13
PHILOSOPHY OF PSYCHOLOGY

Kelby Mason, Chandra Sekhar Sripada, and Stephen Stich

Introduction

The twentieth century has been a tumultuous time in psychology, a century in which the discipline struggled with basic questions about its intellectual identity, but nonetheless managed to achieve spectacular growth and maturation. It is not surprising, then, that psychology has attracted sustained philosophical attention and stimulated rich philosophical debate. Some of this debate was aimed at understanding, and sometimes criticizing, the assumptions, concepts, and explanatory strategies prevailing in the psychology of the time. But much philosophical work has also been devoted to exploring the implications of psychological findings and theories for broader philosophical questions such as: Are humans really rational animals? How malleable is human nature? and Do we have any innate knowledge or innate ideas?

One particularly noteworthy fact about philosophy of psychology in the twentieth century is that, in the last quarter of the century, the distinction between psychology and the philosophy of psychology began to dissolve as philosophers played an increasingly active role in articulating and testing empirical theories about the mind and psychologists became increasingly interested in the philosophical underpinnings and implications of their work. It isn't possible to distinguish sharply between philosophy of psychology and philosophy of mind but, roughly speaking, philosophy of psychology at the end of the twentieth century and start of the twenty-first has two distinctive features: (a) as just mentioned, it is naturalistic, presenting itself as continuous with natural science; and therefore (b) it typically starts from psychological research, and attends to it closely, where philosophy of mind often starts from folk psychological observation or more general theoretical considerations.[1] We present here a survey of five important themes in twentieth-century psychology which have been the focus of philosophical attention and have benefited from philosophical scrutiny.

Perhaps the two most important events in the history of psychology in the twentieth century were the emergence of the *behaviorist* approach, which dominated

psychology for the first half of the century (see "Philosophy of mind," Chapter 12), and its displacement by *cognitivism* as the century drew to a close. Philosophers played an important role in both of these events, developing a philosophical companion to psychological behaviorism, and clarifying the nature and assumptions of the cognitivist approach to psychological theorizing. Our first section will be devoted to the transition from behaviorism to cognitivism.

The linguist Noam Chomsky (1928–) is widely considered one of the founders of the cognitivist approach in psychology; he was also the central figure in a second major movement in contemporary psychology, *nativism*, which is the topic of our second section. Nativism and empiricism have traditionally been philosophical doctrines regarding the structure of the mind and the sources of the justification of belief. Contemporary cognitive psychology complements and extends traditional philosophical inquiry by providing a sophisticated methodology for investigating nativist structures in the mind. But nativism is a problematic notion, and in this section we discuss philosophical attempts to clarify both what nativism claims and what it entails about human nature.

Nativism is closely related to, but importantly distinct from, another topic which has been center-stage in both psychology and the philosophy of psychology for the last two decades: *modularity*. Jerry Fodor's seminal *Modularity of the Mind* (1983) suggested that an important structural principle in the organization of the mind is that at least some cognitive capacities are subserved by specialized, largely independent subsystems. Recent work in psychology has sought to delineate the modular structure of a number of important cognitive capacities, including language and mathematical cognition. Philosophy of psychology has contributed to this endeavor by helping to clarify the notion of a module. Philosophers have also debated whether modularity, if it is true, forces us to abandon traditional views about the transparency of the mental and to reconsider prevailing accounts of epistemic justification. These are some of the issues we'll consider in the third section "Modularity," below.

A fourth theme which attracted a great deal of attention during the twentieth century, both in philosophy and in psychology, is *rationality*. Philosophers have traditionally debated the nature and the extent of both theoretical rationality, or the rationality of belief, and practical rationality, or the rationality of action. During the last three decades of the twentieth century psychologists became increasingly interested in these topics as well, and a large experimental literature emerged exploring the ways in which people actually reason and make decisions. Much of what they found was both surprising and troubling. In the fourth section, we recount some of the more disquieting empirical findings in this tradition and consider how both philosophers and psychologists have attempted to come to grips with them.

The final topic we consider is *intentionality*. Cognitive psychology, as it is currently practiced, appears to be committed to the existence of intentional states, like beliefs, desires, plans, goals, and fears, which are conceived of as being *representational* – they are *about* states of affairs in the world. Thus it would seem that cognitive psychology must grapple with "Brentano's problem," the problem of how to accommodate intentional notions within a naturalistic view of the world (see the discussion in "Philosophy

of mind," Chapter 12). In our final section we discuss debates surrounding attempts by philosophers to "naturalize" the intentional. It may well be the case that more philosophical ink has been spilled on the topic of naturalizing the intentional than on the preceding four topics combined. Despite this, we suspect that the question of how to naturalize the intentional is not well posed, and the importance of the answer is far from obvious. Perhaps a new century of philosophy of psychology will decide that the question, once made more precise, was not worth all the fuss.

From behaviorism to cognitivism

A great deal of contemporary psychological research and theorizing is unabashedly *mentalistic*. Psychological theories explain outward behavior by positing internal psychological states and structures such as beliefs, desires, perceptions, memories, and various and sundry other kinds of mental states. In the first half of the twentieth century, however, a very different ethos prevailed. Psychological theorizing was dominated by *behaviorist* thinking in which the positing of unobservable mental entities was explicitly shunned. Perhaps the most important event in the history of psychology in the twentieth century was the demise of behaviorism and the rise of *cognitivism*, a thoroughly mentalistic approach to mind. We'll begin our survey of key issues in twentieth-century philosophy of psychology by discussing this transition.

Logical behaviorism

There are actually two quite distinct versions of behaviorism that flourished in the first half of the twentieth century, one primarily in philosophy and the other primarily in psychology. *Logical behaviorism*, which prevailed primarily in philosophical circles, is a thesis about the meaning of mental state concepts (see also "The development of analytic philosophy: Wittgenstein and after," Chapter 2). According to logical behaviorists, mental state concepts such as belief or desire don't refer to hidden, and potentially mysterious, internal states of a person. Rather, talk of mental states is actually talk about dispositions to behave in certain ways under certain circumstances. For example, consider the claim that Paul has a headache. Superficially, this claim appears to refer to some inner state of Paul, perhaps some "achy" subjective sensation. But according to logical behaviorists, this appearance is mistaken. Talk of Paul's headache actually refers to a complex set of dispositions, for example the disposition to groan, wince, avoid bright lights, reach for the aspirin, say "Ouch" when he moves his head too quickly, and so on. (See Ryle 1949 for the classic exposition of logical behaviorism.)

One of the main attractions of logical behaviorism is that it provides an account of the reference of mental state concepts without positing anything metaphysically spooky or mysterious, such as a Cartesian *res cogitans*. If mental state concepts are about behavior, then their materialistic bona fides cannot be denied. A second attraction of logical behaviorism is epistemological. People routinely attribute states such as beliefs and desires to others. If beliefs and desires are understood as hidden

inner states of a person, then it is hard to see how we might come to have knowledge of these states in others, and the potential for skepticism regarding other minds looms large. On the other hand, if mental states are understood as dispositions to behave in certain ways, as logical behaviorists contend, then our knowledge of these states is readily explained and skepticism about other minds is dispelled.

Despite its attractions, logical behaviorism ultimately foundered. A key stumbling block for the program is that there appears to be no *straightforward* connection between mental states and dispositions to engage in certain behaviors in the way that logical behaviorism appears to require. Rather, connections between mental states and behavior invariably appear to be mediated by a number of factors, most notably by other mental states. Let us return to the example of Paul who has a headache. The logical behaviorist says that Paul's having a headache means that Paul is disposed to engage in a host of behaviors, including, among others, reaching for the aspirin. But it seems that whether Paul does in fact reach for the aspirin depends critically on Paul's other mental states, including his other beliefs, desires, preferences, etc. For example, suppose that Paul believes that aspirin will upset his stomach, or that Paul dislikes medicines and prefers natural remedies, such as a massage. In each of these cases, Paul will not reach for an aspirin.

The lesson for logical behaviorism, as philosophers such as Donald Davidson and Jerry Fodor emphasized, is that beliefs, desires, and other mental states are embedded in a dense network of other mental states, and these states invariably act only in *concert* in the production of behavior (Davidson 1963; Fodor 1968). The systematic causal interdependency of mental states in the production of behavior makes it impossible to assign to each mental state its own unique set of behavioral ramifications in the way that logical behaviorists envisioned.

Psychological behaviorism

Even if one rejects logical behaviorism's claims about the centrality of behavior in the meaning of mental state concepts, one might still insist on the centrality of behavior in the formulation of psychological *explanations*. This latter view is at the core of the position often called *psychological behaviorism* or *methodological behaviorism*. According to psychological behaviorism, psychologists should restrict themselves to describing relationships between observable external features of the organism, for example relationships between histories of stimuli impinging on the organism and behavioral responses, without invoking hidden internal states of the organism. Psychological behaviorism is independent of logical behaviorism, since one can be a psychological behaviorist and still maintain that mental states such as beliefs and desires exist and that our concepts of belief and desire and other mental states do in fact refer to these hidden causes of behavior. But one might also be a psychological behaviorist and deny these claims, that is, one might be a psychological behaviorist *and* a logical behaviorist, and in fact many theorists were.

There are two important reasons why psychological behaviorists wanted to dispense with talk of mental states. The first reason is broadly epistemological. Many psychological behaviorists were attracted to the positivist doctrine that scientific theories should only invoke explanatory entities that are publicly observable. Since mental states are "inner" events that are not publicly observable, and there is no observational test to determine when such states occur, mental states are not the sorts of things that should be invoked by scientific psychological theories.

The second motivation for psychological behaviorism is in many ways more interesting and more deeply revealing about why psychological behaviorism ultimately failed as a research program. Many psychological behaviorists saw the task of psychology as formulating law-like generalizations about behavior. In pursuing this endeavor, they viewed the positing of mental states that mediate between environmental inputs and behavioral outputs as *explanatorily superfluous*. The basic idea is that if we assume that there are lawful connections between environmental stimuli (or histories of exposure to environmental stimuli) and inner mental states, and lawful connections between inner mental states and the production of behavior, then there will be lawful connections between environmental stimuli and behavior that can be stated without adverting to inner mental states as mediators. The positing of inner mental states does no additional explanatory work, and given that these inner states are epistemically inaccessible, psychology is better off without them. B. F. Skinner (1904–90) puts the point succinctly:

> If all linkages are lawful, nothing is lost by neglecting a supposed nonphysical (mental) link. Thus, if we know that a child has not eaten for a long time, and if we know that he therefore feels hungry and that because he feels hungry he then eats, then we know that if he has not eaten for a long time, he will eat. (Skinner 1953: 59)

Psychological behaviorism rose to prominence with John Watson's (1878–1958) influential 1913 manifesto "Psychology as the behaviorist views it" (Watson 1913), and the movement held sway, especially in North America, for about five decades. Among the most significant of the behaviorists' accomplishments was the formulation of a number of important learning rules. An example of such a rule is Thorndike's Law of Effect, named after the behaviorist psychologist Edward Thorndike (1874–1949), which says very roughly that if an organism performs some behavior X and X is followed by reinforcement, then the probability of the organism performing X again in the future increases (Thorndike 1911: 244). Behaviorists formalized many such learning rules and generated a substantial body of empirical results demonstrating how these rules successfully predict animal behavior, at least in certain experimental contexts. But despite their successes, there were two fundamental problems with psychological behaviorism that eventually led to the movement's mid-century demise.

