome is affectively positive or negative. The rationale behind this extension is that in the context of cognitive control, behavior is not selected, as such, but is governed by the interaction of prepotent stimulus-driven behaviors with top-down goals that may conflict with automatic responses. It is not, for example, sufficient to learn that naming the color of the font is more valuable than reading the word in the Stroop task, since value is

only one factor in determining behavior. In order to successfully deploy cognitive control, it is necessary to learn the likelihood that a certain, automatic response will be produced, independent of how valuable that response may be. Secondly, and related to the first point, is that while prediction errors in standard RL reflect better-than or worse-than expected outcomes, prediction errors in the PRO model indicate the extent to which an observed outcome was expected to occur (again, independent of the affective valence of the event itself). In standard RL both the occurrence of a reward that was not predicted as well as the non-occurrence of a predicted aversive stimulus both result in positive prediction errors—better-than expected outcomes. In the PRO model, these outcomes result in positive and negative prediction errors, respectively, indicating an outcome occurred that wasn't expected (positive surprise) and an outcome that was predicted failed to occur (negative surprise).

Although the PRO model departs from the error likelihood model in adopting a RL-based formulation, a key intuition underlying the error likelihood model, that ACC activity reflects the frequency with which errors are observed in a given stimulus context, is preserved. The PRO framework extends this intuition by allowing for the representation not only of error frequency, but also the frequency with which correct trials are observed. The activity of individual units within the PRO model, each representing a specific response-outcome conjunction, reflects the frequency with which that conjunction is observed, analogous to the SOM that formed the basis of the error likelihood model. Indeed, for typical cognitive control tasks in which correct trials are more frequently observed than error trials, activity in prediction-related units is greater than activity in units related to predicting error outcomes. This aspect of the PRO model corresponds well with the intuition developed in the previous section regarding how ACC neurons may predominantly represent reward and reward-related processes: in the course of experience with a task, neurons learn to represent events based on the frequency with which those events are observed.

16.6.2 Reconciling Existing Data

How then can the PRO model simultaneously account for the predominance of reward-related neurons in ACC as well as effects of error, error likelihood, and conflict observed at the level of populations of neurons? Above, we introduced the notions of positive and negative surprise as they relate to standard RL and the extensions introduced in the PRO model. In particular, negative surprise, the unexpected non-occurrence of a predicted outcome, is found to provide a plausible mechanism by which the ensemble activity of reward-related neurons produces apparent error effects. As previously discussed, correct trials are far more frequently observed in typical cognitive control tasks, leading—in the PRO framework—to proportionally greater activity in units predicting reward. Negative surprise, calculated as

$$\omega_t^N = \sum_i MAX\left(Expected_i - Actual_i, 0\right) \tag{16.2}$$

i.e., the predicted outcome minus the observed outcome, suggests that when a highly reliable event such as a correct, rewarded trial fails to occur, observed neural activity should be greater than when a marginally probable event, such as an error, fails to occur. The PRO model thus reinterprets error effects within ACC as the surprising non-occurrence of a predicted (rewarding) outcome. Furthermore, the same logic can additionally reconcile previous findings that did not previously fit within any explanatory framework, including findings of greater activity in ACC for infrequent vs. frequent errors [18, 24], and error effect-like activity for unexpected successes on low probability gambles [26].

Is activity in ACC primarily predictive or evaluative? The negative surprise signal which we deploy to explain the diverse array of effects observed in ACC is, by definition, evaluative in that it compares predicted outcomes to observed outcomes. In order to perform such an evaluation, however, it is necessary that predictions be maintained somewhere within the brain; one possibility is that ACC is involved in both evaluation and prediction, and that these two functions may be either spatially segregated or that individual units performing each function may be intermingled. Another possibility is that extracingulate areas maintain predictions regarding likely outcomes which are used by ACC to calculate the deviation between predicted and observed outcomes. Distinguishing between these two hypotheses may be problematic, however, in that the timed prediction signal assumed by the PRO model as the basis for computing negative surprise is correlated with the negative surprise signal. Following the presentation of a stimulus that reliably predicts future outcomes, both the prediction and the negative surprise signal are identical. This follows from equation (2), in which negative surprise is calculated as the current predicted outcome ($Expected_i$) minus the current observed outcome ($Actual_i$); prior to the occurrence of an outcome, $Actual_i = 0$, and negative surprise is equal to the current prediction. Following the occurrence of an outcome, however, the signals are predicted to diverge, particularly in the case of the occurrence of a likely outcome, where *Expected* and *Actual* values are similar, resulting in low negative surprise but high predictive activity. This may suggest one manner in which areas of the brain involved distinctly in prediction, and not evaluation, may be distinguished from areas within ACC whose activity is consistent with the negative surprise signal suggested by the PRO model.

16.6.3 PRO Model Stage 2: Maturity—Motivating New Experiments

A key goal for computational models is not only to account for previously observed data, but also to suggest additional research questions which may be empirically tested. In both respects, the PRO model has performed well. The PRO model has motivated tests of surprise effects in ACC by others [25], as well as a number of experiments in our own research group. In work currently in preparation, we identify overlapping areas in ACC that reflect the surprising absence of a predicted painful stimulus as well as the unexpected non-occurrence of a difficult cognitive task. In additional work, we find that ACC activity in substance-dependent individuals, rather

than reflecting an overall lack of engagement in cognitive tasks [47], is best explained as increased attention to rewarding outcomes relative to aversive outcomes, biasing the overall predictions learned by the PRO model to reflect attenuated predictions regarding possible negative consequences of one's actions. A recent study also suggests that predictive signals within the ACC can be distinguished from evaluative processes [48]. All of these studies were motivated in part by the predictions of the PRO model.

16.6.4 A Unifying Framework

In the PRO formulation, activity in ACC is associated with the relatively straightforward mechanism of "negative surprise." While we show that this mechanism does well at accounting for observed effects in ACC in the context of cognitive control, ACC is a promiscuous region of the brain, being implicated in social interaction, affective and emotional response, and processing aversive stimuli. An open question, then, is whether the simple mechanism of negative surprise can be applied more generally to questions beyond the specific area of cognitive control. Put another way, is negative surprise one of several specialized functions, each devoted to a particular cognitive function, or does ACC implement some form of negative surprise across the many cognitive functions it is involved in?

Preliminary evidence suggests that the PRO model may provide a unifying framework for understanding ACC function not only across different recording methodologies, but also across different subdisciplines of neuroscience. Since its publication, the PRO model paper [46] has been cited in studies taking affective, social, and clinical neuroscience perspectives. Previously, ACC has been variously attributed roles in the processing of painful stimuli, indicating negative affect related to social exclusion, and disengagement of cognitive control in substance-dependent behaviors. Under the interpretation of the PRO model, however, these effects may reconciled as reflecting facets of a single underlying mechanism, that of negative surprise. Single unit neurophysiology and fMRI studies have identified neurons in ACC, as well as ensemble activity, showing increased activity in the region when monitoring the decisions made by others [49, 50], especially when the outcome of those decisions violates expectations. If we were to extend our notion of what constitutes an outcome to include any predictable event, whether it be terminal feedback at the end of a trial, or intermediate states such as the appearance of additional stimuli, we might expect ACC activity to more generally to reflect predictions about states and to signal surprising state transitions [51]. Simulations of the PRO model incorporating this more general definition of "outcome" suggest that the principle mechanism of negative surprise can account for activity observed in ACC related to the prediction of task-related stimuli, as well as activity related to the onset of such stimuli. Perhaps the best example of this is the mismatch negativity (MMN), a negative ERP component elicited when a periodic stimulus, reliably and repeatedly

presented to a subject, is unexpectedly withheld. EEG studies have identified generators of the MMN in visual and auditory cortex, depending on the modality of the stimulus. More recent studies have identified a generator within medial PFC/ACC, consistent with the anatomical localization of the ERN, and independent of modality.

