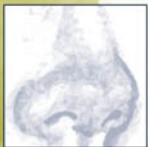# Furnishing the Mind

Concepts and Their Perceptual Basis

Jesse J. Prinz

Furnishing the Mind

Representation and Mind
Hilary Putnam and Ned Block, editors

Furnishing the Mind

Concepts and Their Perceptual Basis

Jesse J. Prinz

To the memory of Joachim Prinz

1

Desiderata on a Theory of Concepts

How comes [the mind] to be furnished? Whence comes it by that vast store which the busy and boundless fancy of man has painted on it with an almost endless variety? Whence has it all the materials of reason and knowledge? To this I answer, in one word, from EXPERIENCE.
Locke (1690, II.i.2)

1.1   Introduction

Without concepts, there would be no thoughts. Concepts are the basic timber of our mental lives. It is no wonder, then, that they have attracted a great deal of attention. Attention, but not consensus. The nature and origin of concepts remain matters of considerable controversy. One item of agreement, however, is that Locke's theory of concepts is wrong. Locke claims that all concepts (or, in his idiom, "ideas") are the products of experience. They are perceptually based. Reason gets its materials from the senses. Those mental states by which we see, hear, and smell the world are used to furnish the faculties by which we think, plan, and solve problems. This position goes by the name "empiricism."

Empiricism is a phoenix in the history of philosophy. It has been perpetually invented, destroyed, and reinvented. In the eyes of some, philosophers resist empiricism only because they suffer from an occupational distaste for the obvious. In the eyes of others, empiricism is a terrible mistake that philosophers have had to rediscover time and again. In the present climate, the latter perspective prevails. Many philosophers and researchers in other fields assume that empiricism has been decisively refuted.

I defend a dissenting view. While certain traditional forms of empiricism are untenable, a properly modernized empiricist account shows

tremendous promise. It turns out that Locke's thesis can be reconciled with, and even supported by, the findings of cognitive science. More to the point, a modernized version of concept empiricism can outperform its rivals. A modernized empiricism can counter objections to older empiricist theories *as well as* objections to nonempiricist theories. Arriving at this heretical conclusion will take some time. It must first be shown that the leading nonempiricist theories are inadequate.

It helps to begin with a neutral characterization of concepts so we can home in on the items of disagreement. One such characterization is found in Locke's apt phrase that concepts are the "materials of reason and knowledge." A similar sentiment is expressed by the assertion that concepts are constituents of thoughts. My thought that aardvarks are nocturnal, for example, contains the concept AARDVARK and the concept NOCTURNAL.[1] This characterization leaves open this possibility that there could be thought constituents that are not concepts, and it leaves open the possibility that concepts can occur outside thoughts. Perhaps one can simply token the concept AARDVARK, as in an episode of free association, without having a full-fledged thought.[2] The characterization also says nothing about what thoughts are and what it is to be a constituent. It is sometimes said that concepts are to thoughts as words are to sentences, but this analogy is misleading if one does not buy into the view that thoughts are sentencelike.

The claim that concepts are thought constituents shows why they are so fundamental to a theory of the mind. Psychological theories seek to explain behavior. In both folk and scientific psychology, this is typically done be ascribing thoughts. We negotiate our environments by thinking about them. Thinking itself subsumes such abilities as planning, reasoning, problem solving, deciding, and recalling. To provide an adequate theory of these abilities, we need a theory of thoughts, and a theory of thoughts requires a theory of what thoughts are made of. If concepts are the constituents of thoughts, then they must play a foundational role in any complete theory of cognition. About this, there is considerable agreement.

There is also considerable agreement about some of the further properties that concepts must have. There are certain phenomena that are widely recognized as explanatory goals for a theory of concepts. These

can be used to form a wish list—a list of desiderata that a theory of concepts would ideally explain. These desiderata can serve as a litmus test for a theory of concepts. It will be my contention that none of the leading theories satisfies all of the desiderata. This opens a space for an alternative.

1.2   Desiderata

The desiderata I present are widely accepted among philosophers and psychologists.[3] They include the phenomena that have motivated the postulation of concepts in the first place. They are the stuff of textbooks: explanatory goals so widely embraced that they are tedious to review.

   Of course, full consensus is too much to hope for. Some might think that no account of concepts can satisfy each of the desiderata I discuss. To insist that an adequate theory of concepts must explain them all would beg the question against those who have more modest explanatory goals. Instead, I offer a conditional thesis: *if* a theory of concepts can accommodate all of the desiderata, then it has an explanatory advantage over its more modest competitors.

1.2.1   Scope
An adequate theory of concepts must have sufficient expressive power or breadth to accommodate the large variety of concepts that we are capable of possessing. The human conceptual repertoire ranges from the sensory to the abstract. We have concepts of readily observable states within ourselves, like PAIN; theoretically derived concepts, such as ELECTRON; and seemingly formal concepts, such as NUMBER. We have concepts of natural kinds, such as FROG; artifacts, such as BOAT; and social kinds, such as MOTHER or DEMOCRACY. This diversity cannot be neglected. Some theories are particularly adept at handling one kind of concept and embarrassingly poor at dealing with others.