One problem with psychological behaviorism was that it dogmatically insisted that the purpose or goal of psychological explanation is, and rightfully should be, the prediction of behavior. But this claim is actually quite puzzling. Even if we grant, for argument's sake, that positing internal mental states and processes is not *necessary* for the prediction of behavior, it does not follow that the elucidation of these states

and processes is thereby rendered uninteresting and unworthy of study. The question of what is, or what is not, interesting or worthy of scientific study simply cannot be legislated in this way. Ironically, in the last several decades, psychologists and neuroscientists have had enormous success in identifying the neural and synaptic mechanisms that underlie several of the learning rules that behaviorist psychologists such as Pavlov, Thorndike, and Skinner originally formulated. (For example, see Rescorla and Wagner 1972 and Kandel et al. 2000.) It seems, then, that psychological behaviorism was needlessly restrictive. The identification of behavioral regularities and the elucidation of internal states that mediate these regularities are *both* worthy projects, and there is simply no reason to privilege one at the expense of the other.

A second problem with psychological behaviorism is the fact that behaviorist explanations appear most plausible when applied to relatively simple animals confronted with certain highly restricted, and indeed somewhat contrived, experimental tasks (for example, pigeons that learn to peck at a lever to obtain food pellets). But when we consider more sophisticated kinds of behavior displayed in more realistic ecological contexts, the idea that this behavior might be explained by simple learning rules such as the Law of Effect begins to look much less plausible. For example, the human capacity to produce and comprehend language is extraordinarily complex, and, as we shall see in the following section, behaviorist accounts of language are particularly implausible.

Noam Chomsky and the rise of cognitivism

Perhaps no figure is more closely associated with the downfall of psychological behaviorism than Noam Chomsky. In a series of influential works, including his 1965 classic *Aspects of the Theory of Syntax*, Chomsky initiated an alternative approach to the study of language that deviated from the explanatory strategy of psychological behaviorism in two crucial respects. (For further discussion of Chomsky, see "American philosophy in the twentieth century," Chapter 5 and "Philosophy of language," Chapter 9.)

First, while behaviorists sought to explain the acquisition of all patterns of behavior by means of just a few domain-general learning mechanisms, Chomsky argued that many kinds of behaviors, or behavioral capacities, are importantly *innate*. Chomsky pressed this claim even in the behaviorists' favored turf of the explanation of relatively simple behaviors in lower animals. For example, in his seminal critique of Skinner's (1957) *Verbal Behavior*, Chomsky noted that many animal behaviors, such as the gaping response of the nestling thrush (a species of bird), appear to be innate in the sense that these behaviors emerge early, reliably, and without the need for learning (Chomsky 1959). In later works, Chomsky assembled a series of arguments that suggested, quite persuasively, that human language too is importantly innate. For example, according to Chomsky's celebrated *poverty of the stimulus argument*, human languages are too complex and the inputs available for the child to learn these languages are too meager for language acquisition to be explicable on the behaviorist model of learning. We'll discuss the poverty of the stimulus argument, and the broader issue of innateness claims in psychology, in more detail in the section entitled "Nativism," below.

The second respect in which Chomsky deviated from psychological behaviorism is that, while behaviorists eschewed positing internal mental states, he argued that in the explanation of complex behavior, the postulation of such states is *unavoidable*. A striking fact, largely ignored by behaviorists, but emphasized by Chomsky, is that language is *productive*; a competent speaker can produce, comprehend, and make judgments about an infinite (or at least unbounded) number of sentences (Chomsky 1965). For example, a competent speaker can judge any of a potentially infinite number of sentences to be grammatical or ungrammatical. Behaviorists are hard pressed to explain the productivity of language, since they would seemingly need to appeal to an infinite number of stimulus-response links. Even worse, speakers can readily produce and comprehend sentences *that they've never heard before*, suggesting that the behaviorist's explanation of these abilities in terms of previously learned stimulus-response links is singularly implausible.

Chomsky argued that the explanation of a productive capacity like language demands that we acknowledge the existence of a sophisticated set of formal processes inside the head of the person. In particular, he postulated the existence of a mentally represented set of rules called a *generative grammar* that underlies speakers' grammaticality judgments. A key feature of a generative grammar is that it specifies *recursive* rules for how words and phrases can be assembled into sentences, thus explaining how speakers with finite minds are able to form judgments of the grammaticality of any of an infinite number of sentences.

Chomsky's work in language was part of a broader effort already under way by other theorists including George Miller in psychology and John McCarthy (1927–), Allan Newell (1927–92), and Herbert Simon (1916–2001) in artificial intelligence (see Gardner 1985). These theorists were heavily influenced by new developments in mathematical logic and computer science, and together, they laid the foundations for an alternative to behaviorism in the study of the mind called *cognitivism*.

The central hypothesis of cognitivism is that the mind can be understood as a kind of computer. Computers process information by executing step-by-step operations, called *algorithms*, on internally encoded bodies of information, called data structures or *representations*. Very roughly, representations are elements within the computer that *stand for* objects and properties in the world. For example, consider a computer that plays checkers. One way such a computer might work is that it might possess internal elements that stand for the various pieces on a checkerboard. By manipulating these internal representations in the appropriate ways, the computer can identify moves that are more likely to lead to unsuccessful outcomes and avoid moves that lead to unsuccessful outcomes.

Cognitivists claim that minds process information in much the same way as computers. Chomsky's explanation of language, which appeals to recursive algorithmic processes that operate over internal representational structures, is a classic example of a cognitivist theory. Cognitivist explanations of other capacities, including perception, categorization, reasoning, and many others, take a similar form (see the chapters in Posner 1989). Cognitivism has flourished since the 1960s and has been an enormously fruitful research program. A central achievement of cognitivism is that it has thrown

off the behaviorists' strictures against talking of the mental and made it possible to develop a rigorous theory of the mind, and not just of behavior. However, in opening up the realm of the mental to theoretical exploration, cognitivism also raises a number of fundamental questions that are of great interest to the philosophy of psychology. In the next four sections of this chapter, we'll explore four important issues in the philosophy of psychology that are, in one way or another, deeply influenced by the cognitivist turn in psychology.

Nativism

The nature of nativism

In the previous section, we noted that Noam Chomsky was a central figure in the development of cognitivism. He also played a pivotal role in the emergence of a second theoretical perspective that has loomed large in contemporary cognitive psychology: *nativism*. Nativists believe that the mind comes equipped with a great deal of innate structure and this innate structure plays an important role in explaining our mature cognitive capacities and abilities. It is useful to distinguish three types of cognitive structure that nativists might endorse as being innate: *concepts* (e.g. "triangle," "God," or "cow"), *beliefs* or *bodies of information* (e.g. mathematical beliefs, geometric beliefs, or beliefs about physical objects) and *mechanisms* (e.g. mechanisms for language acquisition, mechanisms for reasoning). Nativists might endorse the innateness of any the preceding types of cognitive structure, and historically, nativists of various stripes have endorsed the innateness of various combinations of these three (see Cowie's (1999) lucid discussion).

Nativists are opposed by empiricists, who argue that the mind comes equipped with relatively little innate structure, and this structure is relatively unimportant in explaining our mature cognitive capacities and abilities. For example, one particularly extreme version of empiricism, sometimes attributed to the British empiricists of the eighteenth century such as Locke and Hume, claims that the newborn infant's mind is a tabula rasa, or blank slate. According to these theorists, experience is the source of almost all of our mature concepts and beliefs as well as our mature cognitive abilities and capacities. It's important to recognize that even extreme empiricists don't claim that the mind possesses *no* innate structure whatsoever. Rather, empiricists attempt to explain the development of our mature cognitive repertoire by adverting to a minimum of innate structure. Thus, they typically assert that the mind comes equipped with just a few learning mechanisms and these mechanisms are *domain-general*; that is, they operate over a wide variety of cognitive domains. Paradigm examples of empiricist learning mechanisms include associative learning mechanisms (e.g. Pavlovian conditioning) and general-purpose inductive learning mechanisms (e.g. Bayesian approaches to learning).

During the first half of the twentieth century, empiricist ideas dominated psychology and other sciences of human behavior. As we saw in the previous section, behaviorist psychologists such as Watson and Skinner emphasized the role of learning, and in particular, histories of conditioning or reinforcement, in the explanation of behavior.

A similar picture prevailed in the social sciences. Anthropologists such as Franz Boas (1858–1942) and sociologists such as Émile Durkheim (1858–1917) denied that there were rich innate features of the mind shared by all humans. Like behaviorist psychologists, these theorists viewed the human mind as a tabula rasa, although they emphasized the role of social learning and cultural conformity, rather than histories of conditioning or reinforcement, as the primary shapers of human behavior (see Laland and Brown 2002: 53–4).

By the mid-1960s however, the tide had turned. Empiricism no longer enjoyed a position of unquestioned dominance as psychologists increasingly began to emphasize the innate basis for a number of cognitive capacities. Chomsky was a key figure in the resuscitation of nativism. He pointed out that there are a number of features of language that suggest that important aspects of language are in fact innate. For example, Chomsky noted that language is universal in all human groups, and within each human group, it is reliably the case that virtually all normal individuals achieve competence in the native language. These features are predicted on the hypothesis that important aspects of language are innate, and less readily explained on the hypothesis that language is acquired by general-purpose learning.

Perhaps Chomsky's most influential argument for the innateness of language is the *poverty of the stimulus argument*. According to Chomsky, children learn their native language with remarkable ease and rapidity, despite the fact that children are seldom explicitly instructed in their language, and their linguistic experience consists of little more than a fairly limited set of examples of often degraded adult speech. Thus, Chomsky claimed, there is a gap between the learning target achieved by the child, i.e. the child's mature linguistic competence, and the meager inputs available to the child. He argued that the only way to bridge this gap is to postulate that the child antecedently possesses extensive *innate* knowledge of language and brings this knowledge to the language learning task.

Since Chomsky's seminal early investigations of language, nativism has flourished as a theoretical perspective in cognitive psychology. For example, in addition to language, theorists have proposed that there are innate mechanisms or innate bodies of information that subserve our capacities to attribute intentional states to others (Leslie 1994), explain and predict the motion of middle-sized physical objects (Spelke 1988), classify animals and plants (Atran 1998), and many other abilities as well (see the essays in Hirschfeld and Gelman 1994). It's worth emphasizing, however, that although nativist hypotheses were certainly important and influential at the end of the twentieth century, they were by no means uncontested, and vigorous debate between nativists and empiricists raged on (see Elman et al. 1996 for an important defense of broadly empiricist approaches to cognitive development).

Problems with the notion of innateness

As we've seen, the crux of the argument between nativists and empiricists concerns the quantity of innate structure in the mind and the importance of this innate structure in explaining our mature cognitive capacities. Notice, however, that in formulating the

preceding core disagreement between nativists and empiricists, we helped ourselves to the term "innate." What exactly does this term mean? As it turns out, there is no consensus on what it means to say that some element of mental structure is innate, and different theorists often appear to mean quite different things by the term. Given the plethora of meanings, debates between nativists and empiricists are often fraught with confusion, with theorists talking past one another.

An important ongoing project for philosophers of psychology is to clarify the meaning of "innate" and thus help to reduce some of the controversy that surrounds the use of the term in psychological theorizing. This is no easy task since theorists often emphasize different threads of meaning in different contexts. Let us examine a few ways in which the concept of innateness is used in the contemporary literature.