These findings suggest a general function of ACC in signaling unexpected state transitions, consistent with a potential role in providing learning signals for a model-based RL algorithm implemented across distributed across multiple brain regions. If such is the case, we might expect the activity of areas of the brain with a high degree of connectivity with ACC to also reflect additional aspects of such an algorithm. In this regard, dorsolateral PFC (DLPFC) is a likely target for additional modeling and empirical efforts. Previous work has identified DLPFC as being critically involved in maintaining task-dependent rules in working memory, deploying top-down control, and, like ACC, activity in DLPFC correlates with state prediction errors (SPEs) [52]. The PRO model provides critical constraints on the types of information and computations that may plausibly be observed within DLPFC. Although the PRO model undoubtedly is incomplete at some level of detail, the central intuition that ACC maintains multiple, simultaneous predictions regarding future events, is more consistent with existing evidence than competing accounts. In much the same way the heliocentric model of the solar system supplanted the over-complex geocentric model through a single, unifying premise, the goal of future work investigating the function of brain is not only to identify new roles to assign to individual regions, but also to reconcile how those roles may fit within a general framework.

16.7 Conclusion

In the above discussion, we have laid out some principles of model development that have served us well, and we illustrate these principles by tracing out three generations of models of the ACC. The themes we have highlighted include the life cycle of a model, the dialectic between competing models, the dialectic between models and data, and the dialectic between apparently contradictory data. All of these can be viewed as processes that lead to progress and the development of better models. In tracing out the example ACC models here, we do not mean to imply that they are the only fruitful approaches to modeling the ACC, as there are other models that we do not have space to treat fully. Neither do we intend to tell a triumphalist story that the PRO model is the best model. While we find that it provides the best account of the data to date, it is still a model, and is therefore by definition incomplete. The PRO model will have served its purpose well if it inspires a new set of experiments, inspires a deeper understanding of the ACC, and ultimately leads to an even better model.

Exercises

1. What is the advantage of having two different computational neural models of a phenomenon rather than one?
2. Read Platt [39] and Chamberlin [38]. How could you apply the methods they advocate to your own research?
3. What is the life cycle of a model as described here?
4. Name and describe two key dialectics in the process of model-building.
5. Which is a better experimental effort: to try to prove that a computational neural model is true, or to try to prove that it is false or incomplete?
6. Read the paper on the PRO model [46], or another current computational neural model. Try to think of an experiment (or existing data) that could falsify or otherwise challenge the model.

Further Reading

Alexander, W. H., & Brown, J. W. (2010). Computational models of performance monitoring and cognitive control. *Topics in Cognitive Science*, 2, 658–677.

Alexander, W. H., & Brown, J. W. (2011). Medial prefrontal cortex as an action-outcome predictor. *Nat Neurosci*, 14(10), 1338–1344. doi:10.1038/nn.2921

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. C. (2001). Conflict monitoring and cognitive control. *Psychological Review*, 108, 624–652.

Brown, J. W., & Braver, T. S. (2005). Learned Predictions of Error Likelihood in the Anterior Cingulate Cortex. *Science*, 307(5712), 1118–1121.

Acknowledgments Supported by the Intelligence Advanced Research Projects Activity (IARPA) through Department of the Interior (DOI) contract D10PC20023. The US Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI or the US Government.

References

1. Kuhn TS (1962) The structure of scientific revolutions. University of Chicago Press, Chicago
2. Poldrack RA (2008) The role of fMRI in cognitive neuroscience: where do we stand? Curr Opin Neurobiol 18(2):223–7
3. Poldrack RA (2012) The future of fMRI in cognitive neuroscience. Neuroimage 62(2):1216–20
4. Gage GJ, Parikh H, Marzullo TC (2008) The cingulate cortex does everything. Ann Improbable Res 14(3):12–15
5. Falkenstein M et al (1991) Effects of crossmodal divided attention on late ERP components: II. Error processing in choice reaction tasks. Electroencephalogr Clin Neurophysiol 78:447–455
6. Gehring WJ et al (1990) The error-related negativity: an event-related potential accompanying errors. Psychophysiology 27:34

7. Botvinick M et al (1999) Conflict monitoring versus selection-for-action in anterior cingulate cortex. Nature 402(6758):179–181
8. Botvinick MM et al (2001) Conflict monitoring and cognitive control. Psychol Rev 108:624–652
9. Shidara M, Richmond BJ (2002) Anterior cingulate: single neuronal signals related to degree of reward expectancy. Science 296(5573):1709–11
10. Shima K, Tanji J (1998) Role of cingulate motor area cells in voluntary movement selection based on reward. Science 282:1335–1338
11. Chandrasekhar PVS et al (2008) Neurobiological regret and rejoice functions for aversive outcomes. Neuroimage 39(3):1472–84
12. Rainville P (1997) Pain affect encoded in human anterior cingulate but not somatosensory cortex. Science 277(5328):968–971
13. Eisenberger NI, Lieberman MD, Williams KD (2003) Does rejection hurt? An FMRI study of social exclusion. Science 302(5643):290–2
14. Posner MI, Petersen SE, Fox PT, Raichle ME (1988) Localization of cognitive operations in the human brain. Science 240(4859):1627–1631
15. Rudebeck PH et al (2008) Frontal cortex subregions play distinct roles in choices between actions and stimuli. J Neurosci 28(51):13775–85
16. Walton ME, Devlin JT, Rushworth MFS (2004) Interactions between decision making and performance monitoring within prefrontal cortex. Nat Neurosci 7(11):1259–1266
17. Stroop JR (1935) Studies of interference in serial verbal reactions. J Exp Psychol 18(6):643–662
18. Brown JW, Braver TS (2005) Learned predictions of error likelihood in the anterior cingulate cortex. Science 307(5712):1118–1121
19. Holroyd CB, Coles MG (2002) The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. Psych Rev 109(4):679–709
20. Brown JW (2009) Multiple cognitive control effects of error likelihood and conflict. Psychol Res 73(6):744–50
21. Nieuwenhuis S et al (2007) Error-likelihood prediction in the medial frontal cortex: a critical evaluation. Cereb Cortex 17:1570–1581
22. Brown JW, Braver TS (2007) Risk prediction and aversion by anterior cingulate cortex. Cogn Affect Behav Neurosci 7(4):266–77
23. Brown JW, Braver TS (2008) A computational model of risk, conflict, and individual difference effects in the anterior cingulate cortex. Brain Res 1202:99–108
24. Holroyd CB, Krigolson OE (2007) Reward prediction error signals associated with a modified time estimation task. Psychophysiology 44(6):913–7
25. Ferdinand NK et al (2012) The processing of unexpected positive response outcomes in the mediofrontal cortex. J Neurosci 32(35):12087–92
26. Jessup RK, Busemeyer JR, Brown JW (2010) Error effects in anterior cingulate cortex reverse when error likelihood is high. J Neurosci 30(9):3467–3472
27. Oliveira FT, McDonald JJ, Goodman D (2007) Performance monitoring in the anterior cingulate is not all error related: expectancy deviation and the representation of action-outcome associations. J Cogn Neurosci 19(12):1994–2004
28. Grinband J et al (2011) The dorsal medial frontal cortex is sensitive to time on task, not response conflict or error likelihood. Neuroimage 57(2):303–311
29. Yeung N, Nieuwenhuis S (2009) Dissociating response conflict and error likelihood in anterior cingulate cortex. J Neurosci 29(46):14506–14510
30. Forster SE, Brown JW (2011) Medial prefrontal cortex predicts and evaluates the timing of action outcomes. Neuroimage 55(1):253–65
31. Ito S et al (2003) Performance monitoring by anterior cingulate cortex during saccade countermanding. Science 302:120–122
32. Kennerley SW et al (2006) Optimal decision making and the anterior cingulate cortex. Nat Neurosci 9(7):940–947
33. Hayden BY, Pearson JM, Platt ML (2011) Neuronal basis of sequential foraging decisions in a patchy environment. Nat Neurosci 14(7):933–9

