2

A New Framework for the Theory of Moral Cognition

> To search in our common knowledge for the concepts which do not rest upon particular experience and yet occur in all knowledge from experience, of which they as it were constitute the mere form of connection, presupposes neither greater reflection nor deeper insight than to detect in a language the rules of the actual use of words generally and thus to collect elements for a grammar (in fact both researches are very nearly related), even though we are not able to give a reason why each language has just this and no other formal constitution, and still less why any precise number of such formal determinations in general, neither more nor less, can be found in it.
>
> – Immanuel Kant, *Prolegomena to Any Future Metaphysics*

In Chapter 1, I referred to aspects of Universal Grammar to which the theory of moral cognition might be usefully compared. In this chapter, I provide an initial statement of some of these comparisons and indicate which of them I take Rawls to have drawn and his critics to have misunderstood. In order to do so, it will be helpful to introduce and explain some technical terminology from Chomsky's framework, as well as some novel terminology of my own. The bulk of the chapter is therefore devoted to establishing a broad analytical framework for the theory of moral cognition and to clarifying certain philosophical issues that arise within this framework. At the end of the chapter, I draw on this conceptual scheme to provide a road map for the remainder of the book.

Three clarifications are worth making at the outset. First, throughout this chapter and the book as a whole, I often use phrases such as "generative linguistics," "linguistic theory," and "Chomsky's framework" as if they were indistinguishable. Obviously this is not the case: It is perfectly possible to be a linguist – indeed, a great linguist – and to disagree with Chomsky's particular theories of human language or its proper mode of inquiry. It is important to clarify, therefore, that all such references to linguistics in this book unless otherwise indicated are meant to refer *only* to the theoretical framework of

Universal Grammar and to those researchers working more or less within Chomsky's basic paradigm. Second, the following remarks are largely informal in nature. I make no sustained effort to defend the initial comparisons that I make in this chapter, merely to state them clearly, so as to clarify how I will be using certain terminology and to prepare the way for the more detailed discussions of the linguistic analogy that will occur in subsequent chapters. Finally, the comparisons identified in this chapter do not exhaust the interesting or relevant parallels between moral theory and linguistics, nor should they be taken to deny the existence of many important differences between these fields or their subject matters. They are merely some key initial comparisons that I have chosen to emphasize here, in an effort to focus attention on the linguistic analogy and to begin to draw out some of its potential implications for cognitive science, jurisprudence, and moral theory.[1]

2.1 NINE COMPARISONS BETWEEN LINGUISTICS AND MORAL THEORY

2.1.1 The Main Questions

Chomsky's (1986a: 3, 1991a: 6) approach to the study of language is organized around three main questions:

(1) (a) What constitutes knowledge of language?
(b) How is knowledge of language acquired?
(c) How is knowledge of language put to use?

In Chomsky's framework, the answer to (1a) is given by a particular *generative grammar* (or theory of *linguistic competence*): a theory of the steady state of the mind/brain of a person who "knows" or "cognizes" a particular natural language like English, Hebrew, Arabic, or Japanese. The answer to (1b) is given by *Universal Grammar* (UG), a theory of the initial state of the language faculty, assumed to be a distinct subsystem of the mind/brain devoted to language acquisition, along with an account of how the properties UG postulates interact with experience to yield knowledge of a particular language.[2] The answer

[1] I am grateful to Noam Chomsky for many illuminating conversations that have helped me to develop the framework presented in this chapter and the book as a whole. For helpful general introductions to the theory of Universal Grammar, see, e.g., Baker (2001), Cook & Newson (1996), Isac & Reiss (2008), Jackendoff (1994), and Pinker (1994). For more technical studies, see, e.g., Chomsky (1965, 1986, 1995) and Haegeman (1994).

[2] The terms *initial state* and *steady state* have technical meanings in theoretical linguistics that may be unfamiliar. Chomsky explains these terms in the following passage, which also usefully summarizes the research program of Universal Grammar as a whole:

> What many linguists call "universal grammar" may be regarded as a theory of innate mechanisms, an underlying biological matrix that provides a framework within which the growth of language proceeds. There is no reason for the linguist to refrain from imputing

to (1c) is, or would be, given by a theory of *linguistic performance:* a theory of how knowledge of language enters into the actual expression and interpretation of language specimens, as well as into interpersonal communication and other actual uses of language (Chomsky 1965: 4, 1988b: 3–4, 1991a: 6).

As I will attempt to show, the theory of moral cognition is usefully organized around three questions, close analogues to the fundamental questions in Chomsky's framework.

(2) (a) What constitutes moral knowledge?
(b) How is moral knowledge acquired?
(c) How is moral knowledge put to use?

An answer to (2a) would be given by a particular *generative moral grammar* (or theory of *moral competence*): a theory of the steady or acquired state of the mind/brain of a person who possesses a system of moral knowledge, or what one might refer to informally as a "sense of justice," "moral sense," "moral faculty," or "conscience." The answer to (2b) would be given by *Universal Moral Grammar* (UMG): a theory of the initial state of the moral faculty, assumed to be a distinct subsystem of the mind/brain, along with an account of how the properties UMG postulates interact with experience to yield a mature system of moral knowledge. The answer to (2c), if available, would be given by a theory of *moral performance:* a theory of how moral knowledge enters into the actual representation and evaluation of human acts and institutions and other forms of actual behavior.

2.1.2 The General Answers

Chomsky's general answer to (1a) is that a speaker's knowledge of language consists, in part, in her possession of a *grammar:* a complex system of unconscious principles or rules (1980: 51). His general answer to (1b) is that the system is acquired through the unfolding of a specific genetic program, under

> existence to this initial apparatus of the mind as well. Proposed principles of universal grammar may be regarded as an abstract partial specification of the genetic program that enables the child to interpret certain events as linguistic experience and to construct a system of rules and principles on the basis of that experience.
>
> To put the matter in somewhat different but essentially equivalent terms, we may suppose that there is a fixed, genetically determined initial state of the mind, common to the species with at most minor variations apart from pathology. The mind passes through a sequence of states under the boundary conditions set by experience, achieving finally a "steady state" at a relatively fixed age, a state that then changes only in marginal ways. The basic property of this initial state is that, given experience, it develops to the steady state. Correspondingly, the initial state of the mind might be regarded as a function, characteristic of the species, that maps experience into the steady state. Universal Grammar is a partial characterization of this function, of this initial state. The grammar of a language that has grown in the mind is a partial characterization of the steady state attained. (1980: 187–188)

the modest triggering and shaping effects of the environment (1980: 31). In Chomsky's framework, (1c) has two aspects: a *production problem* and a *perception problem*. The former is the problem of how people succeed in acting appropriately and creatively in linguistic behavior and performance. Chomsky's opinion is that serious investigation of this topic lies beyond the bounds of present-day science. He further conjectures that it may even be beyond the bounds of human intelligence in principle, having the status of a permanent mystery for creatures such as ourselves. The reason is that this "creative aspect of language use" (1966: 3–31, 1986b: 519) involves the exercise of free will and voluntary choice, which science is at least presently unable to explain. In any event, Chomsky expressly excludes the production half of (1c) from the set of fundamental questions he takes his theory of language to be addressing. Hence, his general answer to the production half of (1c) is that normal use of language consists in rule-governed yet nondeterministic behavior: in everyday speech production, language users exploit their knowledge of language in the freely chosen construction of linguistic expressions (1986b: 519, 1991b: 40).

The perception problem, also known as the *parsing problem,* is concerned with how the speaker-hearer is able to recognize the properties of form and meaning of linguistic expressions that she encounters. Chomsky's general answer to this question is that when a person is presented an expression in a particular situation, her rule-system assigns it a structural description that in some manner specifies those properties (1991a: 18).

Since (2a)–(2c) are empirical questions, of which there is little scientific understanding at present, whatever general comments one makes about them must be tentative. Nevertheless, for reasons that will become more apparent as we proceed, it seems reasonable to suppose in the case of (2a) that the normal individual's moral knowledge consists in part in her possession of what I will call a *moral grammar:* a complex and largely unconscious system of moral rules, concepts, and principles that generates and relates mental representations of various types. Among other things, this system enables individuals to determine the deontic status of a potentially infinite number and variety of acts and omissions. In the case of (2b), it seems reasonable to assume that this moral grammar is acquired through the unfolding of a specific genetic program, under the relatively modest triggering and shaping effects of the environment. Turning to (2c), it seems plausible to hold that a similar division between the production and perception components of the problem is a useful method of clarifying how moral knowledge is put to use. A solution to the *production problem* in the theory of moral performance would seek to determine how individuals succeed in applying their moral knowledge in their actual, day-to-day conduct. Here it seems possible that Chomsky is correct that, as far as science is concerned, this problem may turn out to be a permanent mystery for creatures like ourselves, since it involves concepts such as free will and voluntary choice that science is so far unable to explain. Given the largely spontaneous and involuntary nature of many human moral intuitions, however, the *perception problem* appears more tractable. Here the question is how individuals are able to recognize the moral properties of the acts and institutional arrangements they encounter. The general answer to this problem, at least in the case of human actions, appears to be substantially the same in case of moral perception as it is in the case of language or vision: when a person encounters or imagines a particular action, performed under a particular set of circumstances, her rule-system assigns it a structural description that in some manner specifies those properties.