1.2.2   Intentional Content
Concepts represent, stand in for, or refer to things other than themselves. My AARDVARK concept is about aardvarks; it refers to all and only aardvarks. Philosophers call this property "intentionality." To say that

concepts have intentionality is to say that they refer, and those things to which they refer, I call their intentional contents. Intentional states can refer to both actual things and merely possible things. I can have a concept that represents unicorns.

An adequate theory should help us understand how concepts attain their intentional contents. Many philosophers are explicit about this desideratum. Psychologists tend to be less explicit, but they almost always assume that concepts have intentionality and identify concepts by their intentional contents. For example, the concept FROG is so called because it is the concept that represents frogs.

The intentionality of concepts plays important explanatory roles. Our ability to represent things contributes to an explanation of our ability to behave in ways that are sensitive to those things. The actions of certain simple organisms might not require the mediation of intentional mental states, because they are fully and directly determined by stimuli present in their environments. But our minds are more powerful. We can act with flexibility and forethought, choosing between different courses of action and anticipating future consequences. These abilities seem to demand representations that stand in for extramental objects. Representations can be manipulated by the mind independently of the things they represent. As a result, we can engage in behavior that is sensitive to extramental objects even when those objects are not present. For example, by manipulating a FROG representation, one can devise, mentally test, and ultimately act on a plan to catch a frog.

Despite the consensus that concepts refer, there is some controversy about what they refer to. These days, most philosophers assume that many of our concepts refer to categories whose boundaries are determined by nature. My FROG concept refers not to the set of things I take to be frogs, but to the set of frogs.[4] The set of frogs, in turn, is determined by nature, not by us. It is a natural kind. Some researchers may be inclined to resist this kind of realism. They think that category boundaries are imposed on the world by our concepts. On this view, so-called natural-kind concepts really pick out categories that depend on human thoughts and practices. For some, being a frog is something like being tasty, being a chair, or being the U.S. president. All these categories are

real in some sense, but they depend on human cognitive abilities, goals, or social practices. Placing natural kinds in this group makes them dependent on us. On a more radical version of the view, my concepts refer only to what I categorize under them, regardless of the practices of other individuals. If I fail to identify some odd-looking species as frogs, despite the fact that science does, my FROG concept simply excludes them.

I presuppose a strong form of realism. I assume that my FROG concept really refers to a naturally delineated category, despite the fact that I might misclassify a few instances. If I were to insist that some odd looking species is not a frog and then subsequently discover that it shares the same underlying properties as the things I admit are frogs, I would change my original judgment. I would not say that the species changed its ontological identity as a consequence of my discovery. It was a frog species all along. I had simply been fooled by appearances. If such changes in my categorization decisions implied that the reference of my concept had changed, there would be no way to explain why I took my original view to be erroneous and changed my mind. If reference is determined merely by how I actually categorize, there could be no such thing as error.

Opponents of strong realism can try to handle such cases by appeal to ideal observation conditions or scientific consensus. I do not think this will suffice. I side with those who say that my FROG concept can refer to frogs even if there are certain instances that *no one* reliably identifies as such, even under ideal conditions. Just as I can imagine my own errors in categorization, I can imagine systematic errors throughout my community, even under the best circumstances. This is not to say that such errors would actually occur. The strong realist intuition only requires that they could occur. It is even conceivable that human cognitive limitations prevent us from ever discovering certain of nature's joints. In such cases, I believe, we can still pick out kinds whose borders are defined by such joints. Little will hinge on this strong realist claim, but I state it for the record. The weaker claim, according to which my concepts can have intentional contents that neither I nor any member of my community can articulate *at present* is important. It is a principle underlying many

scientific pursuits and is implicit in experimentally demonstrable human categorization tendencies.

1.2.3   Cognitive Content

There are well-known reasons for thinking that concepts cannot be individuated by intentional content alone. Two closely related arguments derive from Frege's (1893) philosophy of language. First, Frege draws our attention to the fact that a true identity statement involving two distinct terms can be informative despite the fact that those terms have a common referent. For example, it can be surprising to discover that Lewis Carroll is Charles Dodgson. If we grasped these names by grasping their referents (i.e., intentional contents), the surprise would be inexplicable because their referents are the same. Second, Frege observes that we cannot freely substitute coreferring terms in some linguistic contexts. "Sally believes that Charles Dodgson is a logician" and "Sally believes that Lewis Carroll is a logician" can differ in truth value, even though "Carroll" and "Dodgson" corefer. If reference exhausted the content of terms, these sentences would have the same truth value.

Frege offers these arguments for the limitations of reference in developing an account of linguistic meaning, but parallel examples can be constructed without mentioning language. As with terms, the identification of two coreferring concepts can be informative, and as with sentences, we can have a belief containing one of a pair of coreferring concepts without having a corresponding belief containing the other. This suggests that conceptual content too cannot be exhausted by reference. Peacocke (1992) uses this Fregean insight in discussing identity conditions on concepts. He stipulates that two concepts count as distinct just in case substituting one for the other can render an uninformative thought informative. This is true, he says, even in cases where concepts corefer. This fact about concepts may offer the best explanation of informative identities and substitution failures in language. The linguistic cases may arise as a result of the fact that some coreferential terms are associated with distinct concepts.

