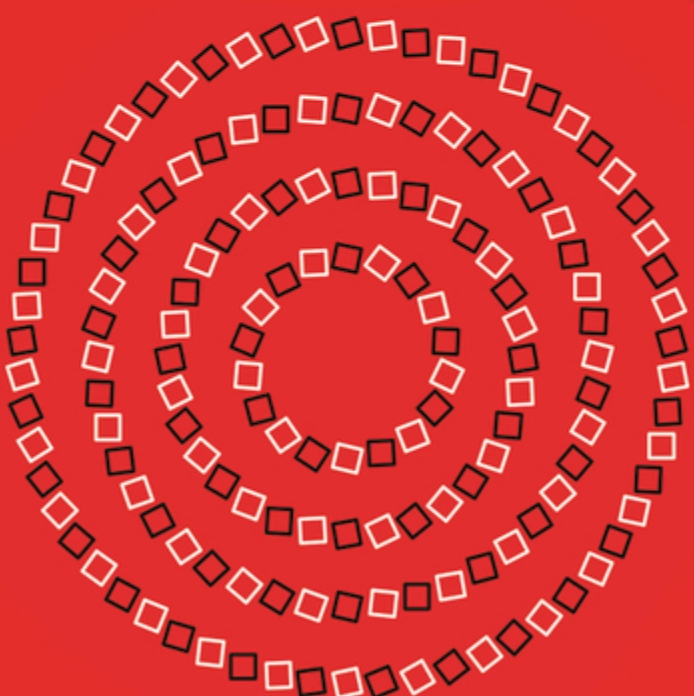
OXFORD

The Predictive Mind



JAKOB HOHWY

THE PREDICTIVE MIND

The Predictive Mind

JAKOB HOHWY

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

First Edition published in 2013

Impression: 1

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data
Data available

Library of Congress Control Number: 2013953488

ISBN 978–0–19–968273–7 (hbk.)
ISBN 978–0–19–968673–5 (pbk.)

As printed and bound by
CPI Group (UK) Ltd, Croydon, CR0 4YY

Contents

Part II The World

Part III The Mind

Preface

My work on this book was made possible through invaluable research support from the Australian Research Council and from Monash University.

I am grateful to many researchers from around the world for inspiration, fruitful discussion, and generous comments.

My colleagues at Monash have influenced me, worked with me, and trained me in science; thanks in particular to my co-authors Bryan Paton, Colin Palmer, and Peter Enticott, to Steve Miller and Trung Ngo for including me in many projects and discussions, and to Naotsugu Tsuchiya, Anastasia Gorbunova, Mark Symmons, George van Doorn, Andrew Paplinski, Lennart Gustavsson, and Tamas Jantvik. Thanks also to my colleagues in philosophy, many of whom have repeatedly been roped in to do pilot studies, and to our participants and the patients who have endured many hours of rubber-hand illusion tapping in the lab.

Andreas Roepstorff's group in Aarhus are expert navigators in blue ocean research, and in many ways initiated and enabled my interest in this field. In addition to Andreas, Josh Skewes deserves thanks for many hours of discussion. Chris and Uta Frith, also sometimes at Aarhus, remain great, generous influences; they are the paradigm of the open-minded academic, especially when data are in the offing.

I am fortunate to have friends in philosophy and neuroscience who are prepared to endure lengthy discussions on predictive coding and the brain. Tim Bayne very early on encouraged me to push on with the book, and he read and commented extensively on the manuscript at various stages; I am extremely grateful for his academic generosity. Thomas Metzinger likewise went well beyond the call of duty and offered generous comments on a draft of the book; I also greatly benefited from many discussions with Thomas' group of colleagues and students while visiting Mainz. During a week at University of Tokyo's Centre for Philosophy I enjoyed very valuable discussions on the book with Yukihiro Nobuhara and his colleagues and students. I have benefited greatly from many stimulating and encouraging discussions and comments on my writings and ideas from Andy Clark. Ned Block offered fruitful and needed resistance to parts of the story. Tim Lane and Yeh Su-Ling and colleagues from Taipei generously discussed many aspects of the book with me. I have had fruitful discussions with Floris de Lange, Sid Kouider, and Lars Muckli. Anonymous reviewers from the Press offered a host of insightful comments and criticisms.

I am especially grateful to Karl Friston whose work in so many ways has inspired the book. On numerous occasions, Karl has patiently offered feedback on my work. He read and commented extensively on every chapter of this book, he has endured long-haul flights to participate in interdisciplinary workshops, and has in many ways contributed to my work and furthered my understanding of the hypothesis-testing brain. It is very encouraging to experience the open-mindedness with which Karl approaches my attempts to translate the framework into philosophy, even as much of the mathematical rigour and detail is lost in translation. I of course remain responsible for any shortcomings.

The book is dedicated to my family: Linda Barclay, for encouraging me to write it, for predicting my errors, and for being with me; and Asker and Lewey, for being terrific rubber-hand guinea pigs and neurodevelopmental inspirations.