34. Amador N, Schlag-Rey M, Schlag J (2000) Reward-predicting and reward-detecting neuronal activity in the primate supplementary eye field. J Neurophysiol 84(4):2166–70
35. Matsumoto M et al (2007) Medial prefrontal cell activity signaling prediction errors of action values. Nat Neurosci 10(5):647–656
36. Cole MW et al (2009) Cingulate cortex: diverging data from humans and monkeys. Trends Neurosci 32(11):566–74
37. Ford KA et al (2009) BOLD fMRI activation for anti-saccades in nonhuman primates. Neuroimage 45(2):470–6
38. Chamberlin TC (1965) The method of multiple working hypotheses. Science 148(3671):754–759
39. Platt JR (1964) Strong inference: certain systematic methods of scientific thinking may produce much more rapid progress than others. Science 146(3642):347–353
40. Amiez C, Joseph J-P, Procyk E (2005) Anterior cingulate error-related activity is modulated by predicted reward. European J Neurosci 21(12):3447–52
41. Scheffers MK, Coles MG (2000) Performance monitoring in a confusing world: error-related brain activity, judgments of response accuracy, and types of errors. J Exp Psychol Hum Percept Perform 26(1):141–51
42. Steinhauser M, Maier M, Hübner R (2008) Modeling behavioral measures of error detection in choice tasks: response monitoring versus conflict monitoring. J Exp Psychol Hum Percept Perform 34(1):158–76
43. Yeung N, Cohen JD, Botvinick MM (2004) The neural basis of error detection: conflict monitoring and the error-related negativity. Psychol Rev 111(4):931–59
44. Holroyd CB et al (2005) A mechanism for error detection in speeded response time tasks. J Exp Psychol Gen 134(2):163–191
45. Alexander WH, Brown JW (2010) Computational models of performance monitoring and cognitive control. Top Cogn Sci 2:658–677
46. Alexander WH, Brown JW (2011) Medial prefrontal cortex as an action-outcome predictor. Nat Neurosci 14(10):1338–1344
47. Goldstein RZ et al (2009) Anterior cingulate cortex hypoactivations to an emotionally salient task in cocaine addiction. Proc Natl Acad Sci U S A 106(23):9453–8
48. Jahn A, Nee DE, Brown JW (2011) The neural basis of predicting the outcomes of imagined actions. Front Neurosci 5:128–128
49. Hillman KL, Bilkey DK (2012) Neural encoding of competitive effort in the anterior cingulate cortex. Nat Neurosci 15(9):1290–7
50. Suzuki S et al (2012) Learning to simulate others' decisions. Neuron 74(6):1125–37
51. Wessel JR et al (2012) Surprise and error: common neuronal architecture for the processing of errors and novelty. J Neurosci 32(22):7528–37
52. Glascher J et al (2010) States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron 66(4):585–595

Chapter 17
Using the ACT-R Cognitive Architecture in Combination With fMRI Data

Jelmer P. Borst and John R. Anderson

Abstract In this chapter we discuss how the ACT-R cognitive architecture can be used in combination with fMRI data. ACT-R is a cognitive architecture that can provide a description of the processes from perception through to action for a wide range of cognitive tasks. It has a computational implementation that can be used to create models of specific tasks, which yield exact predictions in the form of response times and accuracy measures. In the last decade, researchers have extended the predictive capabilities of ACT-R to fMRI data. Since ACT-R provides a model of all the components in task performance it can address brain-wide activation patterns. fMRI data can now be used to inform and constrain the architecture, and, on the other hand, the architecture can be used to interpret fMRI data in a principled manner. In the following sections we first introduce cognitive architectures, and ACT-R in particular. Then, on the basis of an example dataset, we explain how ACT-R can be used to create fMRI predictions. In the third and fourth section of this chapter we discuss two ways in which these predictions can be used: region-of-interest and model-based fMRI analysis, and how the results can be used to inform the architecture and to interpret fMRI data.

17.1 Introduction

In 1973, Newell wrote a commentary in which he caricatured the current psychological practice as "playing a game of 20 questions [with nature]" [1]. While Newell considered the individual experiments and theories presented at the symposium to be "exceptionally fine" (p. 291), he was worried that the results would never be integrated into an overarching theory of the mind. As a solution, Newell proposed the idea of *cognitive architectures* (the actual term is not in his 1973 paper but was well in use at CMU when Anderson arrived in 1978; see for instance [2]).

J. P. Borst (✉) · J. R. Anderson
Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15123, USA
e-mail: j.p.borst@rug.nl

B. U. Forstmann, E.-J. Wagenmakers (eds.), *An Introduction to Model-Based Cognitive Neuroscience,* DOI 10.1007/978-1-4939-2236-9_17

A cognitive architecture is first and foremost a psychological theory: it explains for instance how our memory system works. Instead of being limited to a single psychological construct, however, architectures typically account for complete tasks, from perception to response execution. In addition—and unlike most classical psychological theories—a cognitive architecture is implemented as a computer simulation, which can be used to create cognitive models of specific tasks (e.g., the Stroop task, associative recognition, driving a car). This approach has multiple advantages. First, the models yield precise predictions, for instance reaction times and accuracy measures. Particularly when complete tasks are modeled—often models even interact with the same interface as human subjects—a direct comparison with human data is possible. Second, the underlying psychological components (e.g., memory, vision) are shared by the different tasks, and have to be truly general. If a simulated memory system only works for a single task it probably contains too many task-specific constructs. A cognitive architecture forces one to keep the components general enough to work for many different tasks. Third, because complete tasks are modeled, interactions between perception and central cognition (and between cognitive components themselves) arise naturally from the architecture, which can have a large impact on experimental results [3, 4].