2.1.3 The Fundamental Arguments

As we have seen, Chomsky's answer to (1a) largely equates an individual's knowledge of language with her possession of a particular mental grammar. His answer to (1b) holds that this grammar is acquired through the unfolding of a genetic program under the relatively modest triggering and shaping effects of the environment. These answers are supported by two fundamental arguments. Both have the logical form of what is often called an "abductive" argument (Peirce 1955/1901: 150–156) or an "inductive inference to the best explanation" (Harman 1965). The first is what I will refer to here as *the argument for linguistic grammar.* The second is what I will refer to here as *the argument from the poverty of the linguistic stimulus.* In Ray Jackendoff's (1994: 6) useful formulation, the former argument holds that the best explanation of "the expressive variety of language use" is the assumption that "the language user's brain contains a set of unconscious grammatical principles." The latter argument holds that the best explanation of how children acquire these principles is the assumption that "the human brain contains a genetically determined specialization for language."

The general answers to questions (2a) and (2b) may be defended by two similar arguments. The first is what I will refer to here as *the argument for moral grammar.* In its most general form, it holds that the best explanation of the properties of moral judgment is the assumption that the mind/brain contains a moral grammar. The second is what I will refer to here as *argument from the poverty of the moral stimulus.* In its most general form, it holds that the best explanation of how children acquire this grammar is the assumption that at least some of its core attributes are innate, in Descartes' and Chomsky's dispositional sense (see, e.g., Chomsky 1972: 173; Descartes 1985/1647: 303–305). Put differently, the argument from the poverty of the moral stimulus holds that at least part of the best explanation of how children acquire their unconscious moral knowledge is the assumption that the human genetic program includes instructions for the acquisition of a moral sense.

2.1.4 The Competence–Performance Distinction

In Chomsky's framework, the competence–performance distinction is a technical distinction that refers, in the first instance, to the difference between

(1a) and (1c). These two questions, although interrelated, must be sharply distinguished. (1c) asks a question about a speaker's actual behavior; (1a) represents a question about the knowledge or cognitive system that a speaker's observable behavior presupposes. The competence–performance distinction is, in the first instance, simply the difference between these two questions. A theory of linguistic competence provides an answer to (1a). A theory of performance seeks to provide an answer to (1c).

A less formal way to represent the difference between competence and performance is as the distinction between knowledge and behavior, or between what a person knows and what she does. Linguistic competence, in Chomsky's framework, denotes a person's *knowledge* of her language; linguistic performance refers to how, in actual situations, her knowledge of language gets put to *use.* It is uncontroversial that a speaker's actual linguistic behavior is affected by things other than her underlying competence: her memory structure, mode of organizing experience, perceptual mechanisms and attention span, and a wide range of additional factors. Linguistic competence is thus presupposed by, but is only one factor contributing to, actual performance or language use (Chomsky 1965: 3, 1980: 225).[3]

In my view, the competence–performance distinction is a useful means to distinguish different aspects of the research program illustrated by (2), as well to clarify different aspects of Rawls' linguistic analogy and how that analogy was received by Rawls' critics. In this book, therefore, I adopt this distinction and utilize it in the following manner. I will use the term *moral competence* to refer to an individual's moral knowledge and the term *moral performance* to refer to how that knowledge is put to use. I will say that a theory of moral competence provides an answer to (2a), and that a theory of moral performance, if available, would provide an answer to (2c).

It is important to grasp the significance of the competence–performance distinction. To a certain extent it reflects a division between

[3] Chomsky provides a helpful explanation of the competence–performance distinction in the following passage:

> The person who has acquired knowledge of a language has internalized a system of rules that relate sound and meaning in a particular way. The linguist constructing a grammar of a language is in effect proposing a hypothesis concerning this internalized system. The linguist's hypothesis, if presented with sufficient explicitness and precision, will have certain empirical consequences with regard to the form of utterances and their interpretations by the native speaker. Evidently, knowledge of language – the internalized system of rules – is only one of the many factors that determine how an utterance will be used or understood in a particular situation. The linguist who is trying to discover what constitutes knowledge of a language – to construct a correct grammar – is studying one fundamental factor that is involved in performance, but not the only one. This idealization must be kept in mind when one is considering the problem of confirmation of grammars on the basis of empirical evidence. There is no reason why one should not also study the interaction of several factors involved in complex mental acts and underlying actual performance, but such a study is not likely to proceed very far unless the separate factors are themselves fairly well understood. (1972: 26–27)

two fundamentally different approaches to the study of human nature: *mentalism* and *behaviorism*. Unlike psychological behaviorism, which at least in theory seeks to avoid all references to unobservable mental entities or processes (see, e.g., Skinner 1953; Watson 1925), generative linguistics is mentalistic in the technical sense: it represents a shift of focus from observable behavior to the cognitive structures of the mind entering into behavior (Chomsky 1965: 4, 1986a: 3). As Chomsky observes, the true novelty of his own theoretical approach is its mentalism: its shift of focus from performance to competence, from the study of language regarded from a purely behavioristic point of view to "the study of the system of knowledge of language attained and internally represented in the mind/brain" (1986a: 24). A grammar within a mentalistic framework is not a set of theoretical statements that purportedly describes observable behavior in some fashion; rather, it seeks to describe "exactly what one knows when one knows a language" (1986a: 24). Hence, by distinguishing (2a) and (2c), what I am attempting to do, in effect, is to shift (or begin to shift) the focus of moral theorists away from behaviorism and toward mentalism. In my judgment, the theory of moral cognition has not yet fully recovered from the damage that was done to it by behaviorism. Part of what I hope to accomplish in this book is to begin to repair this damage, by showing how a mentalistic approach to moral cognition might work.

2.1.5 The Distinction between Operative and Express Principles

Linguists do not assume that the normal language user is aware of the system of rules or principles that constitute her knowledge of language, or that she can become aware of them through introspection, or that her statements about them are necessarily accurate. On the contrary, as a result of empirical investigation, the rules of her language are assumed to lie beyond actual and even potential consciousness; moreover, it is taken for granted that her verbal reports and beliefs about her linguistic competence may be in error.

This leads to the following important point: in Chomsky's framework, a linguistic theory attempts to specify the *actual* properties of the speaker's linguistic competence, not what she may or may not report about them. The position linguists adopt in this regard is similar to the position adopted by theorists of visual perception, whose goal is to account for what a person actually sees and the cognitive mechanisms that determine what she sees, rather than to account for her own statements and explanations of what she sees and why (Chomsky 1965: 8).

To mark this difference, I will refer to it as the distinction between *operative* principles and *express* principles. In my view the same distinction should play an important role in moral theory. In this book I will attempt to capture it in the following terms. I will say that a person's *operative* moral principles are those principles that are actually operative in her exercise of moral judgment – the actual principles, in other words, of her moral competence. I will say that a person's *express* moral principles are those statements that a person verbalizes in the attempt to describe, explain, or justify her judgments. From this perspective moral theorists should not assume that the normal individual is aware of the operative principles that constitute her moral knowledge, or that she can become aware of them through introspection, or that her statements about them are necessarily accurate. On the contrary, they should be open to the possibility of discovering that just as normal persons are unaware of the principles guiding their linguistic or visual intuitions, so too are they often unaware of the principles guiding their moral intuitions. In any event, the important point is that, as with language or vision, the theory of moral cognition must attempt to specify what the properties of moral competence actually are, not what a person may report about them.