Introduction

A new theory is taking hold in neuroscience. The theory is increasingly being used to interpret and drive experimental and theoretical studies, and it is finding its way into many other domains of research on the mind. It is the theory that the brain is a sophisticated hypothesis-testing mechanism, which is constantly involved in minimizing the error of its predictions of the sensory input it receives from the world. This mechanism is meant to explain perception and action and everything mental in between. It is an attractive theory because powerful theoretical arguments support it. It is also attractive because more and more empirical evidence is beginning to point in its favour. It has enormous unifying power and yet it can explain in detail too.

This book explores this theory. It explains how the theory works and how it applies; it sets out why the theory is attractive; and it shows why and how the central ideas behind the theory profoundly change how we should conceive of perception, action, attention, and other central aspects of the mind.

THE ARGUMENT

I am interested in the mind and its ability to perceive the world. I want to know how we manage to make sense of the manifold of sensory input that hits the senses, what happens when we get it wrong, what shapes our phenomenology, and what this tells us about the nature of the mind. It is these questions I seek to answer by appeal to the idea that the brain minimizes its prediction error.

My overall argument in this book has three strands. The first strand is that this idea explains not just that we perceive but *how* we perceive: the idea applies directly to key aspects of the phenomenology of perception. Moreover, it is *only* this idea that is needed to explain these aspects of perception. The second strand in my argument is that this idea is attractive because it combines a compelling theoretical function with a simple *mechanical* implementation. Moreover, this basic combination is of the utmost *simplicity*, yet has potential

to be applied in very nuanced ways. The third strand of the argument is that we can learn something *new* from applying this idea to the matters of the mind: we learn something new about the mechanics of perception, and about how different aspects of perception belong together, and we learn something new about our place in nature as perceiving and acting creatures.

The overall picture I arrive at from considering the theory is that the mind arises in, and is shaped by, prediction. This translates into a number of interesting, specific aspects of mind:

> Perception is more actively engaged in making sense of the world than is commonly thought. And yet it is characterized by curious passivity. Our perceptual relation to the world is robustly guided by the offerings of the sensory input. And yet the relation is indirect and marked by a somewhat disconcerting fragility. The sensory input to the brain does not shape perception directly: sensory input is better and more perplexingly characterized as feedback to the queries issued by the brain.
>
> Our expectations drive what we perceive and how we integrate the perceived aspects of the world, but the world puts limits on what our expectations can get away with. By testing hypotheses we get the world right, but this depends on optimizing a rich tapestry of statistical processes where small deviances seem able to send us into mental disorder. The mind is as much a courtroom as a hypothesis-tester.
>
> Perception, action, and attention are but three different ways of doing the very same thing. All three ways must be balanced carefully with each other in order to get the world right. The unity of conscious perception, the nature of the self, and our knowledge of our private mental world is at heart grounded in our attempts to optimize predictions about our ongoing sensory input.
>
> More fundamentally still, the content of our perceptual states is ultimately grounded not in what we do or think but in who we are. Our experience of the world and our interactions with it, as well as our experience of ourselves and our actions, is both robustly anchored in the world and precariously hidden behind the veil of sensory input. We are but cogs in a causally structured world, eddies in the flow of information.
>
> The theory promises not only to radically reconceptualize who we are and how aspects of our mental lives fit into the world. It unifies these themes under one idea: we minimize the error between the hypotheses generated on the basis of our model of the world and the sensory deliverances coming from the world. A single type of mechanism, reiterated throughout the brain, manages everything. The mechanism uses an assortment of standard statistical tools to minimize error and in doing so gives rise to perception, action, and attention, and explains puzzling aspects of these phenomena. Though the description of the mechanism is statistical it is just a causal neuronal mechanism and the theory therefore sits well with a reductionist, materialist view of the mind.

A theory with this kind of explanatory promise is extremely exciting. This excitement motivates the book. The message is that the theory delivers on the promise, and that it lets us see the mind in new light.

I am confident that many other aspects of this approach to the brain and the mind can and will be explored. This book by no means exhausts the impact of this kind of approach to life and mind. I focus on key issues in perception but largely leave out higher cognitive phenomena such as thought, imagery, language, social cognition, and decision-making. I also mostly leave aside broader issues about the relation of the theory to sociology, biology, evolutionary theory, ecology, and fundamental physics. This still leaves plenty of work to do in this book.

PLAN

The book has three parts. Part I relies on the work of researchers in neuroscience and computational theory, in particular that of Karl Friston and his large group of collaborators. In a series of chapters the prediction error minimization mechanism is motivated, described, and explained. We start with a very simple Bayesian conception of perception and end with a core mechanism that makes Bayesian inference sensitive to statistical estimation of states of the world as well as their precisions, while making room for context-sensitivity and model complexity. The overall view is attractive in part because it appeals to just this one mechanism—this is thus a very ambitious unificatory project.