For decades, models developed in cognitive architectures were validated using response times, accuracy measures, and sometimes eye movements [e.g., [5]]. However, behavioral data does not always provide enough constraints to distinguish between different models [6; Chap. 13]. For example, the time leading up to a response typically consists of multiple cognitive steps, which can be arranged in different ways. Researchers turned to neuroimaging data for additional constraints and guidance in developing architectures [e.g., [6, 7]]. Cognitive architectures are well-matched to fMRI data: One cannot ignore any of the perceptual, cognitive, or motor components of a task when designing or interpreting fMRI experiments (because they all show up in brain activity) and a cognitive architecture requires that the modeler address all of these components (to get a running model).

In this chapter we describe how the cognitive architecture ACT-R can be used in combination with fMRI data. We will first explain ACT-R in some detail. Then, based on an example task, we will demonstrate the different steps of generating fMRI predictions from an ACT-R model. Subsequently, we discuss two different ways of using these predictions: a region-of-interest analysis and a model-based fMRI analysis. We conclude with a short section on how the two methods complement each other.

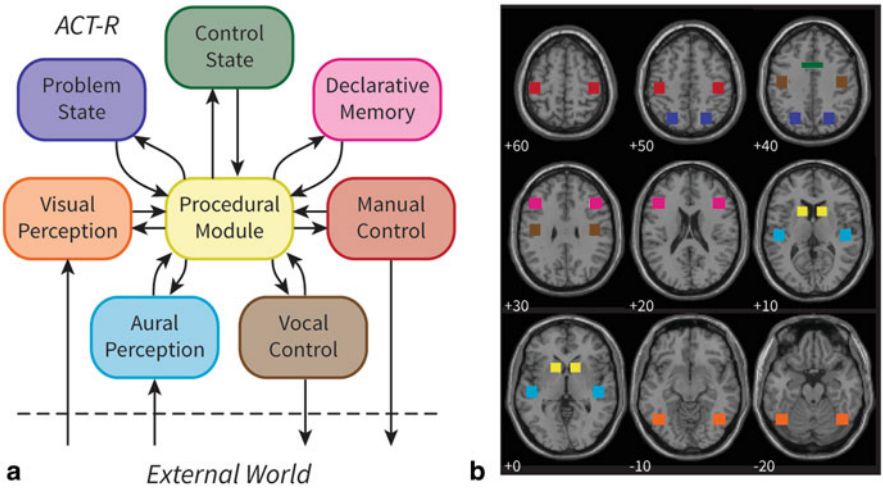


Fig. 17.1 The main modules of ACT-R **a** and associated brain regions **b**. Numbers indicate the z-coordinate of each slice (MNI coordinates); the colors of the regions correspond to the colors in **a**

17.2 ACT-R

Currently, several cognitive architectures are in use, for example SOAR [2], ACT-R [6], EPIC [8], and 4CAPS [7]. In this chapter we will focus on ACT-R[1], because it has an explicit mapping between components of the architecture and brain regions. However, most ideas in this chapter are also applicable to other architectures.

ACT-R consists of a set of independent modules that function around a central procedural module (Fig. 17.1a). There are modules for perception (visual and aural) and action (manual and vocal), and several central cognitive modules (for details on the individual modules, see [6], or [22]). The modules interact with the procedural module through buffers of limited size. The procedural module consists of rules that specify what cognitive action to take given the contents of the buffers. For instance, a rule might request the retrieval of the meaning of word encoded in the visual buffer. An ACT-R *model* consists of such rules and of knowledge in declarative memory (e.g., the meaning of the word 'chair'). Thus, ACT-R itself can be seen as the fixed hardware—the architecture—of the mind, while the models function as software that runs on this hardware. The modules of ACT-R have been mapped onto small regions in the brain, which are shown in Fig. 17.1b. These regions are assumed to be active when the corresponding module is active (see the section on region-of-interest analysis).

[1] For the range of tasks (and associated publications) that have been modeled with ACT-R, see http://act-r.psy.cmu.edu/. ACT-R can also be downloaded from this website.

17.3 Using ACT-R to Predict fMRI Data

In this section we will describe how ACT-R can be used to predict fMRI data. First, we describe the task that we will use as an example throughout this chapter. We will then introduce the model, followed by how it can be used to generate fMRI predictions. The Lisp code for the model and Matlab code to generate the predictions can be downloaded from http://act-r.psy.cmu.edu/, under the title of this chapter.

17.4 The Example Task: Associative Fan

To illustrate the analysis we will use a previously published experiment with an associated ACT-R model (Experiment 2, [9]). This experiment was designed to test the assumption that declarative memory activity is reflected by a region in the prefrontal cortex (see Fig. 17.1b, the pink regions), while representational activity of the problem state module (roughly comparable to a capacity-limited working memory store, e.g., [10]) is reflected by a region in the posterior parietal cortex (Fig. 17.1b, dark blue regions). To this end, memory and representational requirements were independently manipulated in an associative recognition task.

Figure 17.2a shows the basic procedure. A trial started with a 2 s fixation screen, followed by a 6 s study presentation of a paired-associate. Subjects were asked to memorize the paired-associate that was presented, in this case 'band—2'. The study probe was followed by a 6-second fixation screen, after which a test probe was shown for a maximum of 6 s or until the response was given. The test probe consisted of a word (i.e., 'band'); subjects had to respond with the associated number (i.e., '2').

Memory requirements were manipulated within-subject by varying the delay between study and test items. The trial in Fig. 17.2a is an example of having a study and test item in the same trial, but they could be as far as 7 trials apart. There were three levels: no delay, short delay (1–2 trials), and long delay (6–7 trials). Representational requirements were manipulated between-subject by contrasting a 'paired' with a 'generated' condition. Figure 17.2a shows the paired condition; Fig. 17.2b the generated condition. Instead of showing the paired-associate directly, in the generated condition a word phrase was given: 'b-nd—id = adhesive strip'. Subjects were asked to solve the phrase by finding a single letter to complete it. At test, 'band' was shown, and subjects had to respond at the recall test with the position of the letter they had filled in (i.e., '2'). Thus, responses were identical in the paired and the generated conditions. The assumption was that subjects would show greater representational activity in the generated case because they had to solve the phrase at study and extract the response at test.

Figure 17.3 shows the behavioral results, accuracy on the left and response times on the right [for details see [9]]. The effect of the delay manipulation is clear: subjects made more errors and were slower to respond when the delay between study and test was longer (response times of correct responses are shown). The effect of the representational manipulation on behavior was more modest: no effects on accuracy,