The distinction between operative and express principles is a traditional philosophical distinction. One finds reasonably clear expressions of it in the writings of many philosophers, including Leibniz (1981/1705), Hutcheson (1971/1728), Rousseau (1979/1762), Kant (1964/1785), Whewell (1845), Bradley (1962/1876), and Brentano (1969/1889), to name a few.[4] Hutcheson's manner of expressing the distinction in the opening pages of his *Illustrations on the Moral Sense* (1728) is especially elegant and perspicuous. Hutcheson writes:

> Let this also still be remembered, that the natural dispositions of mankind may operate regularly in those who never reflected upon them nor formed just notions about them. Many are really virtuous who cannot explain what virtue is. Some act a most generous disinterested part in life who have been taught to account for all their actions by self-love as the sole spring. There have been very different and opposite opinions in optics, contrary accounts have been given of hearing, voluntary motion, digestion, and other natural actions. But the powers themselves in reality perform their several operations with sufficient constancy and uniformity in persons of good health whatever their opinions be about them. In the same manner our moral actions and affections may be in good order when our opinions are quite wrong about them. True opinions, however, about both, may enable us to improve our natural powers and to rectify accidental disorders incident unto them. And true speculations on these subjects must certainly be attended with as much pleasure as any other parts of human knowledge. (1971/1728: 106)

In this passage Hutcheson reminds his reader that the point of a theory of moral cognition is to describe the "regular operations" of the moral sense, not to determine whether a person "has formed just notions" about it, or can "explain what virtue is" (cf. Reid 1969/1785: 726). Despite its evident concern with being scientific, modern moral psychology has largely ignored Hutcheson's sensible warning not to conflate these distinctions. Indeed, under the apparent influence of behaviorism, the twentieth century's leading

[4] The crux of the distinction is familiar to lawyers in the somewhat different but nonetheless related contrast between *ratio decidendi* ("reason for deciding") and *obiter dictum* ("something said in passing") (see generally Mikhail 2002a).

moral psychologists, Piaget (1965/1932) and Kohlberg (1981, 1984), and their followers have concerned themselves primarily with charting the development of an individual's ability to express an articulate opinion about moral problems, and to give a coherent account of her own moral intuitions, rather than with providing a theory of the moral sense itself. While these abilities are important, they should not be confused with the primary subject matter of the theory of moral cognition. As is the case with a theory of language or a theory of vision, the theory of moral cognition must attempt to describe the operative principles of moral competence, not what an experimental subject may or may not report about them.

The specific relevance of the distinction between operative and express principles for our topic is that, unlike Piaget and Kohlberg, Rawls is careful not to conflate this distinction. On the contrary, he takes its significance fully into account. Indeed, this is one of the main reasons why I believe his conception of moral theory is superior to theirs and should be of special interest to cognitive scientists. A central element of Rawls' linguistic analogy is his recognition that the subject matter of generative linguistics is linguistic competence, not linguistic performance, and that a correct account of linguistic competence requires theoretical constructions that go well beyond anything a nonlinguist can formulate for herself. In *A Theory of Justice,* Rawls suggests a similar situation presumably holds in moral philosophy. Like Hutcheson, therefore, he warns us not to assume that a *theory* of her sense of justice is something that the normal individual can express on her own (see, e.g., Rawls 1971: 47, 491). Rawls' view of this matter appears sound. The distinction between operative and express principles, as it is defined here,[5] is simply taken for granted in the study of language, vision, musical cognition, face recognition, and other cognitive domains. There seems little reason to suppose that the situation should be any different for a cognitive capacity as complex as the moral sense.

2.1.6 Levels of Empirical Adequacy

The terms *observational adequacy, descriptive adequacy,* and *explanatory adequacy* were introduced into linguistics and cognitive science by Chomsky in the early 1960s. They have specific meanings in his framework that are somewhat different from their ordinary connotations. For our purposes these may be rendered as follows. A theory of language is *observationally adequate* with respect to the data of an observed corpus of utterances if it correctly describes that data in some manner or other, for example, by listing them. A

[5] The terms *operative principles* and *express principles* originate with the nineteenth-century philosopher, William Whewell (1845). I have adapted them to the present context, however, and am using them somewhat differently then he does. For a useful introduction to Whewell's moral philosophy, see Schneewind (1977), especially pp. 101–117.

linguistic theory is *descriptively adequate* with respect to a particular individual's system of linguistic knowledge to the extent that it correctly describes that system in its mature or steady state. A linguistic theory meets what Chomsky terms the condition of *explanatory adequacy* to the extent that it correctly describes the initial state of the language faculty and correctly explains how the properties of the initial state interact with the child's primary data to yield mature knowledge of language (Chomsky 1964: 28–29; 1965: 24–27; see also Haegeman 1994: 6–11).

To clarify, the distinction between descriptive and explanatory adequacy corresponds to the difference between (1a) and (1b). Hence, in Chomsky's framework, descriptively adequate means the same thing as "provides a correct description of linguistic competence" or "provides a correct answer to (1a)." Explanatorily adequate means the same thing as "provides a correct explanation of linguistic competence" or "provides a correct answer to (1b)."

Chomsky's distinction between the problems of descriptive and explanatory adequacy is potentially confusing because correct answers to *both* problems are *both* descriptive and explanatory in the usual sense. A solution to the problem of descriptive adequacy – that is, a correct answer to (1a) – is a *description* of the mature speaker-hearer's linguistic competence; at the same time it is an *explanation* of the speaker's linguistic intuitions. Likewise, a solution to the problem of explanatory adequacy – that is, a correct answer to (1b) – is a *description* of the initial state of the language faculty; at the same time it is an *explanation* both of the speaker-hearer's acquired competence and (at a deeper level) those same intuitions.[6]

[6] Chomsky explains the distinction between descriptive and explanatory adequacy in the following passage. As I explain in Chapter 3, Chomsky and other linguists often use the term "grammar" with an acknowledged systematic ambiguity, to refer both to the linguist's theoretical description of the speaker-hearer's knowledge of language and to that knowledge itself. To help the reader who may be unfamiliar with this practice, I have inserted what I take to be the type of grammar Chomsky has in mind (theoretical or mental) in brackets.

> In a good sense, the [theoretical] grammar proposed by the linguist is an explanatory theory; it suggests an explanation for the fact that (under the idealization mentioned) a speaker of the language in question will perceive, interpret, form, or use an utterance in certain ways and not in other ways. One can also search for explanatory theories of a deeper sort. The native speaker has acquired a [mental] grammar on the basis of very restricted and degenerate evidence; the [mental] grammar has empirical consequences that extend far beyond the evidence. At one level, the phenomena with which the [theoretical] grammar deals are explained by the rules of the [mental] grammar itself and the interaction of these rules. At a deeper level, these same phenomena are explained by the principles that determine the selection of the [mental] grammar on the basis of the restricted and degenerate evidence available to the person who has acquired knowledge of the language, who has constructed for himself this particular [mental] grammar. The principles that determine the form of [mental] grammar and that select a [mental] grammar of the appropriate form on the basis of certain data constitute a subject that might, following traditional usage, be termed "universal grammar." The study of universal grammar, so understood, is a study of the nature of human intellectual capacities. It tries to formulate the necessary and sufficient conditions that a system must meet to

Despite these potential misunderstandings, I believe that the concepts of observational, descriptive, and explanatory adequacy are a helpful means by which to refer to different aspects of the research program illustrated by (2). They are also a useful method of identifying and clarifying different aspects of Rawls' linguistic analogy and the arguments of Rawls' critics. Throughout this book, therefore, I will adopt these terms and utilize them in the following way. I will say that a moral theory is *observationally adequate* with respect to a set of moral judgments to the extent that it provides a correct description of those judgments in some manner or other, for example, by listing them. I will say that a moral theory is *descriptively adequate* with respect to the mature individual's system of moral knowledge to the extent that it correctly describes that system – in other words, to the extent that it provides a correct answer to (2a). Finally, I will say that a moral theory meets the condition of *explanatory adequacy* to the extent it correctly describes the initial state of the moral faculty and correctly explains how the properties of the initial state it postulates interact with experience to yield a mature system of moral competence – in other words, to the extent that it provides a correct answer to (2b).

2.1.7 Two Additional Questions

(1a)–(1c) do not exhaust the basic questions concerning human language about which scientists would like to achieve theoretical insight and understanding. In particular, Chomsky (1995b) identifies two additional questions:

(1) (d) How is knowledge of language physically realized in the brain?
(e) How did knowledge of language evolve in the species?