This is essentially Frege's position. He solved his puzzle cases by introducing the notion of sense. "Carroll" and "Dodgson" have different

senses, but the same referent. In grasping these terms, we grasp the sense, and thereby fail to discover the identity of their referents. The sense of an expression is a part of its content other than its referent. Beyond that, it is not always clear what Frege meant by "sense." What is the ontological status of a sense? How do senses relate to reference? Frege uses the term "sense" in several ways (see Dummett 1981). On one standard interpretation, which is discussed in the next chapter, senses are definitional abstract objects that determine what our concepts refer to. This view has come under attack. For example, some philosophers of language argue that reference is determined by a term's causal history (e.g., Kripke 1980, Donnellan 1972, Putnam 1975). On this theory, a term refers to what its originators were pointing two at the moment the term was introduced. I call this the etiological theory. Defenders of the etiological theory often claim that the information we associate with a term is not part of its meaning. Meaning is exhausted by reference. Defenders of other recent semantic theories share this opinion.

Even if these reference-based accounts are correct, Frege is surely right to say that reference cannot exhaust our *understanding* of terms. When we consider the *psychology* of language, or when we consider the non-linguistic cases just described, the need for a kind of content that transcends reference is manifest. Even if one insists that such contents should play no part in a theory of linguistic semantics, they are indispensable for understanding the concepts we deploy in thought. I return to this point below.

To say that we need a construct that individuates concepts more finely than referents does not entail that we must adopt Frege's notion of sense. In particular, it does not mean that we must say that all concepts are associated with definitional abstract objects. We do, however, need *some* kind of content other than reference, or intentional content, as it was called in the last section. I call this further requirement, "cognitive content." Cognitive content is what allows two coreferential representations, be they terms or concepts, to seem semantically distinct to a cognitive agent.

This is only a first approximation. In addition to explaining how coreferential terms can seem different, cognitive content is needed to explain

how concepts that are not coreferential can seem alike. Putnam (1975) introduced a celebrated counterpart to the cases introduced by Frege. He imagines a world, Twin Earth, which is almost exactly like our own world. Every one here has a doppelgänger on Twin Earth. The only difference is that the stuff that has all the superficial properties of water on Twin Earth is not H_2O but some other compound called XYZ. On Twin Earth, XYZ is the clear, tasteless liquid that fills rivers and streams. The concept that I express by "water" refers to H_2O, even if I am ignorant of chemistry, because I apply it to stuff that happens to be H_2O. The concept that my Twin Earth doppelgänger expresses by his word "water" refers to XYZ because that is the local waterlike stuff in his world. Nevertheless, there is some intuitive sense in which our two extensionally distinct concepts are alike. They have the same cognitive content. An adequate account of concepts should explain how coreferential concepts can differ and how divergently referential concepts can be alike.

1.2.4   Acquisition

A fourth desideratum is that a theory must ultimately support a plausible explanation of how concepts are acquired. This requirement has two facets. On the one hand, we need to accommodate ontogenetic acquisition. How does an individual come to possess a given concept? Concepts that are thought to be learned rather than innate must be learnable. A theory of concepts must allow for this.

In addition, an adequate theory must be commensurable with a phylogenetic story. It must lend itself to an explanation of how innate concepts (if such exist) entered the human genome and how we evolved to be able to acquire those concepts that are not innate. Just as we must be able to explain how the language faculty evolved, we must be able to tell a story about how the conceptual faculty evolved. The difficulty of meeting this requirement is proportionate to the degree to which one's theory links concepts to faculties whose evolution is already well understood.

This is not to say that a theory of concepts must come prepackaged with a theory of concept evolution or even a theory of concept learning. A theory of concepts should merely lend itself to such acquisition theo-

ries. If one theory of concepts is compatible with a more plausible, independently motivated theory of acquisition then it should be preferred over an incompatible theory.

1.2.5   Categorization

Reference is a semantic relation, but it also has an epistemic counterpart. In addition to the set of things to which a given concept refers, there is a set of things to which a concept is *taken to* refer. We have mechanisms for forming beliefs about what things fall under our concepts, mechanisms of categorization. Concepts are often identified with such mechanisms.

As I use the term, categorization encompasses two different, but closely connected abilities. "Category identification" is manifested when a person identifies the category under which an object belongs. Various kinds of experimental tasks involve category identification. In some experiments, subjects are asked to verbally identify a verbally described or visually presented object. In other experiments, subjects are asked to confirm a categorization judgment, as in "True or false: canaries are birds."

Second, "category production" is manifested when a person identifies which attributes an object possesses if it is a member of a given category. Experiments often assess this by asking subjects to describe categories or to decide whether members of a given category have some specific attribute. In other experiments, subjects are asked to rate the similarity between two categories or to draw inferences about features possessed by category members.

Recognition (identification) and production can work in concert. For example, when tracking an object, one must recognize it across different transformations, but to do that, one must often anticipate what form those transformations will take. Such strategic tracking depends on category production. Another combination of these abilities is category hypothesis confirmation. Once one has tentatively identified the category of a partially concealed object by using available attributes, one can produce knowledge about concealed attributes, and then confirm the original identification by searching for those concealed attributes.