This area of research is mathematically heavy, and this is indeed part of the reason for its increasing influence: the mathematical equations provide formal rigour and the possibility of quantifiable predictions. However, my exposition is done with a minimum of technical, formal detail. I appeal to and explain very general Bayesian and statistical ideas. This neglects mathematical beauty but will make the discussion accessible and more easy to apply to conceptual and empirical puzzles in cognitive science and philosophy.

My main concern is to bring out the key elements of the prediction error minimization mechanism, in particular, the way prediction error arises and is minimized, how expectations for the precision on prediction error are processed, how complexity and context-dependence factor in, and how action is an integral part of the mechanism. Furthermore, I set out how this mechanism is re-iterated hierarchically throughout the brain. These are the elements needed to explain everything else and they can be fairly conveyed without too much formal detail. I do provide a brief primer of Bayes' rule at the end of Chapter 1, and describe some rudimentary formal detail in the notes to Chapter 2. I also at times provide some very minimal formal expressions, which mainly serve as reminders for how the more complex points relate to simpler Bayesian expressions; these more formal elements are not essential to

the flow of the overall argument but they do serve as an indication of the vast mathematical backdrop to this theory.

The prediction error minimization framework can also be generalized to a basic notion of *free energy minimization*. I do obliquely appeal to this notion when I go beyond the simple, epistemically focused version of the problem of perception in terms of prediction error minimization, but I do not in general use this broader notion of free energy in my discussion, nor have I delved much into the wider consequences of it. This is because the aspects of mind I concentrate on make best sense by first and foremost appealing to the more directly epistemic notion of Bayesian hypothesis testing. Fundamentally however there is no difference between these formal frameworks.

Part II looks at the consequences of the basic prediction error minimization mechanism for some long-standing debates in cognitive science having to do with our perception of states of affairs in the world: the binding problem, and the debate about how much our prior beliefs shape perception. The hypothesis-testing brain theory is able to chart interesting routes through these debates. This part of the book then sets out a multifaceted view of reality testing and fine-tuning of prediction error minimization, which is in turn related to mental disorder.

I will be dealing with these issues at a level of detail that is fine-grained enough to establish the framework as fruitful for understanding them, though I do not give full accounts of every aspect of the vast literature on the binding problem, cognitive impenetrability, mental illness, and so on. I illustrate many of these discussions with examples of empirical research from psychology and cognitive neuroscience, including some that I have been directly involved in myself.

This Part of the book demonstrates that even though there is just one, basic mechanism in this account of the mind, the explanatory reach is both very impressive and illuminating. The final chapter in Part II continues this project in a set of more squarely philosophical debates about misrepresentation, rule-following, representation, and understanding.

In Part III, I explore what the prediction error minimization mechanism can tell us about some intriguing aspects of our mental lives that have fuelled deep and recalcitrant debates in philosophy and cognitive science. Again, given the extreme explanatory ambition of this theory—it is supposed to give the fundamental principle for the brain—we should expect it to apply to all aspects of the mind.

First I apply it to attention and its ill-understood relation to conscious perception. Then I appeal to the theory in an account of the unity of conscious perception, which is an intriguing and puzzling aspect of our perceptual lives. In the penultimate chapter I explore how the theoretical framework can give a sense of our overall place, as perceiving and acting creatures, set over

against the world. Finally I loosen the reins more and speculate about how the framework might be extended to emotion, introspection, the privacy of consciousness, and self.

The simple notion of prediction error minimization at the heart of the theory is capable of both addressing these kinds of deep issues with interesting results, and, importantly, seems able to unify these very diverse aspects of our mental lives under one principle.

Overall, this edges us closer to a unified, naturalistic account which affords a new and surprising understanding of many puzzling aspects of the mind. Conceiving of the brain as a hypothesis tester enables us to re-asses, re-calibrate, and reconceive a whole host of problems and intuitive ideas about how the mind works and how we know the world.

BACKGROUND

Although the formal machinery surrounding the prediction error minimization account has been developed only recently, the core ideas are not new. It was anticipated a millennium ago by Ibn al Haytham (Alhazen) (ca. 1030; 1989), who developed the view that "many visible properties are perceived by judgement and inference" (II.3.16). There is certainly also a distinct Kantian element to the idea that perception arises as the brain uses its prior conceptions of the world (the forms of intuition of space and time, and the categories etc.) to organize the chaotic sensory manifold confronting the sensory system (Kant 1781). The relation between our thinking (or inference) and the manifold content delivered from the senses is captured in the Kantian slogan that thoughts without content are empty, intuitions without concepts are blind: "The understanding can intuit nothing, the senses can think nothing. Only through their unison can knowledge arise" (A51/B75).

But it was Hermann von Helmholtz who first seized on the idea of the brain as a hypothesis tester, in a direct reaction to Kant. He was worried about how it is, on a Kantian way of thinking, "that we escape from the world of the sensations of our own nervous system into the world of real things" (Helmholtz 1855; 1903; for the relation to Kant, see Lenoir 2006). His answer was, basically, that we are guided by the answers nature delivers when we query it, using unconscious perceptual inference based on our prior learning (Helmholtz 1867). It is this kind of inference that anchors perception in the world.