Although (1d) and (1e) are the focus of much ongoing research, Chomsky (1995b: 17) has cautioned that they might be "beyond serious inquiry for the time being," much like many other far-reaching topics in the cognitive sciences. With respect to (1e), for example, although he is a leading proponent of a naturalistic approach to the study of language (the so-called biolinguistic perspective), Chomsky has been critical of those researchers, such as Pinker & Bloom (1990), who argue that the evolution of the human language faculty is best explained by Darwinian natural selection alone. Rather, along with other commentators such as Gould & Lewontin (1979), Lewontin (1990), and Darwin (1958/1859) himself, Chomsky has argued that a better explanation probably rests in some combination of selectional and nonselectional factors,

qualify as a potential human language, conditions that are not accidentally true of existing human languages, but that are rather rooted in the human "language capacity," and thus constitute the innate organization that determines what counts as linguistic experience and what knowledge of language arises on the basis of this experience. Universal grammar, then, constitutes an explanatory theory of a much deeper sort than particular [theoretical] grammar, although the particular [theoretical] grammar of a language can also be regarded as an explanatory theory. (1972: 27) including various ecological and historical contingencies and the space of possible options afforded by physical laws (see, e.g., Chomsky 2000, 2002; see also McGilvray 2005).[7]

In a similar fashion, one can formulate the following two questions in the theory of moral cognition, in addition to the three questions already identified:

(2) (d) How is moral knowledge physically realized in the brain?
(e) How did moral knowledge evolve in the species?

Many researchers began to investigate (2d) and (2e) (or broadly similar questions) from a naturalistic perspective during the last few decades of the twentieth century, including Blair (1995), Damasio (1994), and Damasio et al. (1994) in the case of (2d) and Alexander (1987), De Waal (1996), Trivers (1971), and Wilson (1975) in the case of (2e). More recently there has been an explosion of interdisciplinary research on these and related topics that shows no signs of abating (see, e.g., Sinnott-Armstrong 2008). Despite my keen interest in these issues, I will mostly put them aside in what follows, since within the framework I have articulated thus far any attempt to answer (2d) or (2e) would be premature. Just as posing well-framed and well-motivated questions about the neurophysiological and phylogenetic properties of human language largely depends on an adequate grasp of plausible answers to (1a)–(1c), so too, in my judgment, does the proper formulation of (2d) and (2e) largely depend on plausible answers to (2a)–(2c). Put differently, we cannot seriously ask how moral knowledge is realized in the brain or how it evolved in the species until what constitutes moral knowledge and how it is acquired and put to use by each individual are better understood.

2.1.8 Commonsense and Technical Concepts of Language and Morality

Chomsky draws a fundamental distinction between ordinary, commonsense, or pretheoretical concepts like language, knowledge of language, or linguistic competence, on the one hand, and various artificial or technical elaborations of those concepts, on the other. He suggests that all serious scientific approaches to the study of language must eventually replace the former concepts with technical substitutes, since the latter are more suitable for empirical study.

[7] As Chomsky (2000: 63) observes, "Darwin firmly denied that he attributed 'the modification of species exclusively to natural selection,' emphasizing in the last edition of *Origin of Species* that 'in the first edition of this work, and subsequently, I placed in a most conspicuous position – namely, at the close of the Introduction – the following words: "I am convinced that natural selection has been the main but not the exclusive means of modification." This has been of no avail. Great is the power of steady misrepresentation.'"

In Chomsky's case the technical concept most recently adopted (which simultaneously replaces both "language" and "knowledge of language," in their ordinary sense, and "linguistic competence," as this term was used by linguists during the two decades following publication of Chomsky's *Aspects of a Theory of Syntax*) is the concept *I-language.* A technical concept of language represents an instance of I-language, in Chomsky's view, if it characterizes a language as "some element of the mind of the person who knows the language, acquired by the learner, and used by the speaker-hearer" (1986a: 22). I-language thus refers to the language system in the mind/brain of a person who, in the ordinary sense, "knows" a language like English or Japanese. The "I" in I-language signifies at least three properties the I-language has: it is *internalized* in the sense that it is internal to the mind/brain. It is *intensional* in the sense that it may be regarded as a specific characterization of a function considered in intension that assigns a status to a range of events. It is *individualized* in the sense that the standpoint adopted toward I-language is that of individual psychology: I-languages are construed as something individual persons have, and different persons may be thought of as having in some sense different I-languages. A fourth reason the "I" in I-language is significant is that I-language represents an *idealization,* in several important respects (Chomsky 1986a: 3, 1986b: 513).

The general relevance for moral theory of Chomsky's distinction between commonsense and technical concepts of language emerges when we consider how a philosopher or scientist might go about replacing the more or less obscure, intuition-bound concepts ("sense of justice," "moral sense," "moral knowledge," "morality," "conscience," "moral competence," and so forth) in which her theoretical questions are initially posed with an artificially constructed technical concept, more suitable for scientific study.[8] An *I-morality* approach to moral theory would characterize its primary object of inquiry as some element of the mind/brain of a person who, we might ordinarily say, possesses a sense of justice ("moral sense," "moral knowledge," etc.). I-morality would thus refer, in its most neutral sense, to the moral system of the human mind/brain. The "I" in I-morality would signify at least four properties the system has: it is *internalized* in the sense that it is internal to the mind/brain. It is *intensional* in the sense that a theory of I-morality may be regarded as the characterization of a function considered in intension, which assigns a status to a range of events. It is *individualized* in the sense that the standpoint adopted toward I-morality is that of individual psychology: I-moralities are construed as something individual persons have, and different persons

[8] For instructive discussions of the conceptual difficulties in formulating a coherent account of the ordinary or intuitive concept of morality, see, e.g., Edel (1970: 285f.), Frankena (1976: 125–132, 168–183), Pareto (1935: 231f.), and Perry (1954: 86f.). As Perry aptly observes, "there is something which goes on in the world to which it is appropriate to give the name of 'morality'. Nothing is more familiar; nothing is more obscure in its meaning" (1954: 86).

may be thought of as having in some sense different I-moralities. Finally, I-morality represents an *idealization* in a number of important respects: it is a constructed model of a given biological object – the moral faculty of the human mind/brain.

The more specific relevance of I-morality for our topic is that while Rawls does not use this terminology in *A Theory of Justice,* he conceives of the subject matter of moral theory in precisely this way: that is, as a mental capacity that is internal, intensional, individual, and ideal, in the sense described. In my opinion, Rawls' remarks leave little doubt on this matter. Nonetheless, what appears to me equally true, and what I will argue below, is that many of Rawls' early critics appear to have misinterpreted him on just this point.

2.1.9 Theoretical Goals

Chomsky is often credited with asking more ambitious questions about the structure of language than his predecessors. While this is true, it is important to realize that, in another sense, his theoretical goals were more modest than theirs, and that it was the greater modesty of his goals that enabled him to ask more ambitious questions. In *Syntactic Structures,* Chomsky adapted a set of concepts from mathematical logic and distinguished three approaches to the metalinguistic problem of justifying grammars. He noted that the strongest requirement that could be asked of a linguistic theory is to provide what he called a *discovery procedure* for grammars: that is, a practical and mechanical method by which the linguist could actually construct the grammar, given a particular body of data (for example, a corpus of grammatical and ungrammatical utterances). Chomsky identified a weaker requirement of linguistic theory to be to provide a *decision procedure* for grammars: that is, a mechanical method for determining whether or not a proposed grammar is the *best* grammar of the language from which a given corpus is drawn. Finally, Chomsky identified an even weaker requirement of a linguistic theory to be to provide an *evaluation procedure* for grammars: that is, a mechanical method for determining which of two proposed grammars, G_1 and G_2, is the *better* grammar of the language from which a given corpus is drawn (1957: 49–60).

In *Syntactic Structures,* Chomsky adopted the position that it was unreasonable to expect a linguistic theory to provide anything more than an evaluation procedure for grammars. He thus adopted the weakest of the three positions described above. The notion that the overriding goal of linguistic theory is to evaluate which of two or more competing grammars is the better explanation of a given body of data has been a cornerstone of theoretical linguistics since that time.

In my view, the theory of moral cognition should follow Chomsky in distinguishing clearly and explicitly among possible theoretical goals. The strongest requirement that could be placed on a theory of moral cognition is that it provide what we might think of as a *discovery procedure* for moral principles: that is, a practical and mechanical method for actually constructing a correct set of moral principles on the basis of a given body of data (for example, a set of ordered pairs consisting of a moral judgment and a description of the circumstances that occasion it). A weaker requirement would be to demand that the theory provide a *decision procedure* for moral principles: that is, a mechanical method for determining whether or not a proposed set of moral principles is correct, or valid, or the best set of principles, with respect to a given class of judgments. Finally, a still weaker condition would be to require that the theory provide an *evaluation procedure* for moral principles: that is, a rational method of determining which of two or more proposed sets of principles is the better alternative, given a particular body of data.