Psychologists widely believed that a theory of concepts should explain these abilities in a way that is consistent with empirical findings. For example, experiments have shown that not all members of a category are created equal. Instead, one finds many "typicality effects." People readily rate some category members as more typical than others (Rosch 1973, Rosch 1975, Mervis, Catlin, and Rosch 1976). These typical members are categorized faster and produced more readily during category production tasks (Smith, Shoben, and Rips 1974, Rosch 1978). Likewise, some attributes are rated as more typical for a category and are produced more readily (Rosch and Mervis 1975). Category members possessing many typical attributes are rated as highly typical (Hampton 1979).

Furthermore, not all categories are created equal. Any given object can fall under many different categories. For example, a single object can be a rottweiler, a dog, an animal, a living thing, and so on. It turns out that the intermediate level of abstraction is privileged (Brown 1958, Berlin and Kay 1969, Rosch, Mervis, Gray, Johnson, and Boyes-Braem 1976). Rosch and her colleagues call this the "basic level" of categorization. Subjects can usually identify an object as a dog faster than they can identify it as a rottweiler (the subordinate level) or as an animal (the superordinate level). They also seem to acquire the concept DOG earlier in development. An adequate theory of categorization should predict and explain these asymmetries.

It might be objected that the categorization desideratum, and these results in particular, introduces an unfair bias in favor of certain theories of concepts. Philosophers rarely try to accommodate such psychological findings when developing their theories. In fact, some philosophers think that a theory of concepts need not explain categorization at all. The constituents of thoughts, they contend, may have little to do with the mechanisms by which we classify objects under the conditions psychologists explore. It is certainly conceivable that we have one set of representations for forming thoughts about frogs and another for picking them out in a crowd (Armstrong, Gleitman, and Gleitman 1983). If concepts are thought constituents, it might be best to remove the explanation of categorization from the list of desiderata.

I address such dissenting opinions more fully in chapters 2 and 4. As a preliminary response, however, I emphasize two points. First, psychologists have found evidence that the effects found in categorization studies appear in other contexts as well. For example, studies have shown typicality effects in inductive inference (Sloman 1993, see also Smith 1989). We are more likely to draw an inference about a property from one subordinate-level category to another if the former is a typical instance of the basic-level category that subsumes them. Such findings support the contention that categorization representations coincide with representations used in thinking.

Second, eliminating the categorization desideratum would strongly bias the case against psychological theories of concepts. In psychology, an enormous amount of the research on concepts has focused on categorization. Concepts are often stipulated to be the cognitive mechanisms by which we categorize. If a theory of concepts were absolved of its obligation to explain categorization, most psychological accounts would be rendered moot. Categorization certainly stands in need of an explanation. If psychological theories were to satisfy all desiderata including categorization and philosophical theories were to satisfy all but categorization, psychological theories would have an explanatory advantage. This does not mean that we should disqualify theories that cannot explain categorization. Instead, we should say that the ability to explain categorization is an asset and that theories lacking this asset are able to defeat their rivals only if they outperform them on other desiderata.

A further objection against the categorization desideratum is that some concepts refer to classes whose members we cannot directly recognize. For example, most of us who possess the concept ELECTRON cannot recognize electrons. Perhaps, then, it is too stringent to demand that a theory of concepts explain categorization. I think this objection fails. First, to take the present example, an inability to recognize electrons would not rule out our ability to engage in electron-categorization behavior broadly conceived. Categorization includes category production, an ability possessed by many of those who could never recognize electrons. If a theory of concepts explains why someone with an ELECTRON concept is likely to characterize electrons as negatively charged particles, it

satisfies the desideratum under discussion. Second, even if some concepts never involve categorization (no examples come to mind), the desideratum would not be threatened. *When categorization does occur*, it involves concepts: categorizing something is placing it under a concept or characterizing it by means of concepts. It is natural to hope for and give preference to a theory of concepts that accounts for these abilities.

1.2.6 Compositionality

In many important respects, our cognitive capacities are unbounded. There appears to be no upper limit on the number of distinct beliefs we can entertain, plans we can devise, and sentences we can comprehend. Every day we entertain a breathtaking number of novel thoughts. These are induced by our experiences, tailored by our goals, or awakened by our casual musings. This hyperfertility is achieved using finite means. As finite beings, we have finite minds. Finite minds can only store a limited stock of concepts. Myriad thoughts must somehow be derivable from that limited stock. There is a highly plausible explanation of this. A finite set of concepts can engender a boundless capacity for unique thoughts if those thoughts are derived by combining concepts compositionally.

Concepts are compositional just in case compound concepts (and thoughts) are formed as a function of their constituent concepts together with rules of combination. For example, a compositional system allows one to form the thought that aardvarks are nocturnal by combining one's AARDVARK concept with one's NOCTURNAL concept using the very same combination rule used for forming other thoughts, such as the thought that cows are herbivorous, or that politicians are egomaniacal. Likewise, the very same concepts, AARDVARK and NOCTURNAL, can be used to form other thoughts in a compositional system, e.g., the thought that aardvarks eat insects and bats are nocturnal. The same rules and the same stock of primitives can be used to form different combinations.