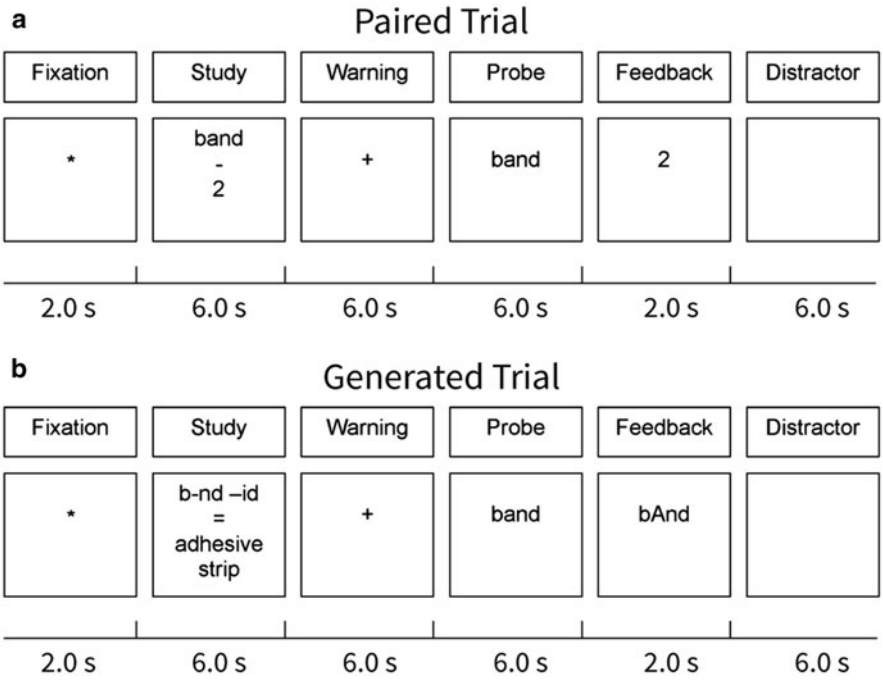


Fig. 17.2 Experimental procedure. (Adapted from Fig. 1 in Anderson et al. [9] by permission of the publisher. Copyright 2008 of the original by Oxford University Press)

and only a marginal effect on response times ($F(1,17) = 2.99$, $p = .10$), with the generated condition leading to slightly slower responses. This illustrates why behavioral data are often not detailed enough to constrain computational models: in these data there is almost no difference between the paired and the generated condition. However, the fMRI results will show that there are clear differences between these conditions that are captured in the ACT-R model of the task. Before we turn to the fMRI results we will discuss that ACT-R model, and how such a model can be used to generate fMRI predictions.

17.5 The Model

As explained above, an ACT-R model consists of procedural rules and declarative knowledge that 'runs' on the cognitive architecture. Anderson et al. [9] presented a model that performs the associative recognition task.[2] Figure 17.4 shows a schematic

[2] A version of the model that was adapted for this chapter can be downloaded from http://act-r.psy.cmu.edu, under the title of this chapter.

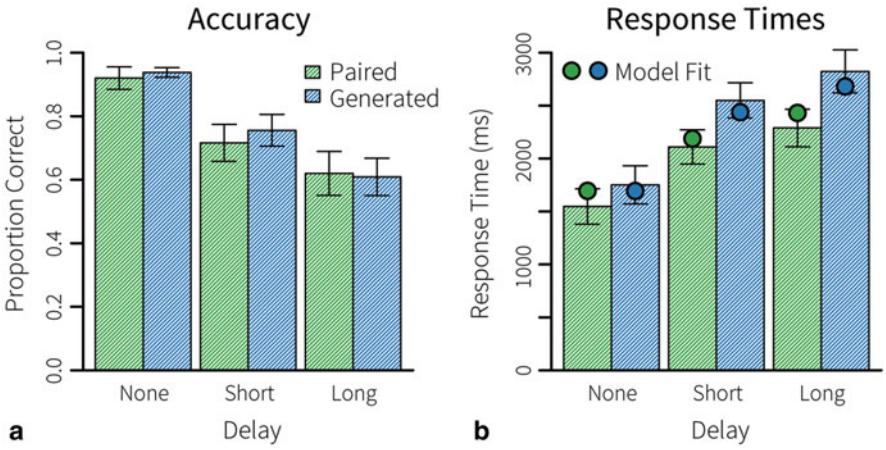


Fig. 17.3 Behavioral results. Error bars indicate standard errors

of model activity for four different trial types. In all conditions, the model starts with encoding the start fixation. When the pair or the phrase is presented two seconds later, it also encodes those, and represents the information in the problem state module (ACT-R's capacity-limited working memory store). In the paired condition, the model then actively stores the pair in declarative memory. In the generated condition, it completes the phrased based on information retrieved from declarative memory, and stores the completed phrase in memory [see 9 for details]. As the figure indicates, the model assumes no difference in memory activity between the paired and the generated condition in the study phase. However, in the generated condition an extra problem state action is performed to extract the position of the letter that was filled in (i.e., bAnd $=>2$), resulting in more representational activity.

In the immediate test conditions (i.e., study and test are in the same trial, cf. Fig. 17.2), the model retrieves the pair (in the paired condition) or the position (in the generated condition) from memory. It then represents this information in the problem state module and generates a response. Thus, both memory and representational requirements are the same in the immediate test phase of the paired and generated conditions. If there is a delay between study and test, it is harder to retrieve the pair from memory, which results in longer declarative memory activity than in the immediate conditions. In addition, in the generated delay condition it is assumed that the model cannot directly remember the position that it filled in, but that has to retrieve the phrase from memory. As a consequence, an extra representational step has to be performed to extract the position again [see 9 for the rationale behind this]. Thus, in the generated delay condition additional representational activity is predicted in the test phase as compared to the other three conditions.

One of the advantages of using a cognitive architecture is that it provides us with latency information of the modules (e.g., visually encoding a stimulus, representing

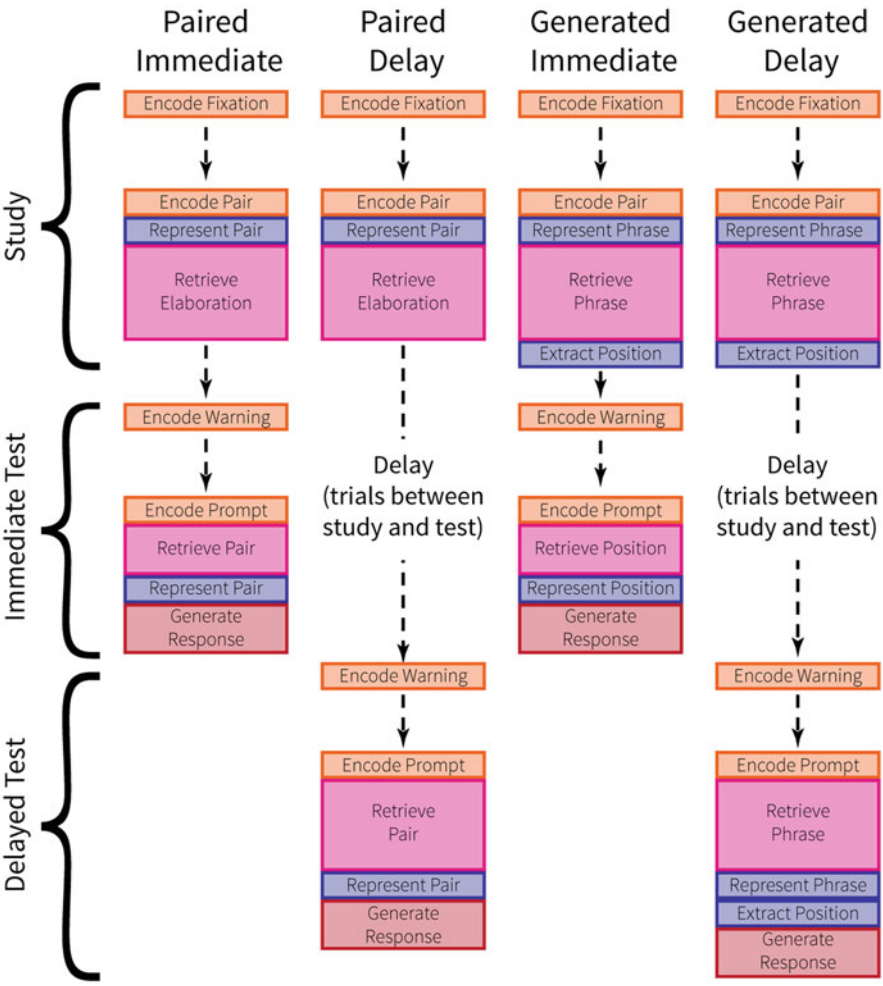


Fig. 17.4 An overview of model activity in four representative conditions. Boxes are not drawn to scale, but indicate the general pattern of module activity. Colors correspond to Fig. 1 **a** *orange* corresponds to visual activity, *blue* to problem state, *pink* to declarative memory, and *red-brown* to motor activity. (Adapted from Fig. 6 in Anderson et al. [9] by permission of the publisher. Copyright 2008 of the original by Oxford University Press)