The immediate significance of these distinctions for our topic emerges when we consider the evolution of Rawls' conception of metaethics from 1950 to 1971. In his early writings, Rawls explicitly rejects the notion that the appropriate goal of moral theory is to provide a discovery procedure for moral principles. In *Outline,* for example, he writes: "There is no way of knowing ahead of time how to find and formulate these reasonable principles. Indeed, we cannot even be certain that they exist, and it is well known that there are no mechanical methods of discovery" (1951a: 178). Hence Rawls' stated aim in *Outline* (as the title of that paper implies) is to construct a "decision procedure" for "validating or invalidating given or proposed moral rules," and he conceives of "the objectivity or the subjectivity of moral knowledge" to turn on whether such a procedure exists (1951a: 177). By 1971, when *A Theory of Justice* was published, Rawls' theoretical goals had apparently changed. His apparent objective in that book is to satisfy the weaker requirement of providing an evaluation procedure for moral principles, in particular to determine whether justice as fairness is a *better* theory of the sense of justice (I-morality) than utilitarianism (see, e.g., Rawls 1971: vii–viii, 17–18, 49–50, 52–53, 581). In my opinion, this is the most coherent way to interpret Rawls' metaethical commitments in *A Theory of Justice.* Nonetheless, as I will argue in Part Three, some of Rawls' early critics appear to have misunderstood this fundamental point.

2.2 PRELIMINARY CLARIFICATIONS ABOUT RAWLS' LINGUISTIC ANALOGY

Thus far I have identified the following questions in the theory of moral cognition:

(2) (a) What constitutes moral knowledge?
    (b) How is moral knowledge acquired?
    (c) How is moral knowledge put to use?
    (d) How is moral knowledge physically realized in the brain?
    (e) How did moral knowledge evolve in the species?

These five questions are simply the outline of a research program, all bones and no flesh. One of the primary aims of this book is to begin to sharpen these questions and to fill in some of the relevant details. The main advantage of introducing these problems and an appropriate terminology for pursuing them at this early stage of our investigation is that it affords the chance to make some initial clarifying remarks about the scope and limits of both Rawls' linguistic analogy and the research based on it that I seek to undertake here. It also enables me to summarize the remaining chapters of this book in a more perspicuous form.

What I am calling Rawls' linguistic analogy is actually a series of comparisons that Rawls makes in Section 9 of *A Theory of Justice* (1971: 46–53) between moral theory and parts of theoretical linguistics, as expressed primarily in the first few pages of Chomsky's *Aspects of the Theory of Syntax* (1965: 3–9). In my opinion Rawls' remarks have been widely misunderstood, due primarily to the failure of his readers to take seriously the distinctions that Rawls draws, implicitly in Section 9 and explicitly elsewhere (e.g., Rawls 1950, 1951a, 1975), between at least five general kinds of question that moral philosophy seeks to answer. The first are *empirical* questions about the *mature* or *steady* state of I-morality – roughly those questions corresponding to (2a). The second are *empirical* questions about the *initial* or *original* state of I-morality – namely, those questions corresponding to (2b). The third are *empirical* questions about how I-morality is put to use – that is, those questions corresponding to (2c). The fourth are *normative* questions about moral principles, in particular the question of which moral principles are justified. Finally, the fifth are *metaethical* questions about particular moral judgments and moral principles, such as the question *whether,* and if so *how,* they may be said on rational grounds to be justified.[9]

Rawls' fourth and fifth questions are not represented by (2a)–(2e). Instead, in Rawls' framework they correspond more closely to (2f) and (2g):

(2) (f) Which moral principles are justified?
    (g) How can moral principles be justified?

In what follows I will refer to (2a)–(2g) as *descriptive, explanatory, behavioral, neurocognitive, evolutionary, normative,* and *metaethical* questions, respectively (see Table 2.1). Following Chomsky, I will continue to refer to (2a) as *the problem of descriptive adequacy* and to (2b) as *the problem of explanatory adequacy.* Because it will be convenient to have analogous phrases to refer to (2f) and (2g), I will refer to them as *the problem of normative adequacy* and *the problem of metaethical adequacy,* respectively. Likewise, I will refer to

[9] All five questions must be distinguished from *practical* questions of political philosophy, such as the question that motivates *Political Liberalism:* How can a stable and just society of free and equal citizens live harmoniously when deeply divided by a plurality of incompatible and irreconcilable, though reasonable, comprehensive doctrines?

TABLE 2.1. *Seven Main Problems in the Theory of Moral Cognition*

No.	Problem	Theoretical Goal
(2a)	What constitutes moral knowledge?	Descriptive adequacy
(2b)	How is moral knowledge acquired?	Explanatory adequacy
(2c)	How is moral knowledge put to use?	Behavioral adequacy
(2d)	How is moral knowledge physically realized in the brain?	Neurocognitive adequacy
(2e)	How did moral knowledge evolve in the species?	Evolutionary adequacy
(2f)	Which moral principles are justified?	Normative adequacy
(2g)	How can moral principles be justified?	Metaethical adequacy

(2c)–(2e) as *the problem of behavioral adequacy, the problem of neurocognitive adequacy,* and *the problem of evolutionary adequacy,* respectively.[10]

Now, as I interpret it, Rawls' linguistic analogy is centered primarily on the comparison between the problem of descriptive adequacy in linguistics and the problem of descriptive adequacy in ethics, and, to a lesser extent, between the problem of explanatory adequacy in linguistics and the corresponding problem in ethics. The primary grounds on which the analogy is criticized, however, is its failure to solve (or to play a part in solving) the problem of *normative* adequacy. Therefore, the central issue between Rawls and his critics turns out to be, not the value of the linguistic analogy for moral theory, as one might initially assume, but the manner in which Rawls conceives the overall structure of an ethical theory, and in particular the relationship between its empirical and normative branches.[11]

In his early writings, Rawls' approach to moral theory appears to rest on two main assumptions about the relationship between the problems of descriptive

[10] In relying on the concepts of descriptive and explanatory adequacy, I largely follow Chomsky (1964, 1965). The remaining terms are constructed by analogy. Note that by defining the problem of metaethical adequacy to be how moral principles can be justified, Rawls departs from other familiar conceptions of metaethics, such as those conceptions that define metaethics to be the linguistic analysis of ethical concepts, or those conceptions that are concerned with the epistemological, metaphysical, or ontological status of moral truths or moral facts. I return to this topic in Chapter 6.

[11] Descriptive ethics, normative ethics, practical ethics, and metaethics are often held to be distinct fields of inquiry. However, what their exact boundaries are, and how much each draws from the others, are not entirely clear, or at least have never been clearly and convincingly stated. Hare's (1960: 100) observation that "no generally accepted terminology for making the necessary distinctions has yet emerged" arguably remains true today. For a series of recent taxonomies, see generally Lamont (1946), Rawls (1950, 1975), Hare (1952, 1960, 1963, 1973, 1981), Mandelbaum (1955), Nowell-Smith (1954), Brandt (1959), Stevenson (1963), Frankena (1963), Findlay (1970), Harman (1977), Scanlon (1982, 1992), Broad (1985), and Brink (1989).

and normative adequacy. First, Rawls assumes that there is an *order of priority* between these two problems, according to which the descriptive takes precedence over the normative. Second, Rawls assumes that a descriptively adequate moral theory constitutes a *presumptive* solution to the problem of normative adequacy, given the nature of the evidence that a descriptively adequate theory explains (see, e.g., Rawls 1951a: 182–184, 186–188; 1971: 46–53).

As I interpret him, Rawls employs the linguistic analogy primarily to clarify certain aspects of the problem of descriptive adequacy. In particular, Rawls draws on the theory of language to accomplish at least eight distinct but overlapping objectives. The first two objectives are (i) to formulate the problem of descriptive adequacy by (ii) defending the empirical assumption that the sense of justice is a cognitive system of sufficient coherence and complexity to make it interesting and worthwhile to describe. Rawls defends this assumption by means of what I refer to in this book as *the argument for moral grammar*. This argument may be usefully thought of as the moral analogue to the *argument for linguistic grammar*. In my terms – not Rawls' – the argument for moral grammar holds that the best explanation of the observable properties of moral judgment is the assumption that the mind/brain contains a moral grammar, or set of unconscious moral rules or principles.