This brilliant and very simple idea remains the core of the modern, formal and empirical explorations of the hypothesis-testing brain. Helmholtzian ideas were taken up and developed in various tempi throughout the 20th Century. Jerome Bruner's 'New Look' psychology considered the influence of prior beliefs on perception (Bruner, Goodnow et al. 1956), which was in turn

challenged by Jerry Fodor and Zenon Pylyshyn, though they both accept the basically Helmholtzian notion of (low-level) unconscious inference (Fodor 1983; Pylyshyn 1999). Ulrich Neisser (1967) developed the notion of analysis by synthesis, which has a distinctive Kantian feel; Irvin Rock (1983) developed such ideas further, and Richard Gregory (1980) explicitly modelled his account of perception on Helmholtz's appeal to hypothesis testing (see Hatfield 2002 for overview and discussion). The formal apparatus for harnessing these ideas was presaged by Horace Barlow (Barlow 1958; Barlow 1990) and developed by many working in computational neuroscience and machine learning, in particular Rao, Ballard, Mumford, Dayan, Hinton, and others, while Bayesian approaches to perception were explored and developed by Kersten, Yuille, Clark, Egner, Mamassian, and many others (useful introductions and texts include (Knill 1996; Dayan and Abbott 2001; Rao, Olshausen et al. 2002; Doya 2007; Bar 2011)). There are also recent expositions and discussions of the framework in work by for example (Bubic, Von Cramon et al. 2010; Huang and Rao 2011; den Ouden, Kok et al. 2012). Chris Frith's terrific *Making Up the Mind* (2007) discusses many aspects of the hypothesis-testing brain and provides very many examples of relevant empirical studies.

A related historical undercurrent to the prediction error minimization story concerns developments in our understanding of causation and inductive inference. David Hume is a pivotal character in this regard, as he defined "a cause to be an object followed by another, and where all the objects, similar to the first, are followed by objects similar to the second. Or in other words where, if the first object had not been, the second never had existed" (Hume 1739–40: 146). To Hume, causation is thus both about extracting statistical data and about imagining what happens when the world is intervened upon in a controlled manner. This dual definition was emphasized by Lewis in his counterfactual treatment of causation (Lewis 1973) and developed in a full-fledged analysis of causation in terms of invariance under intervention by Woodward (2003), in tandem with Pearl's seminal works on both aspects of Hume's basic idea (Pearl 1988, 2000). The notion of extracting statistical regularities and the notion of modelling intervention both loom large in the notion of unconscious perceptual inference.

Perhaps it is the confluence of modern developments of both the causal story and the hypothesis-testing story that has made it possible to develop and now apply the notion of prediction error minimization to such a degree that we can see it transform our conception of the mind.

Though there has been discussion of aspects of this kind of theory in philosophy of mind and cognition throughout the last 60–70 years, very little philosophical work has been done on the newest incarnations of the theory. Working partly on the basis of Dretske's influential approach (Dretske 1983), the pioneers in terms of connecting the statistical idea of representation with

traditional philosophical debates have been Chris Eliasmith and Marius Usher (Eliasmith 2000; Usher 2001; Eliasmith 2003, 2005). Rick Grush has contributed in a similar vein and developed compelling theories, for example, of temporal consciousness on this basis (Grush 2004, 2006). Andy Clark is currently developing the framework in extremely interesting ways, showcasing its wide ramifications and in crucial ways taking the framework in different directions than what I argue for in this book (Clark 2012a, 2013). Within epistemology, there is an interesting, related line of research focused on Hans Reichenbach's (1938) example of inferring the existence of the external world from inside a cubical universe, discussed recently by Eliott Sober (2011) in Bayesian and causal terms that anticipates key elements of the framework.

The prediction error minimization theory is in many respects difficult to classify: it is both mainstream and utterly controversial. On the one hand, with al-Haytham and Helmholtz and others it sits at the very historical core of psychology and neuroscience, and with Gregory, Rock, Neisser, and many others this kind of approach has major contemporary support. On the other hand, it has such extreme explanatory ambitions that there are relatively few who would support it beyond accepting that our expectations and prior knowledge do shape or guide perception. Many would agree with the general idea that predictions play a role in perception but few would agree that prediction error minimization is all the brain ever does and that action and attention is nothing but such minimization. Fewer still would agree that this is all an expression of the way organisms like humans self-organize, and, further, that this is something that essentially is founded in variational free energy with a direct link to statistical physics!