Compositionality explains the extreme fertility, or, as it is often called, productivity, of thought, because in principle a finite set of concepts and a finite set of combination rules can be used to generate an infinite number of distinct compounds. Compositional combination becomes infinitely productive when the rules allow for an endless variety of novel combinations. The simplest examples of such rules are recursive func-

tions in logic and grammar. The word "and," for instance, can be iterated indefinitely; we can say "*A* and *B*," "*A* and *B* and *C*," "*A* and *B* and *C* and *A*," and so on. With a handful of recursive rules and a stock of primitives, the variety of possible thoughts becomes staggering. The ability to form novel thoughts can be explained within a compositional framework. If a person knows the constituent concepts and the combination rules, then she can use them to form thoughts that have never been entertained before. Chomsky (1968) gives a seminal presentation of this kind of argument for compositionality in the context of language. Fodor has argued aggressively for its inclusion among the nonnegotiable conditions on a theory of concepts (e.g., Fodor 1981, 1994, 1998).

Fodor and his colleagues also offer another argument for compositionality. They say that it provides the best explanation for what they call the systematicity of thought (e.g., Fodor 1987, Fodor and Pylyshyn 1988, Fodor and Lepore 1992). Our ability to form certain thoughts, such as the thought that Oscar ate the squid, seems to carry with it the ability to form certain others, such as the thought that the squid ate Oscar. Anyone who can entertain the first thought can entertain the second. This fact can be explained if concepts are compositional. The ability to think that Oscar ate the squid co-occurs with the ability to form the thought that the squid ate Oscar because these two thoughts are comprised of the same concepts and generated using the same combination rules. If we needed to learn separate combination rules for forming each thought, such systematic relations would not arise. A similar insight underlies Evans's (1982) defense of what he calls the Generality Constraint on concept possession. According to Evans, a person possesses the nominal concept *a* and the predicative concept *F* only if she can form the thoughts that *a* is *G* for any possessed predicate concept *G* and that *b* is *F* for any possessed nominal concept *b*. It follows naturally from the Generality Constraint that anyone who can form certain thoughts is able to form other thoughts by use of the same concepts and combination rules.

The compositionally requirement stands in need of one clarification. I said that concepts are compositional if compounds are generated as a function of their constituent concepts. Some standard formulations are stated in terms of contents. It is often said that the *content* of a thought

(or compound concept) is compositional just in case it is a function of the *contents* of the concepts constituting that thought together with rules of combination. In other words, the claim that compounds are formed from constituents carries with it the idea that constituents contribute their contents to compounds. But what does "content" mean here? This question is complicated by the fact that I distinguish two distinct kinds of content, intentional and cognitive. What kind of content does compositionality pertain to?

I answer that both intentional and cognitive content must be compositional because both kinds of content are implicated by the productivity and systematicity of thought. Saying that we are capable of entertaining an unbounded number of distinct thoughts implies that we can entertain an unbounded number of thoughts with distinct intentional and cognitive contents. Saying that certain thoughts exhibit systematicity implies that our ability to have thoughts with certain intentional and cognitive contents carries with it the ability to entertain other thoughts with distinct cognitive and intentional contents. Thus, the compositionality desideratum carries with it two component requirements: intentional-content compositionality and cognitive-content compositionality.[5]

1.2.7   Publicity

The final desideratum is publicity. Concepts must be capable of being shared by different individuals and by one individual at different times. This requirement has been emphasized by many (e.g., Rey 1983, Peacocke 1992, Fodor 1998). It must be satisfied if concepts are to play some of their most important explanatory roles. Two of these roles stand out (Fodor and Lepore 1992). First, it is almost universally assumed that concepts play a pivotal role in linguistic communication. According to the standard picture, people understand each other's words in virtue of the fact that they associate the same (or quite nearly the same) concepts with those words. If no two people associate the same concepts with their words, then communication is impossible. Therefore, concepts must be sharable.

A second reason for thinking concepts are public is that concepts are implicated in intentional explanations of behavior. An intentional expla-

nation of behavior is one that explains what a person does by appeal to her mental states. For example, Mary opened the liquor cabinet because she desired a glass of scotch and believed that she could find some there. As this example illustrates, typical intentional explanations make reference to propositional attitudes, and attitudes are composed of concepts. Perhaps the most striking feature of intentional explanations is their apparent generality. A single intentional explanation can subsume many different people. Felix, Hugo, and Greta might all open their respective liquor cabinets for precisely the same reason that Mary did. But, actions can be motivated by the same attitudes only if those attitudes are composed of the same concepts. If intentional explanations generalize, concepts must be sharable.

As with compositionality, the publicity requirement must be explicated along two dimensions. Concepts must be sharable in both intentional content and in cognitive content. Explaining communication clearly requires that intentional contents be sharable. To understand what someone is saying, we must know what their words and thoughts refer to. For this, their concepts and ours must refer to the same things. It would be a mistake to assume that we can satisfy the publicity requirement by merely establishing that people can have coreferential concepts. It is equally important to show that concepts can share their cognitive contents.

For example, consider Twin Earth cases. There is a strong intuition that I share something with my doppelgänger on Twin Earth when he and I think about the stuff in our respective rivers and lakes. My concept refers to H_2O and his refers to XYZ, but these concepts can arguably be subsumed by some of the same psychological laws. My desire to drink the stuff I call "water" disposes me to the same behaviors as his desire to drink the stuff he calls "water." Thus, there is reason to think that concepts can be importantly alike despite differences in intentional content. Shared cognitive contents provide the best explanation.