As Norman Daniels (1979: 258, 1980: 22–23) observes, the argument for moral grammar can be restated in somewhat different but essentially equivalent terms as the claim that, like linguistic theory, moral theory is faced with a *projection problem*, that is, the problem of explaining how ordinary individuals are capable of applying their moral knowledge to new and often unprecedented cases. I explain this terminology at more length in Chapter 3 (see Section 3.1.1). For now, what is important to recognize is simply that the argument for moral grammar (or alternatively the recognition that moral theory is faced with a projection problem) constitutes the first feature of what I refer to in this book as Rawls' linguistic analogy.

As I interpret him, the six remaining elements of Rawls' linguistic analogy are meant primarily to clarify what the problem of descriptive adequacy involves. Like Chomsky, Rawls distinguishes between (iii) descriptive adequacy and observational adequacy, (iv) operative principles and express principles, (v) descriptive adequacy and explanatory adequacy, and (vi) moral competence and moral performance. He also draws on the framework of theoretical linguistics to explain (vii) the theory-dependence of the competence–performance distinction and (viii) the importance of idealization in addressing the problem of moral diversity and in solving problems of empirical adequacy more generally.[12]

[12] In this book for ease of expression I will sometimes use the term *empirical adequacy* in place of the phrase "descriptive and explanatory adequacy." In other words, an empirically adequate grammar in my usage refers to a grammar that satisfies both descriptive and explanatory adequacy, in the sense defined in the text. Empirical adequacy in its broadest sense also includes behavioral, neurocognitive, and evolutionary adequacy, of course, but my focus here will be descriptive and explanatory adequacy.

Thus far I have said little about (2f) and (2g), the problems of normative and metaethical adequacy. In *A Theory of Justice,* Rawls advances a straightforward answer to (2f). At least within the context of his main contractual argument for principles of social justice, the principles he purports to justify are those he calls *the special conception of justice.*[13] In the case of (2g), however, the situation is more complex. As I interpret him, Rawls presupposes a complicated answer to the general problem of justifying moral principles, which turns on at least three potentially unrelated ideas: first, that moral principles can be presumptively justified by showing that they are a solution to the problem of descriptive adequacy; second, that descriptively adequate moral principles can be further justified by showing that they are part of a solution to the problem of explanatory adequacy; and third, that moral principles that meet the demands of descriptive and explanatory adequacy can be justified to an even greater extent by showing that the adoption of such principles can be proven as a formal theorem in the theory of rational choice.[14] As I understand it, Rawls' notion of *reflective equilibrium* is intended to suggest that these three apparently disparate ideas can, in fact, be reconciled. In other words, Rawls assumes as a general matter that the same set of moral principles can be part of a single, comprehensive solution to the problems of descriptive, explanatory, and normative adequacy simultaneously.

Needless to say, this idea of Rawls' has been the source of considerable interest and debate. His perceived use of reflective equilibrium as a method for justifying moral principles has drawn sharp criticism from some philosophers and has been resourcefully defended by others. Rawls himself compares aspects of reflective equilibrium both with Nelson Goodman's (1983/1955) influential account of the justification of principles of deductive and inductive inference and with Chomsky's (1965) account of descriptive and explanatory adequacy (Rawls 1971: 18–22, 46–53; see also 120, 491). These comparisons have contributed to further uncertainty over whether Rawls conceives of moral theory as an empirical discipline – a branch of psychology – and, if so, whether its pretensions to normativity can be maintained.

[13] As I discuss in more detail in Chapter 6, the conception of social justice Rawls defends in *A Theory of Justice* comes in two forms, one general and one more specific. The two principles of the more specific conception – what Rawls calls the "special conception of justice" – are the following.

First Principle

Each person is to have an equal right to the most extensive total system of basic liberties compatible with a similar system of liberty for all (1971: 250).

Second Principle

Social and economic inequalities are to be arranged so that they are both (a) to the greatest benefit of the least advantaged and (b) attached to offices and positions open to all under conditions of fair equality of opportunity (1971: 83).

[14] For a helpful discussion of the rational choice element in *A Theory of Justice,* see generally Wolff (1977).

In my opinion, once the theory-dependence of the competence–performance distinction and the appropriately modest goal of moral philosophy to provide an evaluation procedure for moral principles are taken into account and given their due weight, Rawls' notion that the same set of moral principles can be part of a solution to (2a), (2b), and (2f) seems plausible. But this statement must not be interpreted too simply. As Rawls observes, reflective equilibrium is a name given to "the philosophical ideal" (1971: 50). It refers to the state of affairs that is reached once the theorist has discovered the principles to which her set of considered judgments conform and, in turn, the premises of those principles' derivation (1971: 20). As with any other science, the conclusions of moral theory are always provisional and hence may be modified as a result of further investigation. This emphasis on the provisional or presumptive character of moral theory is important. It implies that whether a moral theory is correct, and thus whether a given set of moral principles is justified, is always an open question. Nevertheless, the question of justification is largely settled at any given time, so far as it can be, by showing that a particular set of moral principles satisfies these three tests better than any available alternative. In this manner the theorist attempts to show that the principles that do in fact describe and explain human moral competence are also *rational,* in the sense that free and equal persons would choose to adopt them to govern their relations with one another, if they were given that choice.[15]

The fact that one of the three elements of the problem of metaethical adequacy is the requirement that justifiable principles be shown to be rational might seem like a sharp disanalogy with linguistics. Although I believe that this disanalogy is real, I would again caution against interpreting it too simply. It is true that linguists do not ask whether grammars are rational in the sense that Rawls has in mind. That is, they do not ask whether the principles of an empirically adequate grammar would also be chosen in a suitably characterized contractual situation. However, it is important to recognize that linguists do attempt to show that grammars of particular languages can be viewed, in a sense, as theorems derivable from principles of higher generality (i.e., Universal Grammar). Further, it is important to recall that, in *A Theory of Justice,* Rawls considers the original position to be not merely a procedure for proving the rationality of principles of justice, but also a model that explains how the sense of justice is acquired (see, e.g., 1971: 47, 120, 491). Moreover, Rawls evidently believes that by pursuing the problem of empirical adequacy,

[15] As Rawls makes clear in a passage of his dissertation that anticipates *A Theory of Justice*, his contract argument is hypothetical, not historical. Thus, he does not claim "that at any time in the past the rules of common sense have been explicitly discussed and voluntarily adopted in light of the principles, and the nature of man and society. This historical question is not to the point. But it is relevant to say that if men had explicitly discussed the adoption of common sense rules with the principles in mind, then they *would* have adopted those rules which in general they have. That is what I mean when I say that common sense is, in general, justifiable" (Rawls 1950: 107).

the problems of normative and metaethical adequacy may be transformed. This might help to explain his observation that "if we can find an accurate account of our moral conceptions, then questions of meaning and justification may prove much easier to answer. Indeed, some of them may no longer be real questions at all" (1971: 51). Rawls does not specify the conditions under which justifying moral principles may prove to be easier, or may no longer be a real question, but we can readily imagine what he has in mind. By solving the problem of descriptive adequacy, moral theorists may be led to frame, and then to solve, the problem of explanatory adequacy, thereby demonstrating that the morality of common sense is rooted in human nature, as many philosophers, jurists, and cognitive scientists have often assumed.

Having said this, it is important to emphasize that there are, and presumably always will be, many aspects of the problem of justification in ethics that have no clear analogue in linguistics. Even if we have described and explained the acquisition of moral competence, and even if we have discovered its evolutionary origin and its physical signatures in the brain, there will remain many questions that we might wish to ask that have no clear counterparts in the theory of grammar. For example, we may wish to ask whether the moral principles that have evolved in the species (assuming such principles exist) are compatible with the institutional requirements of modern life, and if not, whether our institutions should be changed. We might also wish to ask whether acting out of a sense of justice is conducive to a person's good or whether, as Nietzsche apparently thought, doing so is likely to be a psychological disaster for that person (cf. Scanlon 1982: 218). Above all, we will want to know whether the principles of moral competence are compatible with the requirements of rationality, in whatever sense of rationality seems appropriate. These are complex topics, which, except for a few peripheral remarks, I will not discuss here. Because I agree with Rawls about the importance of not giving way to the impulse to answer questions that one is not yet equipped to examine (Rawls 1975: 10), and because any serious discussion of these topics would be premature within the framework that I have described thus far, I must set them aside for now. I mention these issues only to forestall possible misinterpretations of the nature and scope of this inquiry. The point of this book is not to deny that there are significant differences between moral theory and linguistics. On the contrary, I not only accept, but insist, that these two disciplines, and the theoretical problems they seek to solve, are dissimilar in many fundamental respects.