One quite popular way of characterizing innateness is in terms of *developmental invariance* (the phrase is from Samuels 2002). According to these invariance accounts, a trait is innate if it reliably develops over a wide range of normal environments. Put another way, an innate trait is one that is not dependent on any *specific* developmental environment; it is not the case that the trait develops in one way in one environment and another quite distinct way in another environment. Descartes appears to have had developmental invariance in mind when he suggested that certain diseases which recur in families, such as gout, serve as a model for understanding innateness. In *Comments on a Certain Broadsheet*, he notes that diseases such as gout are not literally inborn in the sense that "babies of such families suffer from these diseases in their mother's womb" (Descartes 1985: 304) Rather, he argues, they are innate in the sense they reliably emerge during the course of development, and their emergence does not depend on any specific environment (see also Stich 1975).

Notice, however, that even if a trait reliably develops over a large range of environments, there will inevitably be certain environments in which the trait does not emerge. For example, if we deprive an immature organism of important nutrients or expose it to drugs like thalidomide, its development will be severely stunted and certain traits of the organism that would otherwise reliably develop might not develop at all. Should we, on this basis, discount these traits from being considered innate? A more reasonable alternative is to relativize our account of innateness to some range of "normal" environments. But which range of environments should we count as normal? One proposal might be to count an environment as normal if it is sufficiently similar to the environment in which the organism evolved. Another proposal counts an environment as normal if it is one that is compatible with the organism's "surviving and thriving" (Kitcher 1996: 243). Neither of these proposals is terribly precise and they both leave a good deal of room for disagreement about what counts as a normal environment.

Invariance accounts capture one conception of innateness that is particularly widely used in the biological sciences, although it is found in psychology as well. There is another important meaning of "innate," however, that is much more specific to psychology. Recall that certain leading empiricists argued that the mind is a tabula rasa at birth and that most of a person's stock of concepts, beliefs, abilities, and capacities are acquired through experience, that is, they are *learned*. Implicit in this doctrine is the idea that elements of mental structure are of one of two types: they are

either innate or learned. This provides us with another characterization of the notion of innateness: an element of mental structure is innate if it is not learned.[2]

A problem with the preceding account, however, is that learning is itself a contested notion, and there are a number of problematic cases in which it is difficult to know whether to classify a particular kind of acquisition process as an instance of learning. For example, consider Chomsky's highly influential model of language acquisition, which he calls the "Principles and Parameters Model" (Chomsky 1988). According to this model, the language faculty is associated with a set of parameters which can be set in various permissible ways (in most cases just two). For example, one putative parameter governs the ordering of phrases in a sentence. English is subject-verb-object, Hindi is subject-object-verb, while virtually no languages are object-subject-verb. According to Chomsky, the linguistic experience that the child confronts sets the parameters associated with the language faculty, thus accounting for the child's mature language competence (1988: 133–4). If language acquisition does in fact consist of parameter-setting, as Chomsky proposes, it is hard to know whether we should count this acquisition process as an instance of learning. Chomsky himself has argued that language is not in fact learned. Rather, he claims, the language organ simply "grows" in much the way the heart or any other somatic organ grows (1988; see Cowie 1999 for a critique).

In addition to the two accounts of innateness discussed above, there are a number of others. The biologist Patrick Bateson (1938–) has identified at least seven important usages of the notion of innateness in the literature. According to him, theorists at different times use "innate" to mean:

- caused by genetic factors
- caused or driven by internal factors
- shared by all members of the species
- adapted by natural selection over the course of evolution (Bateson 1991)

Given that there are a large number of quite distinct accounts of innateness in the literature, it is tempting to ask which of them is correct. In our view, this temptation should be resisted, since it is unlikely that there is a unique right answer. Rather, we think, "innate" has several legitimate meanings, and each meaning has its uses in the context of distinct scientific and explanatory projects. The philosopher Paul Griffiths suggests that in order to avoid confusion, scientists should simply specify which among the many meaning of innateness they intend to signify by their use of the term (Griffiths 2002). For example, a scientist who intends "innate" to connote *not learned* should simply specify this so as to not confuse her meaning of the term with other meanings. We think this is an eminently reasonable proposal.

The malleability of human nature

While we are ecumenical in endorsing a number of different legitimate uses of "innate," there is one use of the term that is worth flagging because it is particularly problematic. It is sometimes assumed that if some element of mental structure is

innate, then it is more or less fixed, implastic, and resistant to change or environmental manipulation. This usage of the term suggests that an innate trait is one that is more or less *immalleable*.

However, it's important to realize that while there may be a connection between a trait's being innate and its being immalleable, this connection must be explicitly defended on a case by case basis, and it is certainly not inevitable. To take a stock example from the literature, phenylketonuria is a genetic disorder in which an affected individual is unable to break down the amino acid phenylalanine from the diet, and the disorder was once routinely life-threatening. However, advances in medical science have made low phenylalanine diets and phenylalanine-free formulas widely available, thus mitigating much of the illness associated with the disorder. The cluster of life-threatening symptoms associated with phenylketonuria is an example of a trait that is innate according to several of the accounts listed above, but it is nonetheless susceptible to substantial environmental manipulation.

Too often, theorists assume that if a trait is innate it is immalleable. This is unfortunate because the question of the malleability of human nature is an emotionally-charged issue with important social, political, and normative implications. The debate between nativism and empiricism is quite distinct from the debate over the malleability of human nature, and keeping these two debates separate will eliminate much unnecessary controversy.

Modularity

In 1983, Jerry Fodor revived a tradition of faculty psychology which he traced back, tongue only partly in cheek, to Franz Joseph Gall (1758–1828), the founder of phrenology (Fodor 1983). On Fodor's view, then-contemporary cognitive science had produced evidence that the human mind contained a number of distinct cognitive mechanisms, which were dedicated to specific tasks; Chomsky (1980) called such mechanisms "mental organs," but Fodor called them "modules." This view, that the mind is at least partly modular, can be contrasted with the view that it is composed only of mechanisms and processes which are entirely domain-general – that is, which can operate indifferently on all tasks and topics. Attention, memory, and reasoning are putative examples of domain-general processes.

Fodorian modules

Fodor did not seek to define the term "module" with necessary and sufficient conditions. Rather, he offered a characterization of the sorts of mechanisms that seemed to be invoked by cognitive science, in particular vision science and linguistics. Fodorian modules are characterized by having a cluster of the following features, each of which is itself a matter of degree:

1 Their operation is mandatory (i.e. they respond automatically to input).
2 Their operation is fast.

3 They are domain-specific (i.e. they operate on a limited range of inputs, defined by some task such as domain like vision or language processing).
4 They are neurally localized.
5 They are informationally encapsulated (i.e. they have limited access to information in other systems).
6 Other mental systems have only limited access to their computations.
7 Their outputs are shallow (i.e. not very conceptually elaborated).
8 They show characteristic and specific breakdowns.
9 Their development shows a characteristic pace and sequence.

The most important of these features, for Fodor, was informational encapsulation. A clear example of encapsulation is the persistence of perceptual illusions like the Müller–Lyer illusion in Figure 1. Even when we know that the two lines are the same length – by measuring them, say, or having drawn them ourselves – they continue to appear of different lengths. The information that they are the same length can't "get into" the visual system.

Figure 1: Müller–Lyer illusion

Because of their informational encapsulation, Fodor argued that modules would only be found at the input and output sides of the mind. To function properly, central processes such as reasoning, decision-making, and belief-fixation must be able to access and integrate information from many different domains and different sources. Moreover, it seems that central systems do in fact have such general access to information; therefore, Fodor maintained, they can't be modular.

Evolutionary psychology and massive modularity

Since the publication of *The Modularity of Mind*, the notion of modularity has been highly influential in cognitive science. And, *pace* Fodor, some psychologists and philosophers have posited the existence of modules in more central domains of cognition. For instance, Scott Atran has argued that there is a module for folk biology—people's intuitive understanding of the living world (Atran 1998); Alan Leslie has argued that there is a module for folk psychology—people's understanding of other people's mental states (Scholl and Leslie 1999); while Simon Baron-Cohen has argued that folk psychology is subserved by several distinct modules (Baron-Cohen 1995).

The most extreme statement of this trend can be found in evolutionary psychology, and in particular the work of Leda Cosmides and John Tooby. Not only have Cosmides

and Tooby posited the existence of modules for central processes such as cheater-detection (e.g. Cosmides and Tooby 1992), but they have also offered a general evolutionary argument that we should expect the mind to be, in Samuels' (1998) phrase, "massively modular" (Cosmides and Tooby 1994). That is, there are supposedly general evolutionary reasons to expect that the mind (including central cognition) will be largely or even entirely composed of modules rather than domain-general processes.

Cosmides and Tooby offer two main reasons: first, our ancestors would have faced different adaptive problems – e.g. foraging, navigation, mate selection – which required different sorts of solutions. An organism with domain-specific ways to solve these problems would have been faster, more efficient, and more reliable than a "jack of all trades" organism which could only solve them in some domain-general way; therefore modular organisms would have been selected over "jack of all trades" organisms in our ancestral lineage. Thus our own evolved cognitive architecture is likely to be massively modular.

The second putative reason to expect massive modularity is that only massively modular minds could have produced "minimally adaptive behavior in ancestral environments" (Cosmides and Tooby 1994: 91). "Jack of all trades" organisms could not have learned by themselves and in their own lifetimes the benefits of avoiding incest or helping kin, especially since what counts as error and success differs from one domain to another. Creatures with domain-specific knowledge of what to do and when to do it would have selective advantage over "jack of all trades" creatures who had to figure it all out themselves.

Even if we grant the general soundness of these just-so stories, they fall far short of establishing that the mind is massively modular in the Fodorian sense of "module," which is that of a *computational* module. Computational modules are distinct computers which only interact with the rest of the mind at the input and output ends, and which contain their own proprietary mental operations.[3] By contrast, there is another, non-computational notion of modularity which Samuels et al. (1999) call a *Chomskian* module. Chomskian modules are mentally represented bodies of domain-specific knowledge which are supposed to underlie our cognitive abilities in various domains. They get their name from Chomskian linguistics, which posits the existence of such a module for grammar: an internally represented body of knowledge of the grammar of our language which explains our ability to comprehend, produce, and make judgments about sentences (Chomsky 1980). As noted in the previous section, developmental psychologists have posited the existence of such domain-specific knowledge for other domains such as intuitive physics (Carey and Spelke 1994) and number (Gelman and Brenneman 1994). Whereas computational modules are distinct little computers, Chomskian modules, by contrast, are merely distinct databases of information about specific domains; these databases may be looked up and used, not by specialized computers, but by general cognitive processes.

If the arguments from evolutionary psychology show anything at all, it is the need for some domain-specific knowledge of the sort contained in Chomskian modules. Perhaps successful organisms do need substantial amounts of knowledge about the adaptive problems their ancestors would have faced. But, as Samuels (1998) has

argued, this does not support the existence of separate computational modules. For it is entirely possible that all this domain-specific knowledge is operated on by the same domain-general cognitive processes. It is one thing to argue that the mind must have a vast library of domain-specific information; it is another thing to show that it must also have a vast network of different computers dedicated to using that information.

Philosophical implications of modularity

Why should philosophers care whether the mind is composed of modules, of either the computational or Chomskian sort? Apart from the inherent empirical interest in uncovering the structure of the mind – and there is a long history of philosophers speculating about this structure – there are more directly philosophical reasons why it matters. For one thing, the issue of modularity is relevant to the nativist–empiricist debate discussed in the previous section. Granted, the mere existence of a module does not itself show that the module is innate, since some features of modules (like informational encapsulation and neural localization) might come about through individual development (Karmiloff-Smith 1992). Even so, if the mind did turn out to be considerably modular, it would be implausible that *none* of that structure was innate. For that would mean that the exact same cognitive architecture and domain-specific information had been constructed in a billion different heads, with a billion different learning histories, by the undirected operation of purely domain-general processes, which is singularly unlikely. Substantial modularity, then, would be supportive of at least some nativist cognitive structure.