Second, consider interpersonal versions of Frege cases (see Aydede 1998). Standard Frege cases involve one person with two coreferential concepts. Interpersonal versions involve two or more people both of whom possess the same pair of coreferential concepts. For example, the ancient Greeks falsely believed that the morning star (Hesperus) is

different from the evening star (Phosphorus). They had two different concepts, and their beliefs involving those concepts can be subsumed under the same generalizations. For example, anyone who wanted to see Phosphorus would go outside in the evening, and not in the morning. These common behaviors cannot be explained by shared intentional contents, because that would obscure the fact that other behaviors with the same intentional contents lead to different behaviors. In particular the desire to see Hesperus makes people go out in the morning. Once again, the relevant kind of concept sharing involves shared cognitive contents. Thus, the publicity desideratum demands that both cognitive content and intentional content be potentially sharable.

The issue of concept sharing raises many thorny questions. There is considerable debate about whether children share concepts with adults (see, e.g., Carey 1985), whether people with radically different beliefs share concepts (Kuhn 1962), whether people with healthy minds share concepts with people who have damaged minds (Stich 1983), and whether humans share concepts with other animals (Sterelny 1990). I do not address these questions here. Even if they are all answered in the negative, there remains ample evidence that concept sharing is possible in ordinary cases. The exotic cases just mentioned are exactly the cases in which the arguments from communication and intentional explanation are least persuasive. For example, it is highly contentious to claim that we truly communicate or fall under the same psychological laws with nonhuman animals. In less exotic cases, we have every reason to believe that concept sharing occurs. This is the claim that an adequate theory of concepts must accommodate.

1.3   Do We Need Language Desiderata?

In laying out these desiderata, I mentioned relations between concepts and language several times. For example, I said that conceptual differences might underwrite the informativeness of linguistically expressed identities, and I said that conceptual publicity is needed to explain linguistic communication. Such remarks raise the question, How exactly are concepts related to language? There are two questions of special

interest here. First, we might ask how concepts are related to linguistic meanings. Second, we might wonder whether one can have concepts without language. I address these in turn.[6]

It is widely assumed that concepts should figure into a theory of linguistic meaning. Some consider this a desideratum on a theory of concepts (e.g., Rey 1983). In the strongest form of this desideratum, one might say that concepts simply *are* the meanings of words. Call this the meaning desideratum. I already hinted at my reasons for leaving the meaning desideratum off the list. I noted that some recent semantic theories restrict meaning to reference. On such reference-based theories, the Fregean and Putnamian data that motivate the intentional-content desideratum should not be taken as evidence that meaning outstrips reference. These data only show that our ways of understanding words outstrip reference. Ways of understanding arguably belong to psychology rather than semantics.

On reference-based theories, concepts may play some role in a full theory of language. First, they may play a role in linguistic epistemology. In this capacity, a theory of concepts may be said to contribute to a theory of how we understand language or how we select linguistic forms in the service of communication. Even though the meaning of a word is exhausted by its referent, our understanding of the word depends on our understanding of its referent. And our understanding of referents is mediated by concepts. Likewise, in communication we typically choose words that refer to the objects or facts that we would like to express. Our knowledge of those objects and facts is, again, mediated by concepts.

Second, concepts may play a role in determining linguistic reference. My utterances of the word "dog" may refer to dogs virtue of being associated with a concept of mine that refers to dogs. This does not mean that the concept itself constitutes the meaning of the word, much less that the concept contributes *all* of its content to linguistic meaning. In particular, the cognitive content of a concept may have nothing to do with meaning. On this view, concepts merely contribute their referents.

All this suggests that it may be inappropriate to saddle a theory of concepts with the responsibility of providing a theory of meaning.

Concepts may play a role in semantic theories, but saying that concepts *are* meanings is not necessarily motivated in light of recent developments in semantic theory.

Of course, reference-based semantics have not gone unchallenged. One worry is that reference-based semantics cannot explain apparent semantic differences between distinct vacuous terms; "unicorn" and "centaur" both refer to the same thing, namely nothing, but they seem to have different meanings. Reference-based theories also offer no easy explanation of apparent restrictions on substitutivity. It is beyond the scope of this discussion to evaluate the success of such objections. It is enough to point out that reference-based semantics are both highly popular and incompatible with the meaning desideratum.

I turn now to the second question about concepts and language. Some philosophers have been tempted to say that public language is necessary for the possession of concepts. To those outside of philosophy the claim may sound absurd. Surely, we know that nonhuman animals engage in behavior that is sophisticated enough to warrant ascriptions of concepts. We know that human infants, chimps, and even parrots can categorize objects. They can identify which things go together. Even more dramatically, aphasics, who suffer serious linguistic deficits, do not exhibit more general cognitive impairments. Their behavior seems entirely appropriate to the situations that confront them, including manifest frustration with their linguistic deficits. Doesn't this show there can be concept possession without linguistic mastery? Another argument against the thesis that concepts depend on language owes to Fodor (1975). He argues that we need concepts to acquire language in the first place. How do we learn a word if not by mapping it onto a previously attained concept?