2.3 OUTLINE OF REMAINING CHAPTERS

The remaining chapters of this book are intended to make a unified whole by supporting each other in the following way. Chapter 3 consists of an interpretation of Section 9 of *A Theory of Justice*. The main aim of this chapter is to call attention to the various features of Rawls' linguistic analogy described above. The conception of moral theory outlined thus far invites at least two kinds of questions. The first is exegetical: Is this conception correctly attributed to Rawls? The second is philosophical: Is this conception sound? Although these two inquiries are logically independent, they are not unrelated to one another. Rawls is among the clearest and most careful thinkers in the field. He has a deep knowledge of his subject matter as well as that of adjacent disciplines. Hence whatever theoretical positions he adopts are likely to be worthy of serious consideration (compare Singer's similar remarks about Sidgwick in Singer 1974). If it can be shown that the conception of moral theory described in this chapter is not merely a restatement of Chomsky's philosophy of linguistics in moral terms, but a restatement of Rawls' conception of moral theory in Chomsky's terms – or, rather, in terms that are intelligible to the community of researchers working more or less within the framework of Universal Grammar – then, insofar one believes, as I do, that Universal Grammar is on the right track, one's conviction that the conception is sound should be strengthened. To say this is not to make a mere appeal, or to rely uncritically, on either Rawls' or Chomsky's authority. Rather, it is to recognize that, like any other science, a theory of moral cognition is more likely to be sound insofar as it builds on the insights of its most successful predecessors, and insofar as it rests on assumptions that are shared by, or at least consistent with, what is known in adjacent fields. Hence, the primary aim of Chapter 3 is to show that the conception of moral theory Rawls describes in Section 9 can be accurately reformulated in the terminology we have adopted and, in fact, that it is even more compelling when restated within this general framework. To do this I first identify and explain the eight basic elements of Rawls' linguistic analogy described above. I then argue that Rawls' conception of moral theory is an I-morality conception in the sense defined in Section 2.1.8.

Let me now turn to a summary of Part Two. Here it helps to begin with some comments of a more general nature. Thus far I have referred to Rawls' conception of moral theory in exclusively favorable terms. I have suggested that Rawls was one of the first philosophers to grasp the potential implications of Universal Grammar for ethics, that *A Theory of Justice* contains one of the most powerful accounts of moral theory ever written, and that Rawls' early work contains the outline of an empirical theory of moral cognition that far surpasses the research programs of Piaget and Kohlberg in terms of depth, coherence, and scientific rigor. Remarks like these may lead the reader to assume that I believe that Rawls' conception of moral theory is somehow beyond criticism, or at least correct in its essentials. Any such assumption, however, would be inaccurate.

In my view, Rawls' early approach to moral theory suffers from at least three major shortcomings, at least from the perspective of someone like myself who wishes to draw on Rawls' early ideas to develop a fruitful research program in the cognitive science of moral judgment. The first shortcoming concerns how Rawls understands the relationship between the moral principles that apply to institutional arrangements and those that apply to the actions of individuals. According to Rawls, the first set of principles is primary, and the second set is derivative (1971: 108–110; cf. Hart 1973; Miller 1974). From a naturalistic point of view, this seems implausible. It seems more likely that the principles that generate considered moral judgments about the basic structure of society are themselves rooted in principles that apply to the acts of conspecifics, rather than the other way around.

The second weakness of Rawls' moral philosophy from a naturalistic perspective concerns his approach to moral psychology and cognitive development. In his early writings Rawls assumes that there are two major traditions in the theory of cognitive development, one stemming from classical empiricism and represented more recently by behaviorism and social learning theory, and the other deriving from classical rationalism and represented more recently by the social constructivism of Piaget and Kohlberg. Rawls places the utilitarians from Hume through Sidgwick, along with Freud, in the former category, and Rousseau, Kant, Humboldt, and J. S. Mill (at least in some of his moods) in the latter (1971: 458–461). While these groupings are conventional, I believe they are inadequate. To the extent that such classifications are helpful (which may be questioned), one must distinguish at least three traditions, not two. In addition to behaviorism and social constructivism, there is a genuine third alternative: the so-called *nativism* of Chomsky, Fodor, Spelke, and others, which is currently one of the major research paradigms in cognitive science and the philosophy of mind.

As this concept will be understood here, nativism is simply the view that early cognitive development is best understood, from a scientific point of view, as a process of biologically determined growth and maturation. As such, nativism assumes that while the acquisition of cognitive systems is triggered and shaped by appropriate experience, and thus depends crucially on cultural inputs, the specific properties of those systems are largely predetermined by the innate structure of the mind. So understood, nativism arguably has a stronger claim than constructivism does to having descended from the classical rationalism of philosophers such as Leibniz, Rousseau, Kant, and Humboldt. I will not take up this historical issue here (for some useful discussion, see, e.g., Chomsky 1966; Piatelli-Palmarini 1980; Spelke 1998). For our purposes what is important is simply to recognize that when Rawls evaluates different accounts of cognitive development in *A Theory of Justice,* he does not adequately consider nativism. I believe that the reasons for this apparent oversight are not too difficult to discern. We must keep in mind when *A Theory of Justice* was published. In 1971, neither academic philosophy nor academic psychology was particularly hospitable to genetic or biologically based accounts of moral or social development, or indeed of cognitive development more generally. Even in the case of language, Chomsky's "innateness hypothesis," as Putnam (1975) called it, was still highly controversial at the time. Moreover, the field of cognitive science as we think of it today was still in its infancy. For example, the first textbook in cognitive psychology did not appear until the late 1960s (Neisser 1967), and many leading journals had not yet begun publication. In general, during the period in which Rawls was writing *A Theory of Justice,* there simply did not yet exist a body of empirical literature or an established theoretical vocabulary on which Rawls could have relied to defend the claim that moral knowledge is innate, even if he had wished to do so.[16]

What I take to be the third inadequacy of Rawls' early approach to moral theory is his failure to take his own metaethical requirements seriously enough. In his early writings, Rawls sets admirably clear and ambitious computational goals for moral theory. For example, he frequently makes provocative observations such as the following: "We should strive for a kind of moral geometry with all the rigor which this name connotes" (1971: 121). Rawls also formulates reasonably clear and, in my view, compelling arguments to explain why moral philosophers interested in normative issues should largely postpone

[16] Indeed, even to find a phrase like "moral knowledge is innate" in recent literature is difficult. To do so, one must return to the Enlightenment. For example, Leibniz observes:

> [M]oral knowledge is innate in just the same way that arithmetic is, for it too depends on demonstrations provided by the inner light. Since demonstrations do not spring into view straight away, it is no great wonder if men are not always aware straight away of everything they have within them, and are not very quick to read the characters of the natural law which, according to St. Paul, God has engraved in their minds. However, since morality is more important than arithmetic, God has given to man instincts which lead, straight away and without reasoning, to part of what reason commands. (Leibniz 1996/1705: 94)

Likewise, Rousseu writes:

> Cast your eyes on all the nations of the world, go through all the histories. Among so many inhuman and bizarre cults, among this prodigious diversity of morals and characters, you will find everywhere the same ideas of justice and decency, everywhere the same notions of good and bad. … There is in the depths of all souls, then, an innate principle of justice and virtue according to which, in spite of our own maxims, we judge our actions and those of others as good or bad. It is to this principle that I give the name *conscience.* (Rousseau 1979/1762: 288–289)

To clarify, I am not suggesting that Rawls actually wanted to argue that moral knowledge is innate but was prevented from doing so because he lacked an appropriate theoretical vocabulary. On the contrary, I believe that Rawls did *not* wish to make this type of argument. This is because his overriding aim was to promote consensus about principles of social justice, and presumably this would have been more difficult if he had also made strong claims about the naturalistic origins of these principles. As it was, Rawls' book generated stiff opposition from commentators who were uncomfortable with his occasional appeals to human nature.

These remarks are not entirely speculative: in my conversations with him, Rawls explained to me that he intended his theory to be consistent with, but not to require, assumptions about the innateness of the sense of justice because he wished to refrain from advancing a more controversial argument when a weaker one might achieve the same ends (cf. Rawls 1971: 495–496).

them until they have achieved the more modest goal of descriptive adequacy. Rawls' actual pursuit of these objectives, however, departs from the approach that he recommends to others. Thus, taking these points in reverse order, beginning in 1957 and continuing throughout the development of his account of justice as fairness, Rawls does not first show that his two principles of justice are descriptively adequate and proceed only afterward to the problem of justification. Rather, he takes their descriptive adequacy as more or less given, and he constructs the original position in order to justify and explain them.[17] Moreover, as many early commentators (e.g., Care 1969; Wolff 1977) observe, and as Rawls (e.g., 1971: 121, 581) readily concedes, Rawls does not actually attempt to prove any "theorems of moral geometry" in *A Theory of Justice* (1971: 126). Rather, his general mode of argument throughout the book is highly intuitive, a feature that draws the ire of many of his sharpest critics (see, e.g., Hare 1973).