Substantial modularity might also seem to undermine the possibility of epistemic justification. Consider, for instance, a coherentist theory of justification, according to which a belief is justified if and only if it belongs to a coherent set of beliefs. This naturally implies that organisms that want to have justified beliefs ought to try to get their beliefs to mutually cohere. But if the organism has a cluster of modules, and if some of the information in those modules counts as *beliefs*, then it may find this task impossible. Whether it has fully computational or merely Chomskian modules, some of its beliefs may well be out of its control, immune to revision or rational consideration. In such a case, Quinean assertions to the contrary notwithstanding (e.g. Quine 1951), not every strand in the web of belief would be revisable, not even in principle.

More generally, modularity raises the prospect of epistemic boundedness, since our cognitive architecture might place regrettable limits on the possibility of our epistemic progress towards the truth. To illustrate this possibility, we adapt an example from Fodor (1983), who likened his input/output modules to a reflex such as protective blinking. Suppose your helpful neighbor sees a mote in your eye and, deciding to remove it, quickly moves her finger to your eye. No matter how much you trust your neighbor, your blink reflex will kick in and you will flinch from her approaching digit. The belief that your neighbor is unlikely to harm you etc., has no effect on your reflexive behavior. To move to a more cognitive case, while staying at the level of

input processes, recall the Müller–Lyer diagram in Figure 1. Here your belief that the lines are of the same length has no effect on your perception; they still look unequal.

So far this is epistemically harmless; indeed, as Fodor points out, it's positively beneficial that creatures' wishful thinking can't impair the accuracy of their perception. But now suppose that certain more central parts of cognition were modular. This could either mean, as in the case of computational modules, that certain processes of belief-fixation are informationally encapsulated and therefore cannot be affected by all the available information in the mind; or, as in the case of Chomskian modules, it could mean that certain beliefs are relatively fixed. Either way, some of your beliefs would not be sensitive to everything you know; and while your modules might be good at dealing with certain problems in their domain, they might be very bad at dealing with other problems they are required to handle. In other words, you would be epistemically bounded.

A modular agent might actually be epistemically bounded in two different ways. First, the agent might be bounded with respect to the thoughts it can entertain. If it has different computational modules for working on different domains, then it might not be able to entertain thoughts with contents which cross those domains. Second, a modular agent might be epistemically bounded with respect to the information it can access in assessing a certain thought. Indeed, this type of boundedness is the very essence of informational encapsulation.

Human beings certainly don't appear to be epistemically bounded in the first way. Not only can we entertain thoughts about minds, bodies, numbers, animals, and so on, we can entertain thoughts that are about any combination of these things. The flexibility of our thought has thus been offered as an argument that our cognition couldn't possibly be *massively* modular and that there must be some sort of domain-general, "central workspace" (this argument is developed in Fodor 2000; see Carruthers 2003 for a response).

By contrast, the extent to which humans are epistemically bounded in the second way is still an open question. On the one hand, our abductive reasoning processes generally seem to be able to use many sorts of information. In trying to predict what Jean Valjean will do next, Inspector Javert might consider facts about Valjean's character, the weather, contemporary social conditions in France, Valjean's basic biological needs, and so on. On the other hand, it is not at all clear that humans always do, or can, take into account all the available and relevant information in forming certain judgments, and this might well be explained by our modular cognitive architecture.

Rationality

Once it became acceptable again in psychology to talk about thoughts, it was also acceptable to study how people string thoughts together, that is, how they reason. Newell and Simon, for instance, ran studies in which subjects worked through logic problems while thinking aloud, in order to investigate the reasoning they went through (Newell and Simon 1963). Unfortunately, just as the headline results of social psychology often showed people behaving badly – e.g. Milgram's (1963) work

on compliance or that of Latane and Darley (1970) on the bystander effect – so too, the study of reasoning soon seemed to show people reasoning badly.

These and other results, two of which we present shortly, widely prompted the question, To what extent are humans rational? Aristotle, of course, defined man as the rational animal, but throughout the centuries cynics and nay-sayers like Hume, Hobbes, and Freud have held a dissenting opinion. It would be good if actual psychological data could help answer this question, but first two complications must be noted.

The first is that the rules of logic and probability are not themselves normative principles. They describe the relations between propositions and probabilities but do not give advice on how to reason or what to believe (Goldman 1986: ch. 5; Harman 1986). There is, however, a natural way to derive normative principles of reason from logic and probability via what Stein (1996) calls the "standard picture of rationality." According to this picture, principles of rationality can be derived, fairly straightforwardly, from logic and probability theory. For instance, the conjunction rule is a basic rule of probability:

> CR: The probability of (A & B) cannot be greater than the probability of either A or B on its own.

From this rule we can plausibly derive a normative conjunction principle:

> CP: Don't assign a greater probability to the conjunction of A and B than to either event alone.

The standard picture also makes the deontological claim that rationality, or good reasoning, is reasoning in accord with these derived principles. If we adopt the standard picture, then the measure of how well (or poorly) people reason is the extent to which they follow principles of reason derived from logic and probability theory.

The second complication is that people's mistakes sometimes fail to show anything about their underlying abilities. Suppose that in the last sentence we had mistakenly written "they're" for "their." We might have done this because we were tired, or distracted, or drunk while writing this chapter, in which case the error wouldn't show that we didn't know the correct spelling. When mistakes are due to such extraneous factors, they are merely performance errors and show nothing about the underlying competence. In our example, our competence with English spelling would be impugned only if we routinely used "they're" for "their," even when fully awake, attentive, sober, etc.

Like our English spelling ability, our reasoning competence can be affected by such relatively extraneous factors as fatigue, level of attention, and intoxication. Our rationality can only be impugned if our mistakes are not mere performance errors but indicative of our competence. Hence the question "to what extent are people rational?" is concerned not with our reasoning performance but with our underlying reasoning competence. There are difficulties involved with spelling out the notion of human reasoning competence (discussed in Stein 1996: ch. 2), but here we will

assume that some such notion is to be had. We now present two of the most famous cases of apparent irrationality, the conjunction fallacy and failure on the selection task, to give a flavor of the psychological data.

Human irrationality?

The conjunction fallacy:

Consider the following character sketch (from Tversky and Kahneman 1982):

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

Now rank the following statements from least to most probable:

1 Linda is a teacher in elementary school.
2 Linda works in a bookstore and takes Yoga classes.
3 Linda is active in the feminist movement.
4 Linda is a psychiatric social worker.
5 Linda is a member of the League of Women Voters.
6 Linda is a bank teller.
7 Linda is an insurance salesperson.
8 Linda is a bank teller and is active in the feminist movement.

Most people rank (8) as being more probable than (6). Alas, this falls afoul of the conjunction principle described above. The probability that Linda is a bank teller *and* that she is a feminist – p(A & B) – simply cannot be greater than the probability that she is a bank teller – p(A) – and so we shouldn't judge it as greater.

The selection task:

Now consider the following task (from Wason 1966): There is a set of cards, each of which has a letter on one side and a number on the other. You are shown four cards which show "A," "K," "4," and "7" (see Figure 2). Which cards do you need to turn over to test the truth of the following rule?

If a card has a vowel on one side, then it has an even number on the other side.

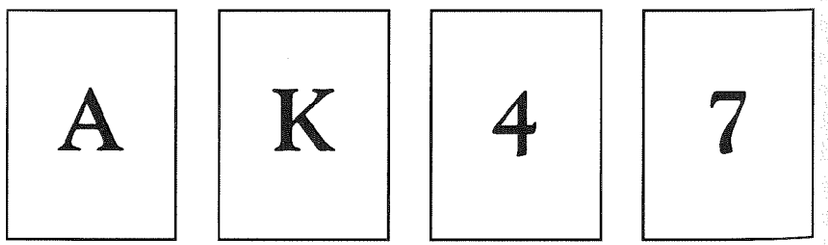

Figure 2

The correct answer is "A" and "7" If the "A" card has an odd number on the back, then the rule is false; if the "7" card has a vowel on the back, then the rule is false. Hence these two cards must be turned over. As for the "4" card, it is irrelevant to the rule; regardless of whether there is a vowel or consonant on the other side, the rule could be true or false, and similarly for the "K" card. Hence these cards do not need to be turned over. Most people, however, choose both "A" and "4" but don't choose "7."

There's plenty more bad news where that came from, with experiments showing overconfidence in the accuracy of one's judgments (Lichtenstein and Fischhoff 1977), hindsight bias (Fischhoff 1975), the gambler's fallacy (Tversky and Kahneman 1971), the perception of illusory correlations (Chapman and Chapman 1967), and neglect of base rate information (Kahneman and Tversky 1973), to mention just a few.[4] Kahneman and Tversky published a series of articles in the early 1970s, many of them collected in Kahneman et al. (1982), which presented some of these apparent failures of rationality and suggested some possible cognitive explanations. This inspired an ongoing research program to do, essentially, more of the same: the *heuristics and biases* program. (See Nisbett and Ross 1980 for a good review of early work, and Gilovich et al. 2002 for a collection of later papers.)

These errors in reasoning were originally studied to cast light on the cognitive mechanisms at work (just as visual illusions can tell us much about the visual system) but, as Kahneman and Tversky (1982) themselves noted, the method quickly became part of the message. Some psychologists, especially in the heuristics and biases tradition, interpreted these and similar results as having "bleak implications" for human rationality (Nisbett and Borgida 1975). Our susceptibility to such errors was taken to show that we are much less rational than we would like to be. The idea here is not, of course, that we are incapable of ever conforming to the principles of reason – we can apply the conjunction principle when it is made explicit and salient. Rather the "bleak implications view" is that much of the time we don't conform to the principles, and not just because of performance errors. Ergo, humans are substantially irrational.

Arguments against irrationality

Although the bleak implications view might seem vague enough to be immune to challenge – how much is much of the time? –there are actually several ways it might be challenged. First, one could try to explain away the data by invoking the influence of pragmatic and other confounding factors in subjects' understanding of the questions they are asked. Perhaps, for instance, subjects in the Linda experiment interpret (6) as meaning "Linda is a bank teller and not a feminist," in which case it could be perfectly rational to assign it a lower probability than (8). This approach can certainly explain away some of the experiments, but it won't work for them all. There are simply too many experiments, many of them designed to control for such confounds, and too many phenomena are robust under variations in experimental design.

Alternatively, one might claim that systematic human irrationality is not something that could ever be demonstrated empirically, because of the very nature of rationality.

One way this argument could go is by invoking ideas about the constraints on interpretation from Quine (1960), Davidson (1984), or Dennett (1987), who provides the clearest expression of the argument. According to Dennett, we attribute to people the intentional states that it would be rational for them to have. (For further discussion of Dennett see "Philosophy of mind," Chapter 12.) Hence, if we could discover that people were massively irrational, then we could no longer attribute beliefs and desires to them; but since we *do* attribute beliefs and desires, we must assume that they are rational. Unfortunately for this argument, in the considerable literature on "mind-reading" or "folk psychology" there is little evidence that considerations of rationality play a role in attribution of mental states; and in our everyday practice we have no trouble attributing inconsistent beliefs to someone, or less than fully rational reasoning.[5] So Dennett's argument will not rule out the bleak implications view.