Despite such obvious motivations for attributing concepts to infraverbal creatures and persons, some philosophers have been tempted to defend the radical view that such attributions are inappropriate. One reason for this odd-sounding claim can be extracted from the philosophy of Wittgenstein (1953). Here is an argument loosely drawn from Wittgenstein's critique of private language. According to Wittgenstein, concepts can be individuated by how they are used. Having a concept is being able to follow a rule for using that concept. For something to count as a rule, there must be criteria for correctness. Merely thinking that one

is following a rule is not the same as following a rule. If rules were private rather than public, there would be no way to confirm that you were conforming to them. If I on one occasion stipulate that I will always use a private concept in some particular way and then use that concept on a subsequent occasion, I will have no way to be sure that I am using it in the same way. I may misremember my initial rule, and there will be no one to correct me, no one to keep me honest. There can be no private criteria for correctness. Correctness and rules only have application in public contexts. Thus, if concepts are individuated by how they are used, the use that matters must be public. Concepts are publicly used in language. Thus, there can be no concepts without language.

If this argument is right, attribution of concepts to infraverbal creatures and persons must be taken with a grain of salt. Such attributions show only that there are some superficial similarities between them and us. An infraverbal creature can appear to possess a concept by, say, sorting things in some way, but there can be no criteria for correctness. If a parrot groups together a bunch of triangles and then includes a square, we cannot say that it made a mistake. Fodor's argument can also be answered. The early language learner may be acquiring words by mapping them onto mental states, but those mental states are not bona fide concepts, because they are not governed by criteria of correctness. Only when a word is in place and anchored to communally determined rules for correct application can the child be said to have a concept. Perhaps we need to have the ability to sort things in order to learn a concept, but sorting only becomes subject to correction, and hence conceptual, when brought under linguistic labels. If these Wittgensteinian considerations are right, then we might want to introduce another desideratum: a theory of concepts must ensure that concept possession requires language.

This brief treatment cannot do justice to Wittgenstein's philosophical outlook. I wish only to show where his position may be vulnerable. Wittgenstein's claim that there can be no private criteria for correctness can be challenged in various ways. First, a number of recent philosophers have proposed "naturalized" theories of error (see chapter 9 for an example). These theories purport to show that correctness need not depend on public policies. Some views explain correctness in terms of

conformity to laws, rather than rules, and others explain correctness in terms of conformity to evolved or designed functions (Fodor 1990, Millikan 1984). If these accounts pan out, Wittgenstein's claim will be refuted.

Second, Wittgenstein's opposition to private criteria for correctness turns in part on considerations about how difficult it would be for an individual to confirm that she was conforming to a private rule. This reasoning is flawed. The fact that a person cannot tell if she is following a rule does not prove that she is not following (or failing to follow) that rule. A criterion for correctness can simply be a correct way of conforming to a rule rather than a method of verifying that one is conforming. Wittgenstein intentionally conflates conformity with knowledge of conformity. The reason for this may stem from the view that rules involve a normative dimension. If someone fails to conform, she can be held accountable. If there is no way for a person to determine whether she is conforming or failing to conform, such accountability is threatened. In response, one might opt for a reliabilist measure of accountability. One might say that a person is justified in thinking that she is following a rule just in case the mechanism by which she reapplies the rule is reliable. For example, if memory systems work reliably, applying a rule from memory is a reliable process. A person can be held accountable for trying to apply a rule using an unreliable process.

A third worry about Wittgenstein's argument is that he may have an inflated picture of how correctness criteria work in public language. What does it mean to say that there are correct and incorrect uses of public words. One possibility, suggested by Chomsky (1991), is that public rules have more to do with authority than correctness. To say that there is a right way to use a word amounts to the claim that some language users use it in that way and will penalize those who do not. Misuse of language can be punished by public correction, social marginalization, and failure of communication. In the private case, there can be no serious threat of penalty. If I have a private rule and threaten to punish myself if I fail to conform, I know that it is within my power to refrain from carrying out that threat. If the difference between public and private rules amounts to the applicability of punishment, it would seem odd to say that public rules are privileged. After all, public rules do not have a

special relationship to justification or accountability on this picture. Moreover, even though I cannot threaten myself, misapplied rules could lead to costly mistakes that serve as punishments. If I misidentify a poisonous plant as another instance of a familiar nutritious plant, I will pay the consequences. If normativity amounts to threat of penalty, apparent differences between public and private criteria for correctness may be exaggerated.

Though not decisive, all these considerations show that the Wittgensteinian argument may be vulnerable. If Wittgensteinian arguments do not go through, many of the remarkable cognitive abilities exhibited by infraverbals can be taken as evidence for conceptual abilities. I assume that infraverbal creatures can have concepts.

Saying that one can possess concepts without language does not imply that language plays no role in our conceptual abilities. Language is, first of all, a dominant means of learning new concepts. People direct each other to new concepts by description, explicit definition, verbal pointing, and so forth. Moreover, concepts that get lexicalized are often more salient and easier to learn. Language also aids in using concepts. Highly complex concepts can be expressed using a single word. Those words can serve as conceptual placeholders in working memory to avoid the burden of processing the corresponding concepts in their full complexity. For linguistic creatures, some concepts may even be constituted by words. The best examples are concepts known only by deference to experts. A person ignorant of physics might arguably be said to possess a QUARK concept by possessing the word "quark" and being disposed to consult physicists about its use (Putnam 1975, Burge 1979).