Within neuroscience research it is rare to see any fundamental concession to the notion; instead textbook accounts explain perception largely in terms of feature detection in the bottom-up sensory signal, without any strong role for the overwhelming amount of backwards connections in the brain, which are thought to mediate predictions on the prediction error scheme. In contrast, textbooks in computational neuroscience and machine learning routinely include chapters on representational learning that in detail go far beyond the aspects of the theory I discuss here. Judging by the increasing number of published studies that are inspired by the idea, or that discuss their results in the light of the idea I think it will not be many years before some version of the theory will dominate in neuroscience, but this prophecy is of course hostage to empirical fortune.

Within cognitive science and machine learning, versions of the prediction error minimization scheme are, as mentioned, widely acknowledged. Parts of the scheme have roots in connectionism, in particular in the construction of neural networks with back-propagation algorithms, which are error correcting ways of classifying incoming data (Rumelhart, Hinton et al. 1986). However,

it differs in central respects from back-propagation in not being supervised (so it does not need labelled training data). Prediction error minimization uses models that generate data in a top-down fashion rather than classify the bottom-up data, in addition use of generative models works much better in a deep hierarchical setting (Hinton 2007). These aspects clearly mark out the scheme as different from earlier connectionist ideas, and they are behind many of the discussions I focus on throughout the book.

The prediction error scheme seems reasonably well-positioned between two opposed trends in cognitive science. On the one hand there is a top-down approach, which begins with conceptual and functional analysis of cognitive processes and then seeks to reverse-engineer a model of the brain. On the other hand there is a bottom-up approach, which builds biologically inspired neural networks and seeks to learn about which cognitive functions such networks implement (Griffiths, Chater et al. 2010; McClelland, Botvinick et al. 2010). As a philosopher, I am naturally inclined to begin with conceptual analysis and indeed the book begins in this vein. However, one of the great attractions of the scheme is that it lends itself to a very mechanistic approach. Though more evidence is needed it sits well with overall anatomical and physiological facts about the brain and how it works. In particular, it is inspired by the overall flow of relatively distinct forwards signalling in the brain, which is met with massive and more diffuse backwards signals; it sits well with the brain's functional segregation and connectivity; and the different functional elements suits the different kinds of plasticity of the brain very well. This appeals to the scientist in me. The combination presents an attractive package.

My own journey towards the hypothesis testing brain began when Ian Gold and I worked on theories of delusion formation (Gold and Hohwy 2000), and took off when Andreas Roepstorff and I began collaborating in Aarhus around 2001. Together with a motley interdisciplinary group we began deciphering the framework, appreciating its explanatory potential, and began thinking about how it would apply across a number of different topics. Inspired by work by Chris Frith I first explored it through issues in neuropsychiatry (Hohwy 2004; Hohwy and Frith 2004; Hohwy and Rosenberg 2005), before looking at broader issues such as self (Hohwy 2007b) and the general consequences for our conception of cognitive and perceptual function (see the special issue of *Synthese* (Hohwy 2007a) with key contributions from Eliasmith (2007) and Friston and Stephan (2007)). Since then I have looked at core functions of visual perception, introspection, and emotion, as well as attention (Hohwy, Roepstorff et al. 2008; Hohwy 2011; Hohwy 2012). In all these cases I have relied on work by Friston, Frith, and others, and developed the consequences for specific issues. The time has clearly come to not only unify many of these themes but also to stand back and get a more overall sense of what the framework says about the mind.

ABOUT THIS BOOK

The book is intended for philosophers, neuroscientists, psychologists, psychiatrists, and cognitive and computer scientists and anyone interested in the nature of the mind. Readers who are not familiar with the framework can gain an appreciation of it through my simplified rendering of the basic mechanism at its heart, and by seeing how it applies to a range of different problem cases. Readers who are already familiar with the framework will be in interested in how it can connect to a wide range of topics in psychology and cognitive science at large as well as with philosophical problems.

I have sought to explain philosophical debates without too much philosophical jargon. Topics from empirical philosophy and neurophilosophy permeate the book, but I have collected much of the more directly philosophical debate in Chapter 8. Sometimes I provide argument in some detail and sometimes more in the shape of promissory notes, or invitations, for further work. The strength of the book, I hope, lies just as much in the particular suggestions as in the combined package of suggestions. I have also made an effort to describe neuroscientific and psychophysics studies in straightforward and accessible terms. Although I do not provide new empirical evidence for the theory here, I believe my treatment does support the theory by providing a broad-ranging, unifying explanation, which resolves and illuminates some recalcitrant problems and debates in philosophy of mind and cognitive science.

At the end of each chapter I have placed Notes. These provide references and suggestions for further reading as well as textual sources and brief reviews of relevant, additional empirical evidence. Some notes contain brief discussions that are not crucial to the main argument of the chapters in question though they do concern important further aspects of the broader topics discussed. I have included them as notes to indicate how they may relate to the main theme of the book. Finally, some notes provide basic descriptions of some of the formal and mathematical machinery on which the conceptual framework rests.

Part I

The Mechanism

1

Perception as causal inference

Our senses are bombarded with input from things in the world. On the basis of that input, we perceive what is out there. The problem that will concern us is how the brain accomplishes this feat of perception.