In a much discussed article, Cohen (1981) offered an even stronger epistemological argument that human reasoning competence must be, not just mostly rational, but fully rational. Stripped to its essentials, Cohen's argument was that the principles of reason are derived from people's intuitions by a process of reflective equilibrium – which is, near enough, the same way we discover people's reasoning competence. Any deviations from the principles of reason must therefore be classed as performance errors. Where this argument founders is its tendency to relativism about rationality. Consider Poor Jim, who bets his entire life savings that the next toss of a fair coin will be heads, because the last ten times the coin landed tails. No matter how much reflective equilibrium we try to induce in Jim, no matter how much statistical training he gets, he persists in committing this gambler's fallacy. Nonetheless, according to Cohen, Poor Jim is not being irrational – at least, not for him. His own principles of reason, derived via reflective equilibrium on things like his intuitions about the next coin toss, countenance the gambler's fallacy. Since Poor Jim is clearly not rational, what this example shows is that Cohen is simply wrong about how the principles of reason are derived.[6]

The third main response to the bleak implications view has come from evolutionary psychology. Its basic thrust is to challenge the ecological validity of most experiments on human reasoning, and to claim they have not been properly designed to elicit responses that reflect the subject's reasoning competence. If instead we considered problems similar to those which our ancestors would have faced in the environment of evolutionary adaptation (EEA) and which human cognition evolved to handle, then we would find much more evidence of human rationality than is allowed by the bleak implications view.

One specific instance of this response concerns the Wason selection task. This task has been extensively studied with a variety of rules, some of which do elicit the correct response. For instance, when the subjects are asked to test the rule "If someone is drinking alcohol, they must be over 21" (see Figure 3), they readily choose the correct responses, "beer" and "18" (Griggs and Cox 1982). To account for the differences between cases that elicit the correct response and cases that don't, Cosmides and Tooby (1992) proposed the cheater-detection hypothesis. According to this

hypothesis, one problem our ancestors would have faced was detecting people who would try to cheat them in social exchanges, and so humans probably have a module devoted to solving this problem. Selection tasks which have the form of a "social contract," as in Figure 3, will trigger this module, which produces the right response. Selection tasks which don't have this form, as in Figure 2, won't trigger the module.[7]

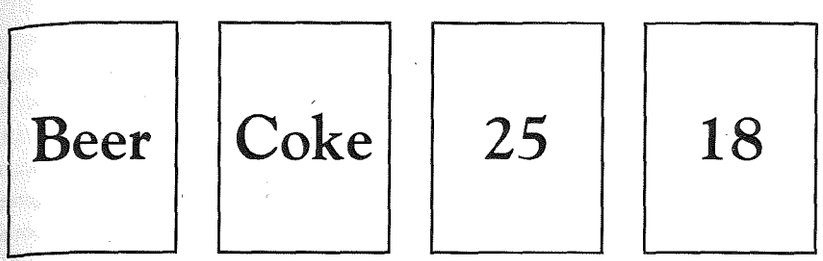

Figure 3

Another application of the challenge from evolutionary psychology concerns the format in which information is presented. Cosmides and Tooby (1996) and Gigerenzer (1994) have argued that, in the EEA, statistical information would have been presented as frequencies rather than probabilities of particular events: "two out of every ten times you go hunting you meet a tiger" rather than "the probability of meeting a tiger on today's hunt is 0.2." It is thus no surprise that people reason poorly about single-event probabilities, but they should be able to reason quite well about frequencies. And, indeed, several studies have shown that subjects can choose normatively correct responses when the information is given as frequencies. For instance, in one study subjects were told that 200 women met the "Linda" description, and were then asked to estimate how many of those women were feminists, bank tellers, and feminist bank tellers (Gigerenzer 1994). Only 13 percent of subjects committed the conjunction fallacy (by estimating that there were more feminist bank tellers than bank tellers), far fewer than the 85 percent who committed the fallacy in the analogous probabilistic version of Tversky and Kahneman (1982).

Evolutionary psychology and the heuristics and biases program have been depicted as necessarily opposed, but this is mistaken (Samuels et al. 2002). There is indeed a difference in emphasis, with evolutionary psychology focusing on reasoning successes and the heuristics and biases program focusing on failures, but this is only a difference in emphasis. Evolutionary psychology is not committed to the Panglossian view that humans are perfect reasoners, especially not in a modern environment so different from the EEA; nor is the heuristics and biases program committed to the bleak implications view.

At a more fundamental level, the challenges from evolutionary psychology haven't yet managed to disprove the bleak implications view. Granted, on the right sorts

of EEA-friendly problems, people answer correctly, but this leaves untouched our failures on all the other problems. Hence it is still an open possibility that, in many of the cases where information is not presented in the appropriate format or where the problem does not resemble a problem from the EEA, humans are indeed as irrational as they have been in the classic studies.

An alternative view of rationality

There is, however, a second, more interesting component to Gigerenzer's program which leads to a deeper issue in evaluating human rationality. The bleak implications view, and the entire enterprise of assessing rationality by seeing whether people follow the principles of reason, relies on the standard picture and its deontological account of rationality. According to the alternative consequentialist account that Gigerenzer advocates, good reasoning is reasoning which (reliably) produces the best outcome. If we adopt this consequentialist approach, then the truth or falsity of the bleak implications view, and thus the extent of human rationality, become contingent on the sort of world we live in and not on our following the principles of reason. Even if people violate principles of reason, and fall short of being perfect Bayesians, our reasoning mechanisms might still manage to produce the best outcomes overall.

For instance, a plausible principle of reason might be "consider all the available evidence when making a decision" but Gerd Gigerenzer (1947–) and colleagues have shown that *not* considering all the evidence can be a better strategy in the right sort of environment (Gigerenzer et al. 1999: Parts II and III). The work of Robyn Dawes (e.g. Dawes 1979) has similarly shown that prediction rules which are prima facie normatively deficient (e.g. rules which weight all predictive factors equally, regardless of their predictive validity) can outperform more complicated rules.

There are good reasons for favoring a consequentialist view of rationality which we won't go into here.[8] To assess the extent of human rationality, on a consequentialist account, we will have to learn much more not only about human psychology, but also about the task environments we actually face (see Martignon 2001). In any event, whether we accept the standard deontological picture of rationality or opt for an alternative consequentialist picture, we are still a long way from being able to say just how rational humans are.

Naturalization of intentionality

We have now considered four central questions in the philosophy of psychology: Do we need to invoke inner representational states to explain behavior? How much innate information and structure does the mind have? How modular is the mind? How rational are we? While all four involve some conceptual and philosophical heavy lifting, they are largely empirical questions. To answer them, at the very least we need to know a lot of facts about human psychology.

Our final topic, the naturalization of intentionality, is quite different. The enterprise here looks considerably more like traditional philosophy: X proposes a theory,

Y comes up with a counterexample, X adds a complication or two to keep her theory going; repeat as necessary. To naturalize intentionality, it seems, we don't need to know a lot of facts about human psychology. The question is not *whether* the human mind has intentionality,[9] or *how much* intentionality it has, but *how* it has it.

Nonetheless, even this philosophical enterprise has been strongly influenced by developments in psychology, at least indirectly. As we have seen, during the 1960s, the ruling orthodoxy of behaviorism was overthrown by cognitivism, and one of the distinctive features of cognitivism was its positing of representational states: mental states which are *about* things. As cognitivism became more widespread, the question thus arose: how to develop a scientifically respectable account of representation? Or, in other words, how to naturalize intentionality?

Intentionality and naturalization

But first the basics: What is intentionality, and what would it take to naturalize it? In our everyday folk psychological reasoning, we attribute thoughts to one another, and these thoughts are *about* things: suppose you want a beer, believe there's a beer in the fridge, wonder whether Guinness is better than Foster's, and so on. Your thoughts are *about* beer, the fridge, Guinness and Foster's, etc. Similarly, cognitivist psychology also posits mental states which are about things. Intentionality is this property of *aboutness*: the property that mental states (as well as things like words and maps) can have of being about one thing or another, of having a particular *content*. The notion of intentionality was reintroduced to philosophy and psychology in 1874 by Franz Brentano, who adopted it from medieval Scholastic philosophy. Brentano claimed that what distinguished mental states from (merely) physical states was that mental states were intentional; hence what came to be known as *Brentano's thesis* that the intentional is the mark of the mental.

Taking up these ideas, Chisholm (1957) argued that mental states were ineliminably intentional, that they could not be described or explained except in intentional terms. As Quine (1960) put it, Brentano's thesis shows either "the indispensability of intentional idioms and the importance of an autonomous science of intention" or "the baselessness of intentional idioms and the emptiness of a science of intention." Quine, the arch-behaviorist, was happy to take the second option, but almost everyone else has preferred the first.

Hence the attempts from the 1970s onwards to naturalize intentionality, that is, to come up with a theory of intentionality which is scientifically respectable. If intentionality is a real feature of mental states as posited by cognitivism, then we want to know how it fits into the rest of our scientific (and physicalist) picture of the world. Intentionality doesn't look to be part of the basic stuff in the world like atoms and quarks, and so we want to know how you get intentionality out of the basic stuff (Fodor 1987). Attempts to naturalize intentionality are attempts to show how you can do this, to explain intentionality in a naturalistic vocabulary which does not itself invoke intentionality (see also "Philosophy of mind," Chapter 12).

In addition to showing how intentionality arises from non-intentional stuff, there are two thorny problems an adequate naturalistic theory must solve: first, the *problem of misrepresentation*. You can believe your chocolate is in the cupboard, even though someone has secretly moved it elsewhere. You can mistake fool's gold for gold. You can believe that unicorns exist. In each of these cases, your mental states do not match up with the way the world really is. A naturalistic theory of intentionality must explain how this mismatch is possible. Second, the problem of *grain*. You can think about Hesperus, or you can think about Phosphorus. These are different thoughts, and have different content, but of course Hesperus just is Phosphorus. A naturalistic theory of intentionality must be able to draw such fine-grained distinctions between related but distinct contents.

Finally, before discussing the theories themselves, a word about narrow versus wide content. Since Putnam (1975), philosophers have debated whether the contents of mental states are to be individuated widely (by including factors which "ain't in the head") or narrowly (by not including such factors).[10] Most of the naturalistic accounts of intentionality are externalist, favoring wide individuation conditions. But even internalists, who favor narrow content, have to explain how mental states can be about things. Hence they too need to provide an account of intentionality (see Botterill and Carruthers 1999).

Three accounts of intentionality

Three main approaches have been taken to naturalizing intentionality: informational theories, teleosemantic theories, and conceptual role semantics. As they have developed in response to various problems and objections, these theories have accrued epicycles like barnacles on a shipwreck. To discuss the finer points of these theories would be well beyond the scope of this chapter (and beside the point, as we argue below). It is enough for our purposes to present them in somewhat simplified form.[11]

The informational approach is most closely associated with Fred Dretske (1981; 1988). This approach analyzes the content of a mental state in terms of the information it carries; your beer-thoughts are about beer because they carry information about beer. Information here is understood as causal co-variance. If there is a reliable law (typically, a causal law) that "whenever X obtains, so does Y," then Xs carry information about Ys. Spots carry information about measles, smoke carries information about fire, and beer-thoughts carry information about beer.

A pure informational approach has trouble both with misrepresentation and with grain. Suppose that Brenda reliably mistakes toads for frogs. Then her frog-thoughts would carry information about toads as much as about frogs, so how can we say that her frog-thoughts are mistaken when they occur in the presence of toads? Aren't they instead perfectly accurate frog-or-toad-thoughts? The informational approach cannot deal easily with misrepresentation. Similarly, Hesperus-thoughts may carry information about Hesperus, but they also carry information about Phosphorus. So why aren't our Hesperus-thoughts equally well Phosphorus-thoughts; that is to say, how can an informational approach solve the problem of grain? Another problem

concerns causal chains. Smoke-thoughts carry information about smoke, but they also carry information about fire (since, generally, where there's smoke, there's fire). So why aren't our smoke-thoughts equally well fire-thoughts?