I leave all of these as open possibilities. I am only committing to the assumption that concept possession can occur without language. All the theories of concepts that I consider have proponents who share this assumption. This does not entail that a complete theory of concepts can be developed without mentioning language. It does suggest, however, that one can present a theory of what concepts are without mentioning language. If some concepts depend on language, then this precept may have to be violated to accommodate the scope desideratum. I allow such violations, but I regard language-dependent concepts as the exceptions.

For this reason, language has a limited role in the chapters that follow. I part ways with those who think that a theory of concepts is motivated primarily in the context of a theory of language. Fortunately, mine is not a renegade position. Many researchers investigate concepts under the assumption that they are language-independent. The omission of a language desideratum in my list reflects a general, if tacit, bias in concepts research. It has been the burden of this section to show that this bias rests on a stable foundation.

1.4   Preview

In the following chapters, I use the desiderata presented above to measure the comparative success of various theories of concepts. Most of these theories claim that typical lexical concepts decompose into representations of features. A lexical concept is a concept expressed by a single word (e.g., BIRD, CAR, JUSTICE). A feature representation, or just "feature" for short, is a representation of some attribute possessed or condition met by objects falling under a concept. Thus, a BIRD concept may decompose into features such as FLIES, HAS WINGS, and so forth. Features are generally construed as concepts in their own right, some of which decompose into further features, and some of which do not (the "primitive" features). Most of the debates between competing theories of concepts concern the nature of features. Three questions can be distinguished. First, one can ask what the features constituting our concepts represent. Different theories claim that concepts decompose into features representing different kinds of attributes. They disagree about what kind of information a typical lexical concept contains about the category it represents. Second, one can ask which features are primitive. At one extreme are researchers who say that primitives are restricted to features representing perceivable properties; at the other extreme are those who say that primitives are roughly word-sized units. On the latter view, lexical concepts cannot be decomposed into more primitive features. The third question concerns the "mental medium" in which our concepts are couched. Are concepts like mental images, mental word lists, or something else? Most theories of concepts focus on the first of these questions, but all are important.

Chapters 2 through 4 evaluate the leading theories of concepts. None of these theories ends up with a perfect score. They each stumble on one desideratum or another. Such widespread failings are likely to invite a degree of skepticism. If none of the top theories of concepts can satisfy all the desiderata, perhaps the list is too demanding. Perhaps concepts cannot do all the things we want them to do.

I combat this skepticism by proposing a theory of concepts that is informed by the strengths and weaknesses of these other theories. That is the task of chapters 5 through 11. Readers familiar with prevailing theories and convinced of their shortcomings are invited to begin with chapter 5. There I define concept empiricism and offer preliminary arguments in its defense. In chapter 6, I describe how the empiricism I endorse differs from its historical ancestors. I also begin to show that this brand of empiricism can accommodate the desiderata by examining publicity and categorization. In chapter 7, I address scope, showing that perceptual representations have far greater expressive breadth than ordinarily appreciated. Acquisition is the subject of chapter 8, where I challenge received opinion that many of our concepts are innate. In chapter 9, I defend a theory of intentional content, building on causal and informational semantic theories. In chapter 10, I suggest an alternative to leading "narrow content" approaches to cognitive content. This leaves only compositionality, which I take up in the final chapter.

Together these chapters form an extended plea for recidivism. The theory that I defend is modern in that it avails itself of contemporary cognitive science and appropriates many insights from recent theories of concepts. However, it also harks back to more traditional accounts that sought to blur the boundary between conception and perception. When brought up to date, such accounts show tremendous promise. We can move forward by glancing backward and embracing the idea that concepts have a perceptual basis.

Notes

Chapter 1

1. I use small capital letters to designate concepts.

2. "To token" is a piece of philosophical jargon that means to produce a token of. A token is an instance. The AARDVARK concept that enters into my current thought that aardvarks are nocturnal is a token of the same type as the AARDVARK concept that entered into my thought last week that aardvarks live in burrows. Other cognitive scientists sometimes use the verb "to activate" where philosophers use "to token." In more ordinary language, my AARDVARK concept is tokened when I am thinking of aardvarks or experiencing something that brings aardvarks to mind.

3. See, for comparison, the "nonnegotiable conditions" presented by Fodor (1998).

4. It may be more accurate to say the concepts refer to properties, e.g., the property of being a frog or "froghood." The set of frogs is unified by the fact that all frogs have this property. The set of things we take to be frogs excludes some frogs and includes some things that are not frogs.

5. It is often also assumed that concepts must be syntactically compositional. The vehicles of compound concepts must contain the vehicles of their constituent concepts as real parts. This has been the subject of controversy. Some connectionists have argued that one can achieve the kind of compositionality necessary to explain productivity and systematicity without syntactic compositionality in this sense (e.g., van Gelder 1990). I do not want to get embroiled in this debate.

6. By "language" and "linguistic" I will be referring to public languages, the languages that we speak, rather than a special languagelike code used only in thought. Fodor defines a language of thought as a system of mental representation with a combinatorial syntax and semantics. Compositionality, at least in its semantic form, is already a desideratum on a theory of concepts, so it would not be surprising to find that concepts must be languagelike in this minimal sense.