This chapter pursues the idea that the brain must use inference to perceive—the brain is an inference mechanism. The first aim is to show why we should agree with this and what the key ingredients of such perceptual inference are. The second aim is to show how inference could underpin the phenomenology of perception.

A very basic and useful formulation of the problem of perception is in terms of cause and effect. States of affairs in the world have effects on the brain—objects and processes in the world are the causes of the sensory input. The problem of perception is the problem of using the effects—that is, the sensory data that is all the brain has access to—to figure out the causes. It is then a problem of causal inference for the brain, analogous in many respects to our everyday reasoning about cause and effect, and to scientific methods of causal inference.

The problem of perception is a *problem* because it is not easy to reason from only the known effects back to their hidden causes. This is because the same cause can give rise to very different effects on our sensory organs. Consider the very different inputs we get from seeing rather than merely touching a bicycle, or seeing it from different perspectives, or seeing it in full view as opposed to being partly obscured behind a bush. Likewise, different causes can give rise to very similar effects on our sense organs. Consider the potentially identical sensory input from different objects such as a bicycle or a mere picture of a bicycle, or a whole bicycle occluded by a bush as opposed to detached bicycle parts strewn around a bush, or more outré possibilities such as it being an unusually well-coordinated swarm of bees causing the sensory impression as of a bicycle.

In our complex world, there is not a one-one relation between causes and effects, different causes can cause the same kind of effect, and the same cause can cause different kinds of effect. This makes it difficult for the brain to pick the one effect (sensory input) that goes with the one cause (object in the

world). If the only constraint on the brain's causal inference is the immediate sensory input, then, from the point of view of the brain, any causal inference is as good as any other. When the input is different, as in the seen and felt bicycle case, the brain would not know whether to infer that the cause of the inputs is the same, or if there are distinct causes, and whether one type of cause is more likely than another.

CONSTRAINTS ON PERCEPTUAL INFERENCE

The key issue is then that without any *additional constraints* the brain will not be able to perform reliable causal inference about its sensory input. We can in fact engage in such inference, since we can perceive. So there must be such additional constraints, but what could they be?

One possibility is that the additional constraints are *mere biases*. Even though the brain cannot reliably infer that it is one rather than another cause, it simply happens to be biased in favour of one. It just so happens that it decides in favour of, say, the bicycle being the cause when it gets a certain kind of input. No doubt there is a describable, law-like regularity in nature such that, in certain to-be-specified conditions, if a system like the brain were to have a certain kind of sensory input caused by a bicycle, then it would be biased towards perceiving it as a bicycle. In principle, various branches of science would be able to discover these biases by systematically exposing systems like the brain to bicycle inputs and tracking the causal chain of events throughout the brain. The brain would seem to cut through the problem of perception by just opportunistically favouring one among the intractably many possible relations between cause and effect.

But even if at some level of description there are these regularities it would not solve the problem of perception as we have conceived it. Such regularities do not afford an understanding of perception as causal *inference*. Inference is a normative notion and brute biases cannot lead us to understand how there could be a difference in quality between an inference back to bicycles rather than, say, swarming bees being the cause of sensory input. What brute regularities in nature give us is a story about what the system *would* do, not what it *should* do in order to get the world right. What is needed, then, is a normative understanding of the role of such regularities. We need to see the additional constraints on causal inference in normative terms.

There is a clear first candidate for an additional constraint with normative impact. It seems obvious that causal inference about things like bicycles draws on a vast repertoire of *prior belief*. This could be what allows us to rank lowly some candidate causes such as it being a swarm of bees that is causing the current sensory impression. Our prior experience tells us that bees are actually

extremely unlikely to form such patterns of sensory input in us. There is in fact little doubt that perceptual causal inference needs to be buttressed with prior knowledge, but doing so is no trivial matter. On the one hand, if the story we tell is that we just find ourselves with a stock of prior beliefs, then we have not after all moved beyond the mere biases type of story. On the other hand, if prior knowledge is itself a product of prior perceptual, causal inference, then we are presupposing what we set out to explain, namely perceptual causal inference—the bump in the carpet has merely shifted.

We can now see what a solution to the problem of perception must do. It must have a bootstrapping effect such that perceptual inference and prior belief is explained, and explained as being normative, in one fell swoop, without helping ourselves to the answer by going beyond the perspective of the skull-bound brain (Eliasmith 2000; Eliasmith 2005). The contours of just such a solution are now beginning to emerge. It is based in probability theory—Bayesian epistemology—which is normative because it tells us something about what we should infer, given our evidence.

PERCEPTION AND BAYES' RULE

Consider this very simple scenario. You are in a house with no windows and no books or internet. You hear a tapping sound and need to figure out what is causing it (Figure 1).