Informational theorists have developed two main ways of dealing with these problems. The first is Fodor's, which is to introduce a requirement of *asymmetric causal dependence* (Fodor 1990). Suppose that there were two laws in Brenda's psychology:

(1) frogs → frog-thoughts

(2) toads → frog-thoughts

Brenda's frog-thoughts are about frogs and not toads if (2) asymmetrically depends on (1). This means that if (1) did not hold, then (2) would not hold either but even if (2) did not hold, (1) still would. If frogs were no longer reliably linked to frog-thoughts, then toads wouldn't be linked to frog-thoughts either, but the converse is not true. The notion of asymmetric dependence, however, has serious problems of its own, which have been well rehearsed elsewhere (Adams and Aizawa 1994; Adams 2003). The main problem is spelling it out in a way which is properly naturalistic and not ad hoc, and this looks to be difficult, if not impossible.

The second way of dealing with these problems for the informational approach, and the way Dretske has taken, is to incorporate teleological elements into the theory. But the teleosemantic approach has also been pursued in its own right, primarily by Ruth Millikan (1984; 1989) and David Papineau (1987; 1993). The basic idea of teleosemantics is that the content of a particular mental state is given by its biological function: frog-thoughts are about frogs just in case their biological function is to carry information about frogs. A thing's biological function is then defined, following Wright's (1973) analysis, as the job that natural selection designed that thing to do. The function of hearts is to pump blood and, similarly, the function of frog-thoughts is to carry information about frogs.

Such a teleological approach deals well enough with the problem of misrepresentation. We could generally be mistaken about toads – indeed, our ancestors could have been as well – and our frog-thoughts would still be about frogs. All that matters is that frog-thoughts were selected to carry information about frogs, not toads. Where the teleosemantic approach runs into trouble is dealing with the problem of grain. For instance, suppose that in our evolutionary history we had only ever come into contact with one particular species of frog. Then what (if anything) makes it the case that our frog-thoughts are about frogs in general rather than that one particular species? Teleosemantic approaches must assume that evolutionary theory will pick out one content rather than the other as the appropriate biological function, but this assumption may not be well founded. Even if an appeal to evolution can solve this problem, it is not clear how it could also solve the Hesperus/Phosphorus problem.

The final approach we shall consider has been variously called conceptual, functional, inferential, or causal role semantics (Block 1986; Field 2001). The idea here is that thoughts are individuated in terms of their inferential or functional role

in a person's overall cognitive economy, and so such an approach has typically been restricted to theories of narrow content (e.g. Field 1978; Loar 1981). But to explain intentionality – the relation between thoughts and the world – and to specify contents, functional role semantics must construe functional role broadly, so as to include relations between mental states and the world. The reason beer-thoughts are about beer as opposed to scotch (or, for that matter, the reason they are about anything at all) surely has something to do with how the world is and how we are hooked up to it. A broad functional role semantics would explain this by including various thought–world interactions as part of the thought's functional role; for instance, part of the functional role of beer-thoughts is that they are produced in the presence of beer, and that they lead to consumption of beer when one so desires.

It is difficult to assess the merits of functional role semantics as a naturalistic account of intentionality. While various philosophers have gestured at the approach as a way to naturalize intentionality, no one has actually offered a detailed account that would assign specific content to mental states. For instance, what is the functional role of the belief that the beer is in the fridge, as opposed to the belief that the scotch is in the fridge, or that the beer is in the cupboard? Moreover, all beer-thoughts are about beer, but what is common to the functional role of beer-thoughts? It is one thing to say that functional roles can answer these questions, but it is another thing to show how and, as yet, no one has come even close to doing this.

Doubts about naturalization

So much for the main approaches to naturalizing intentionality. Each of these approaches has its own problems, which we have only sketched here; they are well summarized in Adams (2003) and Loewer (1997). There are also more general doubts about the overall project of naturalizing intentionality. First, a suggestion made by Horgan (1994) and Loewer (1997). Perhaps intentionality is indeed naturalistically respectable – that is, intentionality is built out of basic physical features of the world – but there is no way for us to understand how this happens. The problem of explaining intentionality may be simply intractable, or beyond our limited cognitive capacities.[12]

This sort of cognitive closure with respect to intentionality is certainly a possibility, but its significance ought not be overstated. Maybe we *can't* provide a fully naturalistic account of intentionality but that's no argument against trying. You've got to be in it to win it. To be fair to Horgan and Loewer, they were not arguing that we should give up on naturalizing intentionality. Rather, the possibility of cognitive closure was offered as a potential explanation of why we haven't succeeded so far, and why we might never succeed. Even so, there is no reason why, having recognized this possibility, we shouldn't keep trying.

There is another general doubt about the project, however, which we think is more serious and which lies in the fuzziness of the demand for naturalization (Stich 1992; Tye 1992; Stich and Laurence 1994; Botterill and Carruthers 1999). What, exactly, would *count* as having naturalized intentionality? As we noted at the start of this section, the debates over intentionality resemble good old-fashioned conceptual

analysis (well, old-fashioned conceptual analysis, anyway). X offers a theory of intentionality, Y brings up a counterexample (often outlandish), and X goes back to the drawing board. So the call for naturalization looks like a call to provide necessary and sufficient conditions for something's being intentional, and for something's having the particular content that it does. That is, it looks like a call for a definition of the term or concept "intentionality."

But if that's what naturalization amounts to, then why suppose that we can naturalize intentionality – or anything else? Fodor himself has argued in a different context that necessary and sufficient conditions won't be forthcoming for even the most innocuous concept like "paint" (Fodor 1981), so it would be surprising if we *could* produce a definition for "intentionality." Indeed, if certain views about conceptual structure are right, it's not just that definitions for concepts are hard to come by, it's that there *aren't* any. If concepts have exemplar or prototype structure, for instance, then they don't have definitions and so cannot be analyzed in terms of necessary and sufficient conditions.[13]

What unites these two general doubts (about cognitive closure and about the very project of naturalization) is their skepticism about the urgent need to naturalize. Fodor predicts, quite literally, apocalypse if the intentional can't be naturalized: it would be "the end of the world" (1990: 156) and "the greatest intellectual catastrophe in the history of our species" (1987: xii). This is because he thinks that intentional irrealism would follow; other putative consequences include the causal/explanatory impotence of intentionality and the impossibility of an intentional psychology.

The prospect of "failure" to naturalize, then, has been seen by some as tantamount to opening the seventh seal. Against this alarmism, we caution: don't believe the hype. Regardless of the actual structure of our concepts, there are plenty of other features of the world for which we lack definitions despite a lot of sophisticated effort aimed at producing them. Necessary and sufficient conditions have not been found for something's being a phoneme, or a gene, or a species, or just about anything except being a bachelor or a triangle. You don't find biologists gnashing their teeth because they can't give necessary and sufficient conditions for "gene" or "species," and this "failure" seems hardly to impugn the realism, efficacy, or scientific respectability of genes or species.

More generally, however the demand to naturalize intentionality is meant, the apocalypse will not ensue if we fail (see Stich and Laurence 1994). This is not to say that there are no interesting questions about intentionality and mental content. It is not yet settled, for instance, what sort of mental states a mature psychology will be committed to, what sorts of intentional properties, if any, those states will have, and how this is related to the intentional properties invoked by everyday folk psychology (Stich 1992). But answering these questions is a long way from "naturalizing intentionality."

Conclusion

Of necessity, this survey of important developments in twentieth-century philosophy of psychology has been selective. There are many issues we have not discussed, or only touched on, including the following.[14]

The nature of concepts

What is the structure of mental representation, or concepts? Philosophers [illegible] have typically assumed that our concepts are defined by a set of necessary and [illegible] cient conditions. Since the pioneering work of Eleanor Rosch (1938–) in the [illegible] (Rosch and Mervis 1975), psychologists have developed at least two other, competing accounts of our conceptual structure. Philosophers have participated actively in the (still unresolved) empirical debate which has ensued, as well as teasing out the philosophical implications of these alternatives. Margolis and Laurence 1999 offers an excellent overview of this work.

Connectionism

According to the "classical" or symbol-manipulation view in cognitive psychology, thinking is computation, i.e. the manipulation of symbolic representations. By contrast, according to connectionism there is no separation between representations and processes operating over those representations. Rather, cognition is modeled as a network of interconnecting nodes, loosely inspired by neural networks, and information is distributed over the entire network. Since the re-emergence of connectionism in 1986 (Rumelhart et al. 1986), some have argued that connectionist models were more plausible descriptions of human psychology; philosophers have debated whether connectionism and classical models are genuine alternatives; and the potential consequences if the human mind were indeed connectionist (see Bechtel and Abrahamsen 2002).

Extended cognition

Psychologists and philosophers have generally conceived of thinking as something that goes on purely "in the head"; both connectionist and classical models of cognition make this assumption. The idea of extended cognition challenges this by recasting cognition as reaching out "into the world": for example, the mobile phone you store important phone numbers in is not just a tool for your thinking, it is a part of your thinking. Andy Clark has advocated this idea in his 2003 book and elsewhere.

The emotions

What are the emotions, and how do they relate to the rest of our thinking? In the late twentieth century, there was extensive work in both philosophy and psychology trying to address this question. Unfortunately, most of the philosophy simply ignored the psychology, although this situation has been partly remedied in later years (e.g. Griffiths 1997; Prinz 2004).

Theory of mind and simulation in folk psychological reasoning

How do we detect and think about other people's mental states? For instance, how do we predict your behavior given what we know about your beliefs and desires? According to the theory theory, we use a mentally represented theory of mind which includes general psychological laws. By contrast, according to simulation theory, proposed independently by philosophers Robert Gordon and Jane Heal (Gordon 1986;

Heal 1986), we must simulate being you; that is to say, we imagine having your mental states, and see what we would do. Much research on the psychological mechanisms for mentalizing was done during the last three decades of the twentieth century, and philosophers actively contributed to this research and the broader debates (Nichols and Stich 2003; Goldman 2006).

Consciousness

Philosophers at least since Dennett (1969) have tried to develop accounts of consciousness informed by psychology; this trend increased in the 1990s with the revival in psychology itself of research into consciousness (see the papers in Baars et al. 2003). Much of the philosophy of consciousness, however, is still only tangentially connected, if that, with the psychology of consciousness.

Eliminativism and the role of the propositional attitudes in psychology

In our everyday folk psychology we naturally ascribe mental states like beliefs and desires to one another, and to ourselves. Will such propositional attitudes appear in our finished scientific psychology and, if not, does that not mean we should reject them as empirically inadequate concepts (like "witch" or "phlogiston")? It has seemed to some philosophers that cognitive psychology or connectionist models leave no room for propositional attitudes, but this has been hotly contested (see Horgan and Tienson 1991).

The rise of evolutionary psychology

One of the most striking facts about psychology since the 1990s has been the increasing influence (not least in the arena of public debate) of evolutionary psychology. Broadly speaking, evolutionary psychology is any psychology influenced by evolutionary theory; more commonly, however, it refers to a specific research paradigm closely associated with Cosmides and Tooby (Barkow et al. 1992). This paradigm has found both philosophical advocates (e.g. Samuels et al. 1999) and opponents (e.g. Buller 2005).