a pair, generating a response). For the current model all parameters were left at their default values, except for the time it takes to retrieve information from memory. This was estimated to fit the model to the behavioral data. The resulting fit is shown in Fig. 17.3b (accuracy was not modeled). The correspondence between model and data is acceptable: the main effects in the data are reflected by the model (originally, the model was compared to data of two experiments, which yielded a better overall fit).

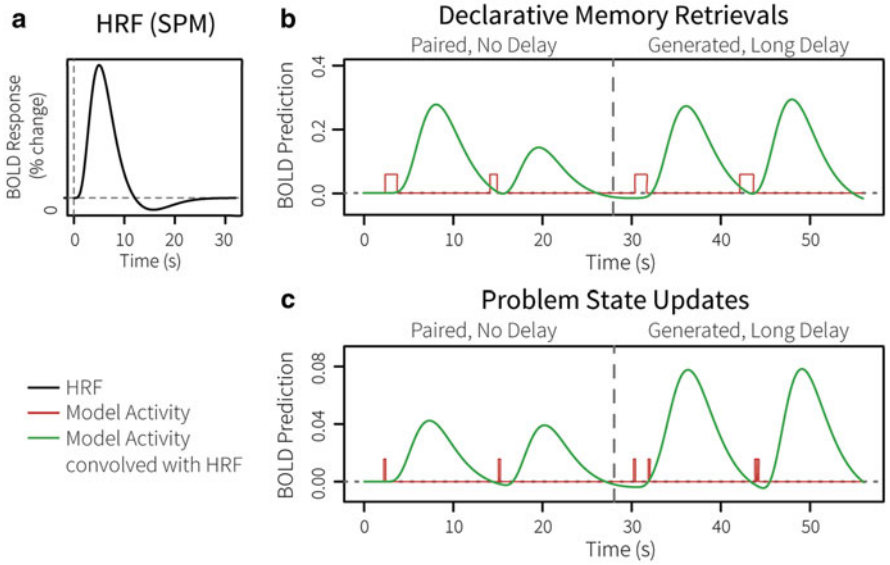


Fig. 17.5 Predicting the BOLD response with ACT-R. Panel **a** shows a typical hemodynamic response function; **b** and **c** show model activity in *red* and the model activity convolved with the HRF in *green*

17.6 Predicting the BOLD Response

The next step in using ACT-R with fMRI data is generating BOLD (blood-oxygen-level-dependent) response predictions. As is known from the fMRI literature, the BOLD response is sluggish with respect to neural activity. This is illustrated in Fig. 17.5a: if there is a spike of neural activity at time 0, the BOLD or hemodynamic response function (HRF) rises slowly to a peak around 5 s, declines again and dips under the baseline until it eventually comes back to baseline at around 30 s. The HRF is often described with a gamma function or a mix of gamma functions [e.g., [11]–[13]]. In this chapter we will use a difference of two gamma functions from the SPM software package [14].

To generate BOLD predictions, we convolve module activity—which resembles neural activity with respect to its direct timing—with the HRF. This is shown in Fig. 17.5b (declarative memory retrievals) and Fig. 17.5c (problem state updates). The red lines indicate the activity of the modules (cf. Figure 17.4). Two different trial types are shown: a paired trial without a delay between study and test on the left, and a generated trial with a long delay on the right. As explained in the model description above, a long delay leads to a slower second memory retrieval than no delay, and the generated condition shows more representational activity than the paired condition. The green lines depict the predicted BOLD response: longer activities lead to longer and higher BOLD predictions (declarative memory), and

multiple short module activations can be added up to a single large BOLD response (problem state).[3] In the next two sections we will describe how these predictions can be used for a region-of-interest analysis and for a model-based fMRI analysis.

17.7 Region-of-Interest Analysis

Most ACT-R/fMRI papers to date have used so-called region-of-interest (ROI) analyses [e.g., [6, 9, 11, 15, 16]]. For this analysis stream, all ACT-R modules have been mapped onto small regions of the brain (Fig. 17.1b; see [6] for Talairach coordinates and details of the regions, or [15] for MNI coordinates). The assumption is that these regions are active when the corresponding module predicts activity. For example, declarative memory retrievals should lead to activity in the prefrontal cortex and problem state updates to activity in the posterior parietal cortex. Note that we do not assume that these are the only regions that are active in response to the modules, and neither that these regions exclusively indicate activity of ACT-R modules.

Using predefined ROIs has a number of advantages. By comparing model predictions and data one can validate and constrain models (i.e., if the predictions are off, the model should be improved), which would not be possible without a predefined mapping. In addition, because only a limited number of predefined regions are inspected the typical multiple-comparison problem of fMRI is avoided. This makes is possible to analyze much smaller differences than is possible with conventional fMRI analyses. The obvious disadvantage of an ROI analysis is that it is constrained to the predefined ROIs and ignores the rest of the brain.

Figure 17.6 (declarative memory) and 17.7 (problem state updates) show the results for our example dataset. The top panels show the model predictions; the bottom panels show the data. For declarative memory, the model predicted no difference between the paired and the generated conditions, but a clear difference between the no, short, and long delay conditions at test. In the data, we see hardly any differences between these conditions at study—as predicted—and some differences—in the right order—at test, especially in the generated condition (if we average over the paired and the generated conditions the effect is clear, see also [9]). In addition, the peak seems to be larger in the generated condition, but this difference was not significant (see [6], for details). With respect to the problem state updates (Fig. 17.7), the model predicted a larger response in the generated condition than in the paired condition, and only an effect of delay in the generated condition at test. These predictions were matched by the data.

We can conclude that the prefrontal region indeed mostly reflects declarative memory retrievals, while the posterior parietal cortex reflects updating problem representations. The model's predictions matched the data reasonably well, although

[3] The amplitude of the predicted BOLD response for declarative memory is much larger than for the problem state module. However, the amplitude is typically fitted separately for each module in a region-of-interest analysis and is not important for the model-based fMRI analysis.