This illustrates the basic perceptual task. You are like the brain, the house is the skull, and the sound is auditory sensory input. As you are wondering about

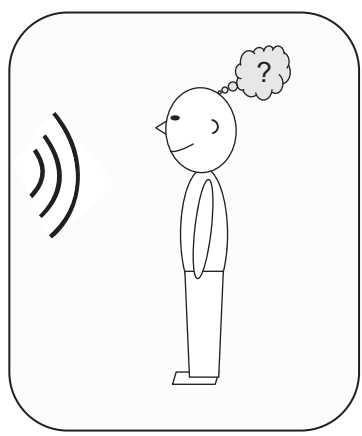


Figure 1. The basic perceptual inference problem: figuring out what caused a sound. This is analogous to the situation for the brain.

the cause of the input, you begin to list the possible causes of the input. It could be a woodpecker pecking at the wall, a branch tapping at the wall in the wind, a burglar tampering with a lock, heavy roadworks further down the street, a neighbour's loud music, or those kids throwing stones; or it could be something internal such as loose water pipes banging against each other. Let your imagination rip: it could be that your house has been launched into space over night and the sound is produced by a shower of meteorites. There is no end to the possible causes. Call each of these possibilities a *hypothesis*. The problem of perception is how the right hypothesis about the world is shaped and selected.

Set aside the problem that once we begin generating hypotheses, there is no clear principle for when we should stop. Consider instead the fact that we *can* generate hypotheses, and that not just any hypothesis will seem relevant. For example, we would not accept that the tapping noise on your house could be produced by a distant mathematician's musings on Goldbach's conjecture, or by yesterday's weather. This means we are able to appreciate the link between a hypothesis and the effects in question. We can say "if it is really a woodpecker, then it would indeed cause this kind of sound". We can say something about how likely it is that the hypothesis fits the effects. This is *likelihood*: the probability that the causes described in the hypothesis would cause those effects. It is clear that assessing such likelihoods is based on assumptions of causal regularities in the world (for example, the typical effects of woodpeckers). Based on our knowledge of causal regularities in the world we can often rank hypotheses according to their likelihood, according to how close their tie is to the effects we are seeking to explain. Such a ranking can be said to capture how good the hypothesis is at accounting for, or *predicting*, the effects. For example, the woodpecker hypothesis may have roughly the same likelihood as the banging pipes hypothesis, and both have higher likelihood than the hypothesis concerning those stone-throwing kids.

We could simplify the problem of perception by constraining ourselves to just considering hypotheses with a high likelihood. There will still be a very large number of hypotheses with a high likelihood simply because, as we discussed before, very many things could in principle cause the effects in question. Just going by the hypothesis with the very highest likelihood does not ensure good causal inference. Here is a hypothesis with very high likelihood: the sound is caused by a tapping machine especially designed by cunning neuroscientists to use you to illustrate perceptual causal inference. This hypothesis fits the auditory evidence extremely well, but it does not seem like a good explanation in very many actual situations. The problem is that the cunning neuroscientist hypothesis seems very improbable when considered in its own right and before you heard the banging sound.

Therefore, we need to take the independent, prior plausibility of hypotheses into consideration, in addition to their likelihood. We need to consider the probability of the hypothesis prior to any consideration of its fit with the

evidence. This is then the *prior probability* of the hypothesis. Perhaps there is some objective truth about how probable each hypothesis is, based on the frequency of the events it describes. This kind of knowledge would be useful but mostly it is not something we have. Instead we will assume you assign probabilities to hypotheses based on your own background beliefs and subjective estimates (making sure the probabilities sum to 1, to make the ranking meaningful).

By appealing to your prior beliefs we have given you two tools for figuring out the cause of the sound: likelihood, which is the probability of the effect you observe in the house *given* the particular hypothesis you are considering right now; and the prior probability of the hypothesis (or just the "prior"), which is your subjective estimate of how probable that hypothesis is independently of the effects you are currently observing.

It seems rational to pick the hypothesis which best fits the observed effects but weighted by the independent probability of that hypothesis. Likelihood and prior are the main ingredients in Bayes' rule, which is a theorem of probability theory and thought by many to be a paradigm of rationality. This rule tells us to update the probability of a given hypothesis (such as the woodpecker hypothesis), given some evidence (such as hearing some tapping sound) by considering the product of the likelihood (which was the probability of the evidence given the hypothesis) and the prior probability of the hypothesis (normalized so probabilities sum to 1). The resulting assignment of probability to the hypothesis is known as the *posterior probability*. The best inference is then to the hypothesis with the highest posterior probability. (A brief primer on Bayes' rule is included at the end of this chapter).

Return now to the sound you hear in the house. With likelihoods and priors you can arrive at a good hypothesis: the one that achieves the highest posterior. If you have experienced many woodpeckers in your area and only a few burglars, and if you don't really think your house is likely to have been launched into space over night, and so on and so forth, then you should end up inferring to the woodpecker hypothesis (Figure 2).