The growing importance of neuroscience

During the closing decades of the twentieth century, technological improvements have made possible increasingly detailed investigation into the actual neurological circuitry of human and animal cognition. Inevitably, this work has received increasing philosophical attention (Bechtel et al. 2001), which is only likely to increase further as more and more is discovered in neuroscience.

This represents just a sample of philosophically relevant research in psychology at the start of the twenty-first century. More so than in the twentieth century even, psychology today moves fast, in hundreds of different directions. Scarcely a week goes by without some new research plowing up fertile ground for philosophical work. What we hope to have shown in this overview is that philosophers of psychology have proved themselves up to the task. Philosophy of psychology is not the handmaiden of psychology. In the last century at least, philosophy and psychology worked hand-in-hand to uncover truths about the human mind. May they stay partners in the next.

Notes

1 We stress that this is only a rough characterization; it is both too broad and too narrow. Ultimately, nothing really hangs on whether a specific issue is considered part of philosophy of psychology or philosophy of mind.

2 Richard Samuels offers an account of innateness that is related to the present account, but is quite a bit more sophisticated. See Samuels 2002 for a lucid discussion.

3 In fact, evolutionary psychologists are generally committed to a specific type of computational module, namely one which was produced by natural selection and is universal across all (normal) humans. Since these further features aren't relevant here, we will pass over them.

4 *Overconfidence*: people's confidence in their own judgments can far outstrip their actual accuracy. *Hindsight bias*: once they know an event has occurred, people typically overestimate how likely they would have thought that event in advance. *Gambler's fallacy*: e.g. after seeing a long run of heads in a series of coin tosses, many people think that tails must be "due" and therefore more likely to occur. *Illusory correlations*: people will often see correlations (in a set of data) that aren't there. *Base-rate neglect*: given information about the base-rate frequency of a particular outcome in a population (e.g. frequency of colon cancer), and diagnostic evidence about a specific case (e.g. Smith's bioscopy), most people ignore the base-rate information when predicting the likelihood that the outcome has occurred in this case (i.e. that Smith has colon cancer).

5 The psychologists Csibra and Gergely have interpreted their results in a Dennettian framework (e.g. Gergely et al. 1995; Csibra et al. 1999) but this seems an over-interpretation. For more on the relation of rationality theories to human rationality and to folk psychology, see Stich 1985 and Nichols and Stich 2003: sect. 3.4.4.

6 Admittedly, the argument here is hardly conclusive. For more extended criticism of Cohen, see Stich 1985; 1990; and Stein 1996.

7 There are other hypotheses to explain the pattern of results, from outside evolutionary psychology (e.g. Cheng and Holyoak 1989; Oaksford and Chater 1994; Cummins 1996). Which of these hypotheses is correct is still a hotly debated issue.

8 They are persuasively put by Bishop (2000) and Bishop and Trout (2004).

9 Certainly, irrealism about intentionality is a possible view, but few contemporary philosophers have adopted it with any gusto (but see the discussion of Quine, below). Dennett's instrumentalism might seem close (e.g. Dennett 1987), but he denies that his position is irrealist (Dennett 1991).

10 Pessin and Goldberg 1996 contains many papers on this externalism/internalism debate.

11 For more detail, see the selections in Stich and Warfield 1994.

12 This suggestion is analogous with what McGinn (1993) has claimed about consciousness.

13 See Margolis and Laurence 1999 and Murphy 2002 for discussion of psychological theories of concepts and categorization.

14 This list is *far* from exhaustive.

References

Adams, F. (2003) "Thoughts and their contents: the structure of cognitive representations." In S. Stich and T. A. Warfield (eds.) *The Blackwell Guide to Philosophy of Mind*, Oxford: Blackwell, pp. 143–71.

Adams, F. and K. Aizawa (1994) "Fodorian semantics." In S. Stich and T. A. Warfield (eds.) *Mental Representation*, Oxford: Blackwell, pp. 223–42.

Atran, S. (1998) "Folk biology and the anthropology of science: cognitive universals and cultural particulars." *Behavioral and Brain Sciences* 21: 547–609.

Baars, B. J., W. P. Banks, and J. B. Newman (2003) *Essential Sources in the Scientific Study of Consciousness*. Cambridge, MA: MIT Press.

Barkow, J., L. Cosmides, and J. Tooby (eds.) (1992) *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford: Oxford University Press.

Baron-Cohen, S. (1995) *Mindblindness: An Essay on Autism and Theory of Mind*. Cambridge, MA: MIT Press.

Bateson, P. (1991) "Are there principles of behavioural development?" In P. Bateson (ed.) *The Development and Integration of Behaviour*, Cambridge: Cambridge University Press, pp. 19–39.

Bechtel, W. and A. Abrahamsen (2002) *Connectionism and the Mind: Parallel Processing, Dynamics, and Evolution in Networks*, 2nd edn. Oxford: Blackwell.

Bechtel, W., R. Stufflebeam, J. Mundale, and P. Mandik (2001) *Philosophy and the Neurosciences: A Reader*. Oxford: Blackwell.

Bishop, M. (2000) "In praise of epistemic irresponsibility: How lazy and ignorant can you be?" *Synthèse* 122: 179–208.

Bishop, M. and J. D. Trout (2004) *Epistemology and the Psychology of Human Judgment*. Oxford: Oxford University Press.

Block, N. (1986) "Advertisement for a semantics for psychology." *Midwest Studies in Philosophy* 10: 615–78.

Botterill, G. and P. Carruthers (1999) *The Philosophy of Psychology*. Cambridge: Cambridge University Press.

Brentano, F. (1973) [1874] *Philosophy from an Empirical Standpoint*. London: Routledge & Kegan Paul.

Buller, D. J. (2005) *Adapting Minds: Evolutionary Psychology and the Persistent Quest for Human Nature*. Cambridge, MA: MIT Press.

Carey, S. and E. Spelke (1994) "Domain-specific knowledge and conceptual change." In L. Hirschfeld and S. Gelman (eds.) *Mapping the Mind*, Cambridge: Cambridge University Press, pp. 169–200.

Carruthers, P. (2003) "On Fodor's problem." *Mind and Language* 18: 502–23.

Chapman, L. J. and J. P. Chapman (1967) "Genesis of popular but erroneous psychodiagnostic observations." *Journal of Abnormal Psychology* 73: 193–204.

Cheng, P. and K. Holyoak (1989) "On the natural selection of reasoning theories." *Cognition* 33: 285–313.

Chisholm, R. (1957) *Perceiving: A Philosophy Study*. Ithaca, NY: Cornell University Press.

Chomsky, N. (1959) "A review of Skinner's *Verbal Behavior*." Repr. in N. Block (ed.) *Readings in Philosophy of Psychology*, Cambridge, MA: Harvard University Press, 1980, pp. 48–63.

—— (1965) *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

—— (1988) *Language and Problems of Knowledge*. Cambridge, MA: MIT Press.

—— (1980) *Rules and Representation*. New York: Columbia University Press.

Clark, A. (2003) *Natural-Born Cyborgs: Minds, Technologies, and the Future of Human Intelligence*. Oxford: Oxford University Press.

Cohen, L. (1981) "Can human irrationality be experimentally demonstrated?" *Behavioral and Brain Sciences* 4: 317–70.

Cosmides, L. and J. Tooby (1992) "Cognitive adaptations for social exchange." In J. Barkow, L. Cosmides, and J. Tooby (eds.) *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, Oxford: Oxford University Press, pp. 163–228.

—— and —— (1994) "Origins of domain specificity: the evolution of functional organization." In L. Hirschfeld and S. Gelman (eds.) *Mapping the Mind*, Cambridge: Cambridge University Press, pp. 85–116.

—— and —— (1996) "Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty." *Cognition* 58: 1–73.

Cowie, F. (1999) *What's Within? Nativism Reconsidered*. New York: Oxford University Press.

Csibra, G., G. Gergely, S. Biro, O. Koos, and M. Brockbank (1999) "Goal attribution without agency cues: the perception of 'pure reason' in infancy." *Cognition* 72: 237–67.

Cummins, D. (1996) "Evidence for the innateness of deontic reasoning." *Mind and Language* 11: 160–90.

Davidson, D. (1963) "Actions, reasons and causes." *Journal of Philosophy* 60/23: 685–700.

—— (1984) *Inquiries into Truth and Interpretation*. Oxford: Oxford University Press.

Dawes, R. M. (1979) "The robust beauty of improper linear models in decision making." *American Psychologist* 34: 571–82.

Dennett, D. (1969) *Content and Consciousness*. London: Routledge & Kegan Paul.

—— (1987) *The Intentional Stance*. Cambridge, MA: MIT Press.

—— (1991) "Real patterns." *Journal of Philosophy* 88: 27–51.

Descartes, R. (1985) *The Philosophical Writings of Descartes*, vol. 1, trans. J. Cottingham, R. Stoothoff, and D. Murdoch. Cambridge: Cambridge University Press.
Dretske, F. (1981) *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.
—— (1988) *Explaining Behavior*. Cambridge, MA: MIT Press.
Elman, J., E. Bates, M. H. Johnson, A. Karmiloff-Smith, D. Parisi, and K. Plunkett (1996) *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge: Cambridge University Press.
Field, H. (1978) "Mental representation." *Erkenntnis* 13: 9–61.
—— (2001) *Truth and the Absence of Fact*. Oxford: Oxford University Press.
Fischhoff, B. (1975) "Hindsight is not equal to foresight: the effect of outcome knowledge on judgment under uncertainty." *Journal of Experimental Psychology: Human Perception and Performance* 1: 288–99.
Fodor, J. (1968) *Psychological Explanation*. New York: Random House.
—— (1981) *Representations*. Cambridge, MA: MIT Press.
—— (1983) *The Modularity of Mind*. Cambridge, MA: MIT Press.
—— (1987) *Psychosemantics*. Cambridge, MA: MIT Press.
—— (1990) *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press.
—— (2000) *The Mind Doesn't Work That Way*. Cambridge, MA: MIT Press.
Gardner, H. (1985) *The Mind's New Science: A History of the Cognitive Revolution*. New York: Basic Books.
Gelman, R. and K. Brenneman (1994) "First principles can support both universal and culture-specific learning about number and music." In L. Hirschfeld and S. Gelman (eds.) *Mapping the Mind*, Cambridge: Cambridge University Press, pp. 369–90.
Gergely, G., Z. Nadasdy, G. Csibra, and S. Biro (1995) "Taking the intentional stance at 12 months of age." *Cognition* 56: 165–93.
Gigerenzer, G. (1994) "Why the distinction between single-event probabilities and frequencies is important for psychology (and vice versa)." In G. Wright and P. Ayton (eds.) *Subjective Probability*, New York: John Wiley, pp. 129–61.
Gigerenzer, G., P. Todd, and the ABC Research Group (1999) *Simple Heuristics that Make Us Smart*. New York: Oxford University Press.
Gilovich, T., D. Griffin, and D. Kahneman (2002) *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge: Cambridge University Press.
Goldman, A. (1986) *Epistemology and Cognition*. Cambridge, MA: Harvard University Press.
—— A. (2006) *Simulating Minds: The Philosophy, Psychology and Neuroscience of Mindreading*. Oxford: Oxford University Press.
Gordon, R. (1986) "Folk psychology as simulation." *Mind and Language* 1: 158–71.
Griffiths, P. (1997) *What Emotions Really Are*. Chicago: University of Chicago Press.
—— (2002) "What is innateness?" *The Monist* 85/1: 70–85.
Griggs, R. and J. Cox (1982) "The elusive thematic-materials effect in Wason's selection task." *British Journal of Psychology* 73: 407–20.
Harman, G. (1986) *Change of View*. Cambridge, MA: MIT Press.
Heal, J. (1986) "Replication and functionalism." In J. Butterfield (ed.) *Language, Mind, and Logic*, Cambridge: Cambridge University Press, pp. 135–50.
Hirschfeld, L. and S. Gelman (eds.) (1994) *Mapping the Mind: Domain Specificity in Cognition and Culture*. Cambridge: Cambridge University Press.
Horgan, T. (1994) "Computation and mental representation." In S. Stich and T. A. Warfield (eds.) *Mental Representation*, Oxford: Blackwell, pp. 302–11.
Horgan, T. and J. Tienson (eds.) (1991) *Connectionism and the Philosophy of Mind*. Dordrecht: Kluwer.
Kahneman, D. and A. Tversky (1973) "On the psychology of prediction." *Psychological Review* 80: 237–51.
—— and —— (1982) "On the study of statistical intuitions." In D. Kahneman, P. Slovic, and A. Tversky (eds.) *Judgment under Uncertainty: Heuristics and Biases*, Cambridge: Cambridge University Press, pp. 493–508.
Kahneman, D., P. Slovic, and A. Tversky (eds.) (1982) *Judgment under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.
Kandel, E., J. H. Schwartz, and T. M. Jessell (2000) *Principles of Neural Science*. New York: McGraw-Hill.
Karmiloff-Smith, A. (1992) *Beyond Modularity*. Cambridge, MA: MIT Press.