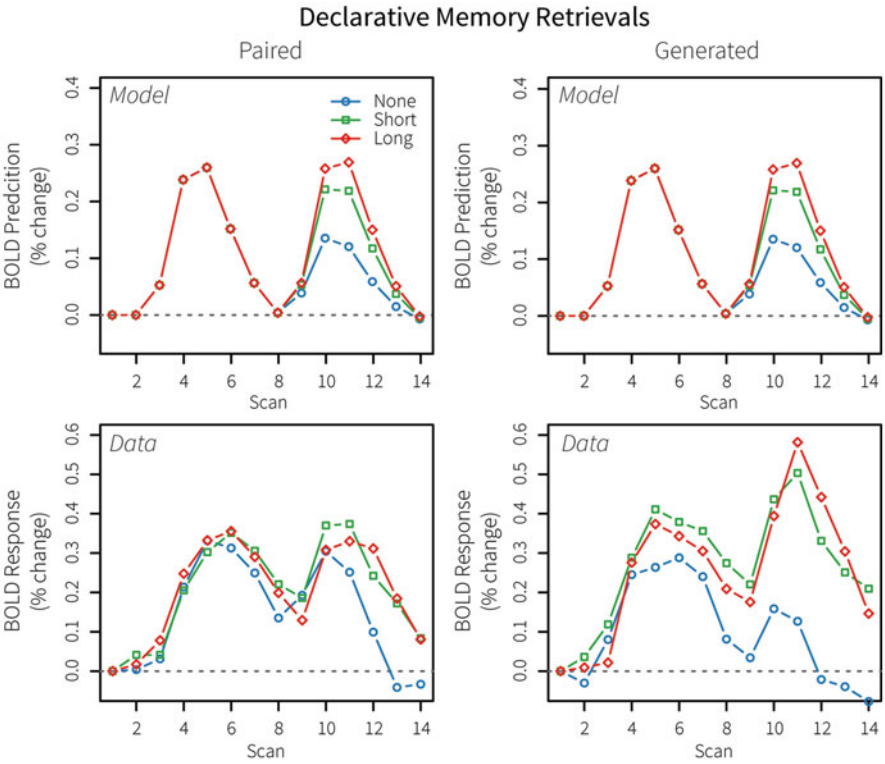


Fig. 17.6 ROI results for declarative memory retrievals. Top panels show model predictions; bottom panels data. One scan = 2 s

there were some discrepancies (e.g., for problem state updates the data shows a larger peak at test than at study in the generated condition, the model did not predict this). Such discrepancies can be due to several different reasons: the model might be inaccurate, the mapping of ACT-R on the brain might be incomplete, or—with respect to the noisy results for declarative memory at test—we might have to test more subjects. In addition, it is known that the shape of the BOLD response is different in different brain regions, as well as between different subjects. Here we presented a priori BOLD predictions based on SPM's HRF, but it would be reasonable to fit the shape and magnitude of the BOLD response separately for each region.

17.8 Model-Based fMRI Analysis

The ROI analysis has one clear disadvantage: it is dependent on the correctness of the predefined mapping. The current mapping was based on a reading of the literature on regional functions, and might therefore not be optimal. To find regions that map best

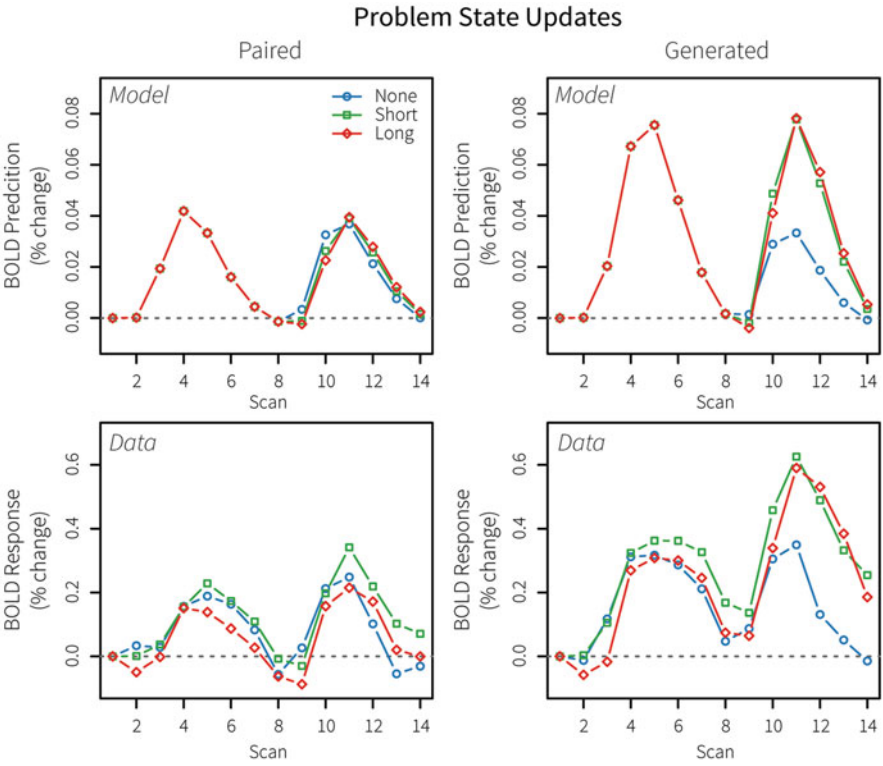


Fig. 17.7 ROI results for problem state updates. Top panels show model predictions; bottom panels data. 1 scan = 2 s.

on our module predictions we can use model-based fMRI [e.g., [17, 18]]. Whereas in conventional fMRI the experimental structure is typically used as the basis for the regressors in the general linear model (GLM, see Chap. 4), in model-based fMRI predictions stemming from a computational model are used. In the case of ACT-R this means that predictions such as the ones in Fig. 17.5 are regressed against the BOLD response in all voxels in the brain. This shows which voxels correlate significantly with the predictions of a module, indicating that these voxels might implement the functionality of the module.

Using model-based fMRI with ACT-R involves generating model predictions for all trials for all subjects, because model activity is regressed against the BOLD response over the whole experiment. This has not typically been done in ROI analysis although it could be. Generating predictions for single trials requires representing any differences that might occur because of the specific stimuli on each trial, as these can lead to different behavior and different model predictions [see 19, for an example]. Second, because we are regressing the model predictions directly against the brain data, it is important to have an exact time-mapping between data and model, to avoid,

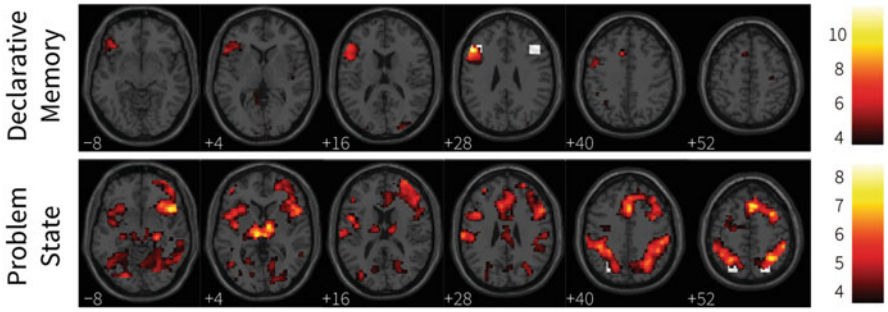


Fig. 17.8 Model-based fMRI results. Statistical maps were thresholded at $p < .001$ (uncorrected). *White* squares indicate predefined ACT-R regions

for instance, comparing a fixation in the data with a key-press in the model [17]. To this end, trial onset and key-presses were lined up between model and data. That is, the model predictions on each trial were subjected to a linear transformation to create a perfect response-time match to the data (i.e. all model activity was increased or decreased in length on each trial; see [19] for details). The resulting predictions were used for the model-based analysis.