Even on this very simplified presentation, Bayesian inference provides a very natural way to think about perception. Of course, the drawback with illustrating the problem as I have done here is that there is no intelligent little person inside the skull consciously performing causal inference. On the story we shall develop, which goes back to Helmholtz, what is really going on is that the neural machinery performs perceptual inference unconsciously. As Helmholtz says about the "psychical activities" leading to perception,

> [they] are in general not conscious, but rather unconscious. In their outcomes they are like inferences insofar as we from the observed effect on our senses arrive at an idea of the cause of this effect. This is so even though we always in fact only have direct access to the events at the nerves, that is, we sense the effects, never the external objects (Helmholtz 1867: 430).

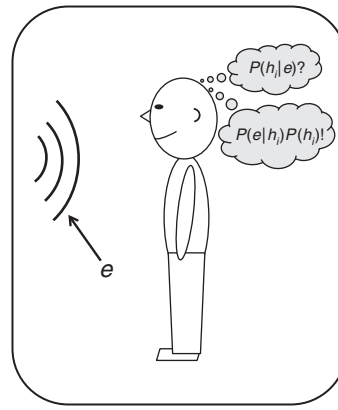


Figure 2. Prior probability of hypothesis h_i: $P(h_i)$. Likelihood that the evidence e would occur, given h_i is the true hypothesis: $P(e|h_i)$. Posterior probability of the hypothesis h_i, given the evidence e: $P(h_i|e)$. Simplified version of Bayes' rule that puts it together: $P(h_i|e) = P(e|h_i)P(h_i)$.

So what we will be talking about is *unconscious perceptual inference*. The job before us is to see how what the system does can be usefully conceived as a form of inference. We just need to accept the Helmholtzian idea that the brain is capable of unconsciously going through the same kind of reasoning that we described for figuring out the cause of the sound heard inside the locked house. The brain infers the causes of its sensory input using Bayes' rule—that is the way it perceives. The core idea is fairly clear and has a pleasing air of generality to it: the problem of perception is not inherently special, something for which an entirely new branch of science is needed. It is, instead, nothing more than a version of the kind of causal inference problem that we are often confronted with in both science and everyday life.

While the Bayesian, inferential approach to perception is attractive many questions quickly arise. Straight off, aligning perception with ideally rational, probabilistic, scientific-style reasoning seems rather intellectualist. It is difficult to learn probability theory and to implement Bayesian inference but perception is unconscious and effortless—it is something adults, children, and animals can do without knowing anything about Bayes. Moreover, there is evidence that we are not very good at explicit Bayesian reasoning—Bayes' rule takes some explaining and exercise so does not seem to come naturally to us (Kahneman, Slovic et al. 1982). There is also something slightly odd about saying that the brain "infers", or "believes" things. In what sense does the brain know Bayes, if we don't?

For that matter, a Bayesian approach to perception does not seem to directly concern the full richness of perceptual phenomenology as much as mere conceptual labelling or categorization of causes of input (it could seem to be not so

much about visually experiencing a bicycle as merely labelling some sensory input "bicycle"). Nor does this approach, with its focus on assigning subjective probabilities, immediately begin to provide a satisfactory explanation of where prior beliefs come from. As we will see in this and the following chapters, the theoretical framework can be developed to deal with all of these issues.

The contrast to the inferential picture of perception is a picture on which perception, rather than being the upshot of inferential processes in a hypothesis-testing brain, is the result of an analytic, bottom-up driven process where signals are recovered from low-level sensory stimulation and gradually put together in coherent percepts. On this alternative, non-inferential approach, perception is driven bottom-up by the features the brain detects in the input it gets from the world. Crudely, changes in input drive changes in perception, and so top-down inference in any substantial, normative sense is not needed.

There is much discussion about the relative virtues of the feature detection approach vs. the more inferentialist, Bayesian approach (for a review and discussion, see Rescorla (in press). One reason for not adopting the feature-detection approach is that it is not very clear how it can help with the problem of perception as we have set it out above. This theoretical debate cannot be resolved conclusively here but in the next section I will give what I think is a very good example of a perceptual effect demonstrating the need for inference.

PERCEPTUAL INFERENCE AND BINOCULAR RIVALRY

In 1593 the Italian polymath Giambattista della Porta reported an intriguing visual phenomenon:

> Place a partition between the eyes, to divide one from the other, and place a book before the right eye, and read; if another book is placed before the left eye, not only can it not be read, but the pages cannot even be seen, unless the visual virtue is withdrawn from the right eye and changed to the left (Porta 1593; quoted in Wade 1998: 281).

Some centuries later Charles Wheatstone invented the stereoscope, which uses mirrors to help split the images presented to the eyes, and in 1838 also described this kind of perceptual alternation between different letters shown to each eye (Wade 1998; Wade 2005). This fascinating effect is known as binocular rivalry and remains, 400 years after Porta, a vibrant focus of much research in vision science. The neural mechanism behind it is still unknown and it keeps throwing up new and intriguing findings. As Porta delightfully puts it, what makes "visual virtue" alternate between the eyes?

It is a surprising effect because one would think that if two different images are shown to the eyes they should just somehow blend in with each