Kitcher, P. (1996) *The Lives to Come*. New York: Simon & Schuster.
Laland, K. N. and G. R. Brown (2002) *Sense and Nonsense: Evolutionary Perspectives on Human Behavior*. New York: Oxford University Press.
Latane, B. and J. M. Darley (1970) *The Unresponsive Bystander: Why Doesn't He Help?* New York: Appleton-Century-Crofts.
Leslie, A. (1994) "ToMM, ToBY, and agency: core architecture and domain specificity." In L. Hirchfeld and S. Gelman (eds.) *Mapping the Mind*, Cambridge: Cambridge University Press, pp. 119–48.
Lichtenstein, S. and B. Fischhoff (1977) "Do those who know more also know more about how much they know?" *Organizational Behavior and Human Performance* 20: 159–83.
Loar, B. (1981) *Mind and Meaning*. Cambridge: Cambridge University Press.
Loewer, B. (1997) "A guide to naturalizing semantics." In B. Hale and C. Wright (eds.) *A Companion to the Philosophy of Language*, Oxford: Blackwell, pp. 108–26.
Margolis, E. and S. Laurence (eds.) (1999) *Concepts: Core Readings*. Cambridge, MA: MIT Press.
Martignon, L. (2001) "Comparing fast and frugal heuristics and optimal models." In G. Gigerenzer and R. Selten (eds.) *Bounded Rationality: The Adaptive Toolbox*, Cambridge, MA: MIT Press, pp. 147–71.
McGinn, C. (1993) *Problems in Philosophy: The Limits of Inquiry*. Oxford: Blackwell.
Milgram, S. (1963) "Behavioral study of obedience." *Journal of Abnormal and Social Psychology* 67: 371–78.
Millikan, R. (1984) *Language, Thought, and Other Biological Categories*. Cambridge, MA: MIT Press.
—— (1989) "Biosemantics." *Journal of Philosophy* 86: 281–97.
Murphy, G. (2002) *The Big Book of Concepts*. Cambridge, MA: MIT Press.
Newell, A. and H. Simon (1963) "GPS, a program that simulates human thought." In E. Feigenbaum and J. Feldman (eds.) *Computers and Thought*, New York: McGraw-Hill, pp. 279–93.
Nichols, S. and S. Stich (2003) *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding of Other Minds*. Oxford: Oxford University Press.
Nisbett, R. E. and E. Borgida (1975) "Attribution and the psychology of prediction." *Journal of Personality and Social Psychology* 32: 932–43.
Nisbett, R. E. and L. Ross (1980) *Human Inference: Strategies and Shortcomings of Social Judgment*. Englewood Cliffs, NJ: Prentice-Hall.
Oaksford, M. and N. Chater (1994) "A rational analysis of the selection task as optimal data selection." *Psychological Review* 101: 608–31.
Papineau, D. (1987) *Reality and Representation*. Oxford: Blackwell.
—— (1993) *Philosophical Naturalism*. Oxford: Blackwell.
Pessin, A. and S. Goldberg (eds.) (1996) *The Twin Earth Chronicles: Twenty Years of Reflection on Hilary Putnam's "The Meaning of 'Meaning'."* Armonk, NY: M. E. Sharpe.
Posner, M. I. (1989) *Foundations of Cognitive Science*. Cambridge, MA: MIT Press.
Prinz, J. (2004) *Gut Reactions: A Perceptual Theory of Emotion*. Oxford: Oxford University Press.
Putnam, H. (1975) "The meaning of 'meaning'." In K. Gunderson (ed.) *Language, Mind and Knowledge*, Minneapolis, MN: University of Minnesota Press, pp. 131–93.
Quine, W. V. O. (1951) "Two dogmas of empiricism." *Philosophical Review* 60: 20–43.
—— (1960) *Word and Object*. Cambridge, MA: MIT Press.
Rescorla, R. A. and A. R. Wagner (1972) "A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement." In A. H. Black and W. F. Prokasy (eds.) *Classical Conditioning II: Current Research and Theory*, New York: Appleton-Century-Crofts, pp. 64–99.
Rosch, E. and C. Mervis (1975) "Family resemblances: studies in the internal structure of categories." *Cognitive Psychology* 8: 382–439.
Rumelhart, D. E., J. L. McClelland, and the PDP Research Group (1986) *Parallel Distributed Processing Explorations in the Microstructure of Cognition*. Cambridge, MA: MIT Press.
Ryle, G. (1949) *The Concept of Mind*. London: Hutchinson.
Samuels, R. (1998) "Evolutionary psychology and the massive modularity hypothesis." *British Journal for the Philosophy of Science* 49: 575–602.
—— (2002) "Nativism in cognitive science." *Mind and Language* 17: 233–65.
Samuels, R., S. Stich, and P. D. Tremoulet (1999) "Rethinking rationality: from bleak implications to Darwinian modules." In E. LePore and Z. Pylyshyn (eds.) *What Is Cognitive Science?*, Oxford: Blackwell, pp. 74–120.

Samuels, R., S. Stich, and M. Bishop (2002) "Ending the rationality wars: how to make disputes about human rationality disappear." In R. Elio (ed.) *Common Sense, Reasoning, and Rationality*, Oxford: Oxford University Press, pp. 236–68.
Scholl, B. and A. Leslie (1999) "Modularity, development and 'theory of mind'." *Mind and Language* 14: 131–53.
Skinner, B. F. (1953) *Science and Human Behavior*. New York: Macmillan.
—— (1957) *Verbal Behavior*. New York: Appleton-Century-Crofts.
Spelke, E. S. (1988) "The origins of physical knowledge." In L. Weiskrantz (ed.) *Thought Without Language*, Oxford: Clarendon Press, pp. 168–84.
Stein, E. (1996) *Without Good Reason*. Oxford: Clarendon Press.
Stich, S. (1975) "The idea of innateness." In S. Stich (ed.) *Innate Ideas*, Los Angeles: University of California Press, pp. 1–22.
—— (1985) "Could man be an irrational animal?" *Synthèse* 64: 115–35.
—— (1990) *The Fragmentation of Reason*. Cambridge, MA: MIT Press.
—— (1992) "What is a theory of mental representation?" *Mind* 101: 243–61.
Stich, S. and S. Laurence (1994) "Intentionality and naturalism." *Midwest Studies in Philosophy* 19: 159–82.
Stich, S. and T. A. Warfield (eds.) (1994) *Mental Representation*. Oxford: Blackwell.
Thorndike, E. L. (1911) *Animal Intelligence*. New York: Macmillan.
Tversky, A. and D. Kahneman (1971) "Belief in the law of small numbers." *Psychological Bulletin* 76: 105–10.
—— and —— (1982) "Judgments of and by representativeness." In D. Kahneman, P. Slovic, and A. Tversky (eds.) *Judgment under Uncertainty: Heuristics and Biases*, Cambridge: Cambridge University Press, pp. 84–98.
Tye, M. (1992) "Naturalism and the mental." *Mind* 101: 421–41.
Warfield, T. A. and S. Stich (1994) Introduction. In S. Stich and T. A. Warfield (eds.) *Mental Representation*, Oxford: Blackwell.
Wason, P. C. (1966) "Reasoning." In B. Foss (ed.) *New Horizons in Psychology*, Harmondsworth: Penguin, pp. 135–51.
Watson, J. B. (1913) "Psychology as the behaviorist views it." *Psychological Review* 20: 158–77.
Wright, L. (1973) "Functions." *Philosophical Review* 82: 139–68.

Further reading

General philosophy of psychology

G. Botterill and P. Carruthers, *The Philosophy of Psychology* (Cambridge: Cambridge University Press, 1999) provides a more in-depth treatment of most topics covered here, as well as others. R. Cummins and D. Cummins, *Minds, Brains and Computers: The Foundations of Cognitive Science* (Oxford: Blackwell, 2000) reprints many seminal papers in cognitive science. Most general anthologies of philosophy of mind also contain selections relevant to topics discussed here (e.g. W. Lycan (ed.) *Mind and Cognition*, 2nd edn., Oxford: Blackwell, 1999).

Cognitivism

H. Gardner, *The Mind's New Science: A History of the Cognitive Revolution* (New York: Basic Books, 1985) remains an accessible history of the development of cognitive science. The first part of N. Block (ed.) *Readings in Philosophy of Psychology* (Cambridge, MA: Harvard University Press, 1980) reprints some of the historically important texts. For a valuable collection of philosophical reactions to Chomsky's work, see L. Antony and N. Hornstein, *Chomsky and His Critics* (Oxford: Blackwell, 2003).

Nativism

L. Hirschfeld and S. Gelman (eds.) *Mapping the Mind: Domain Specificity in Cognition and Culture* (Cambridge: Cambridge University Press, 1994) is an influential collection of nativist papers. F. Cowie, *What's Within? Nativism Reconsidered* (New York: Oxford University Press, 1999) and J. Elman et al., *Rethinking Innateness: A Connectionist Perspective on Development* (Cambridge: Cambridge University Press, 1996) are important critiques.

Modularity

J. Fodor, *The Modularity of Mind* (Cambridge, MA: MIT Press, 1983) is the locus classicus of modularity. Fodor, *The Mind Doesn't Work That Way* (Cambridge, MA: MIT Press, 2000) presents his later view, and R. Samuels, "Evolutionary psychology and the massive modularity hypothesis" (*British Journal for the Philosophy of Science* 49, 1998) is critical in various ways of modularity.

Rationality

The anthologies D. Kahneman et al. (eds.) *Judgment under Uncertainty: Heuristics and Biases* (Cambridge: Cambridge University Press, 1982) and T. Gilovich et al. (eds.) *Heuristics and Biases: The Psychology of Intuitive Judgment* (Cambridge: Cambridge University Press, 2002) provide copious empirical evidence. The alternative approach advocated by Gigerenzer and colleagues is presented in a series of books, including their *Simple Heuristics that Make Us Smart* (New York: Oxford University Press, 1999). E. Stein, *Without Good Reason* (Oxford: Clarendon Press, 1996) is a good overview of the debate.

Intentionality

S. Stich and T. A. Warfield (eds.) *Mental Representation* (Oxford: Blackwell, 1994) collects many of the important papers.