Figure 17.8 shows the results of a model-based analysis for our example task (originally reported in [20]). The top panel shows that declarative memory updates were exclusively reflected by activity in a left prefrontal region, located directly on top of the predefined ACT-R region (indicated by white squares). The bottom panel shows that problem state activity was reflected by a large number of regions. This is not very surprising, given that many regions will be active in response to the task, and we search for correlating regions. A number of regions show a strong response: the largest and most significant regions included the inferior parietal lobule and the anterior cingulate. In addition, we see strong correlations in the thalamus (slice + 4) and inferior frontal gyrus (-8). This illustrates a weakness of model-based fMRI: multiple regions might correlate with the predictions, yielding imprecise results. A meta-analysis combining these results with four additional studies indicated that the parietal and anterior cingulate activity was consistent over the five studies, while the other regions in Fig. 17.8 are probably due to idiosyncrasies of the current task and model [20].

17.9 Concluding Remarks

In this chapter we discussed how ACT-R can be used in combination with fMRI data. We described two different analyses: ROI analysis and model-based fMRI analysis. The remaining question is which analysis to use. This naturally depends on the situation: If one wants to test the predictions of a cognitive model it is more constraining to have pre-specified regions as in the ROI analysis. If one wants to understand how a cognitive function maps onto the brain, the model-based approach

allows one to see the full picture. With respect to model-based fMRI, the analysis is not limited to the current modules—any prediction from a model can be used (e.g., only numerical retrievals or representational activity in response to visual encoding). In effect, model-based fMRI yields a mapping that can be used for ROI analyses. However, the results of model-based fMRI are strongly dependent on the quality of the model and on how well the experiment dissociated different model processes. For this reason it either should be used over multiple different tasks as in [20] or in combination with conventional fMRI analyses.

Exercises

1. Read Newell (1973a). Do you agree or disagree with Newell's diagnosis? Explain why.
2. Do you think cognitive architectures are a solution to Newell's problem?
3. Use the included Matlab code to generate model predictions for the manual and visual modules (like Figs. 17.6 and 17.7). Do the predictions match your expectations?
4. Give three possible reasons for discrepancies between model predictions (such as in Figs. 17.6 and 17.7) and fMRI data.
5. What do you think is more interesting: fitting the BOLD response of the model to the data by changing the parameters of the HRF, or only using a priori predictions? Explain why.
6. Discuss advantages of ROI analysis as compared to conventional exploratory fMRI analysis.
7. Discuss advantages of model-based fMRI analysis as compared to conventional exploratory fMRI analysis.
8. Discuss advantages of conventional fMRI analysis as compared to the methods described in this chapter.

Further Reading

- 'You can't play 20 questions with nature and win' [1] is Allen Newell's commentary in which he argues for a cognitive architecture approach. Four decades later the paper is still a thought-provoking and entertaining must-read for every cognitive scientist. In a companion piece in the same volume he presents his initial production system approach to cognitive architectures [21].
- Chapter 1 of [6] gives a very clear introduction to cognitive architectures and ACT-R. In case you do not have the book available, [22] provides an introduction to ACT-R and its mapping on brain regions. For a more concise introduction to ACT-R's mapping on brain regions, see [11].
- Gläscher and O'Doherty [17] give a general introduction to model-based fMRI analysis. Borst and Anderson [20] show how model-based fMRI can be applied in combination with an ACT-R model.

References

1. Newell A (1973) You can't play 20 questions with nature and win: projective comments on the papers of this symposium. In: Chase WG (ed) Visual information processing. Academic, New York, p 283–308
2. Newell A (1990) Unified theories of cognition. Harvard University, Cambridge
3. Kieras DE, Meyer DE (1997) An overview of the EPIC architecture for cognition and performance with application to human-computer interaction. Hum Comput Interact 12:391–438
4. Van Maanen L, Van Rijn H, Borst JP (2009) Stroop and picture-word interference are two sides of the same coin. Psychon Bull Rev 16(6):987–999
5. Salvucci DD (2006) Modeling driver behavior in a cognitive architecture. Hum Factors 48(2):362–380
6. Anderson JR (2007) How can the human mind occur in the physical universe? Oxford University, New York
7. Just MA, Varma S (2007) The organization of thinking: what functional brain imaging reveals about the neuroarchitecture of complex cognition. Cogn Affect Behav Neurosci 7(3):153–91
8. Meyer DE, Kieras DE (1997) A computational theory of executive cognitive processes and multiple-task performance: Part 1. Basic mechanisms. Psychol Rev 104(1):3–65
9. Anderson JR, Byrne D, Fincham JM, Gunn P (2008) Role of prefrontal and parietal cortices in associative learning. Cereb Cortex 18(4):904–914
10. Borst JP, Taatgen NA, Van Rijn H (2010) The problem state: a cognitive bottleneck in multitasking. J Exp Psychol Learn Mem Cogn 36(2):363–382
11. Anderson JR, Fincham JM, Qin Y, Stocco A (2008) A central circuit of the mind. Trends Cogn Sci 12(4):136–143
12. Cohen MS (1997) Parametric analysis of fMRI data using linear systems methods. Neuroimage 6(2):93–103
13. Friston KJ, Fletcher PC, Josephs O, Holmes A, Rugg MD, Turner R (1998) Event-related fMRI: characterizing differential responses. Neuroimage 7(1):30–40
14. Friston KJ, Ashburner JT, Kiebel SJ, Nichols TE, Penny WD (2007) Statistical parametric mapping. The analysis of functional brain images. Academic Press, London, UK
15. Borst JP, Taatgen NA, Stocco A, Van Rijn H (2010) The neural correlates of problem states: testing fMRI predictions of a computational model of multitasking. PLoS ONE 5(9):e12966
16. Sohn MH, Goode A, Stenger VA, Jung KJ, Carter CS, Anderson JR (2005) An information-processing model of three cortical regions: evidence in episodic memory retrieval. Neuroimage 25(1):21–33
17. Gläscher JP, O'Doherty JP (2010) Model-based approaches to neuroimaging: combining reinforcement learning theory with fMRI data. (Wiley interdisciplinary reviews). Cognitive Science 1(4):501–510
18. O'Doherty JP, Hampton A, Kim H (2007) Model-based fMRI and its application to reward learning and decision making. Ann N Y Acad Sci 1104:35–53
19. Borst JP, Taatgen NA, Van Rijn H (2011) Using a symbolic process model as input for model-based fMRI analysis: locating the neural correlates of problem state replacements. Neuroimage 58(1):137–147
20. Borst JP, Anderson JR (2013) Using model-based functional MRI to locate working memory updates and declarative memory retrievals in the fronto-parietal network. Proc Natl Acad Sci U S A 110(5):1628–33
21. Newell A (1973) Productions systems: models of control structures. In: Chase WG (ed) Visual information processing. Academic, New York, pp 527–546.
22. Anderson JR (2005) Human symbol manipulation within an integrated cognitive architecture. Cognitive Science 29:313–341

Index

B. U. Forstmann, E.-J. Wagenmakers (eds.), *An Introduction to Model-Based Cognitive Neuroscience,* DOI 10.1007/978-1-4939-2236-9