These remarks are not necessarily meant to be criticisms of Rawls. On the contrary, in light of his primary objectives in *A Theory of Justice* – to explicate and justify a comprehensive theory of social justice and to apply it to the basic institutional structure of a well-ordered society – Rawls' strategic choices and intuitive mode of argument seem highly appropriate, and indeed to some extent inevitable. In the first place, it does not seem reasonable to suppose that, for complex problems of social justice, there exists a finite yet complete system of principles that, together with the nonmoral facts, is capable of mechanically settling all possible disputes that might arise to the satisfaction of all relevant observers. Hence, the notion of a computational theory of social justice in this sense seems implausible (cf. Pound 1908). In addition, one must consider the vast ground that Rawls wished to cover. *A Theory of Justice* touches on a wealth of topics beyond those I have already mentioned. These range from constitutional liberties, justice between generations, and tax policy to civil disobedience and conscientious objection. The book also addresses the law of nations, natural duties and obligations, the theory of moral development, and even evolutionary stability. Rawls could not possibly have examined all of these topics in the manner he does while also proving "theorems of moral geometry" with "all the rigor which this name connotes" (1971: 126, 121). Hence, Rawls' actual mode of argument seems both appropriate and inevitable.

Nevertheless, one must recognize that these features of Rawls' work are real shortcomings from the point of view of the theory of moral cognition.

[17] See Rawls (1957), where Rawls introduces his two principles for the first time. He says: "*Given these principles,* one might try to derive them from a priori principles of reason, or offer them as known by intuition. These are familiar steps, and, at least in the case of the first principle, might be made with some success. I wish, however, to look at the principles in a different way" (1957: 655, emphasis added). In this essay and in his other early writings, Rawls does not show that his principles are descriptively adequate; rather, he takes their descriptive adequacy as more or less given, and he constructs an argument to justify and explain them.

Insofar as a philosopher or scientist seeks, as I do, to develop such a theory, and to integrate it into the cognitive and brain sciences, she must begin by focusing attention on much simpler problems than those that occupy Rawls in *A Theory of Justice.* In addition, she must start from empirically plausible assumptions about how the mind works. Finally, she must attempt to make her theory as analytically rigorous as possible.

The purpose of Part Two is to build on the conception of moral theory that Rawls outlines in Section 9 of *A Theory of Justice* while at the same time correcting for these perceived shortcomings. Specifically, I attempt to illustrate the worth of Rawls' linguistic analogy by showing how the conception of moral theory it presupposes can be transformed into a genuine empirical theory that takes the actions of individuals, rather than institutional arrangements, as its primary focus; that is broadly mentalist, modular, and nativist in its basic orientation; and that approaches the problem of descriptive adequacy in a manner consistent with the demanding computational requirements that Rawls articulates in his early writings.

In Chapter 4, I begin this process by taking a closer look at the family of trolley problems that originated with the work of Foot (1967) and Thomson (1986). In a series of experiments that began in the mid-1990s, my colleagues and I began testing these problems, and others like them based on the same basic template, on hundreds of individuals, both adults and children. Our central aim was to pursue the idea of a Universal Moral Grammar and to begin to investigate a variety of empirical questions that arise within this framework. Our basic prediction was that the moral intuitions elicited by at least some of these problems would be widely shared, irrespective of demographic variables such as race, sex, age, religion, national origin, or level of formal education. We also predicted that most individuals would be unaware of the operative principles generating their moral intuitions, and thus largely incapable of correctly describing their own thought processes. These predictions were confirmed, and our initial findings have now been replicated and extended with over 200,000 individuals from over 120 countries (see, e.g., Miller 2008; see generally Section 5.2.1).

After introducing the trolley problems and observing that the properties of the moral judgments they elicit appear to illustrate various aspects of the linguistic analogy, the remainder of Chapter 4 attempts to formulate the problem of descriptive adequacy with respect to these judgments, and to situate this problem within the framework of the contemporary cognitive sciences. Chapter 5 then sketches a provisional solution to this descriptive problem, which I label *the moral grammar hypothesis.* According to this hypothesis, a crucial feature of the trolley problems is that they suggest and can be used to prove that moral judgments do not depend solely on the superficial description of a given action, but also on how that action is *mentally represented,* a critical preliminary step in the evaluative process that jurists have frequently examined (see, e.g., Cardozo 1921; Hutcheson 1929; Oliphant 1928; Radin

1925; see also Grey 1983; Kelman 1981) but, surprisingly, many psychologists have unduly neglected. Hence the problem of descriptive adequacy in the moral domain must be divided into at least three parts, involving the description of (i) deontic rules, (ii) structural descriptions, and (iii) conversion rules. Although the difficulty that most people have in explaining or justifying their judgments implies that they are unaware of the principles that guide their moral intuitions, the judgments themselves can be explained by assuming that these individuals are intuitive lawyers, who possess tacit or unconscious knowledge of a rich variety of legal rules, concepts, and principles, along with a natural readiness to compute mental representations of human acts and omissions in legally cognizable terms. Put differently, the intuitive data can be explained by assuming that ordinary individuals implicitly recognize the relevance of categories like ends, means, side effects, and *prima facie* wrongs, such as battery, to the analysis of legal and moral problems. In particular, the key distinction that explains many of the standard cases in the literature is that the agent commits one or more batteries as a means of achieving his good end in the impermissible conditions (e.g., the Transplant and Footbridge problems), whereas these violations are merely subsequent and foreseen side effects in the permissible conditions (e.g., the Trolley and Bystander problems). Moreover, the structural descriptions that are implied by this explanation can be exhibited in a two-dimensional tree diagram, successive nodes of which bear a generation relation to one another that is asymmetric, irreflexive, and transitive (Goldman 1970a; see generally Mikhail 2000, 2002b, 2005; Mikhail, Sorrentino, & Spelke 1998).

In Chapter 6, I provide a more detailed and formal description of the mental operations implied by the moral grammar hypothesis. In particular, drawing on a diverse set of ideas and traditions, including deontic logic, lexical semantics, the philosophy of action, and the common law of crime and tort, I argue that the manner in which trolley problems are mentally represented can be described in terms of a hierarchical sequence of act-token representations, or *act tree* (Goldman 1970a; Mikhail 2000), which encodes the information relevant to determining a particular action's deontic status. On this basis I propose a novel computational analysis of trolley problem intuitions that appears capable of accounting, in explicit and rigorous fashion, for a broad range of these otherwise puzzling commonsense moral judgments. Finally, throughout Part Two I distinguish the moral grammar hypothesis from the alternative model of moral judgment advocated by researchers such as Joshua Greene and Jonathan Haidt (2002). Unlike Greene and Haidt, I argue that the critical issue in the theory of moral cognition is not whether moral intuitions are linked to emotions – clearly they are – but how to characterize the appraisal system those intuitions presuppose, and in particular whether that system incorporates elements of a sophisticated jurisprudence.

Let me turn next to a summary of Part Three. The main purpose of these chapters is to respond to what I take to be some rather unconvincing criticisms of Rawls' linguistic analogy that have not yet received adequate attention in the philosophical literature, in particular those of R. M. Hare, Peter Singer, Thomas Nagel, and Ronald Dworkin. In Chapter 7, I argue that philosophers such as Hare and Singer who have criticized Rawls' linguistic analogy on the grounds that the conception of moral theory it presupposes is too empirical or insufficiently normative appear to be operating with an unduly narrow and impoverished conception of moral philosophy. On the one hand, these critics apparently wish to exclude empirical questions about the nature and origin of commonsense moral knowledge from the domain of what they identify as moral philosophy. Yet, on the other hand, they appear to beg the very questions that Rawls' research program is designed to answer, inasmuch as they assume a broadly empiricist account of how moral knowledge is acquired. Moreover, neither Hare nor Singer appears to have grasped the relationship in Rawls' conception of moral theory between the problems of empirical and normative adequacy. In any event, they have failed to address Rawls' reasonable contention in *Outline* that a descriptively adequate moral theory constitutes a *presumptive* solution to the problem of normative adequacy, in light of the class of judgments that a descriptively adequate moral theory explains (Rawls 1951). Finally, I argue that their central objection to Rawls' method of explicating common moral intuitions, which I call their *objection from insufficient normativity,* fails to come to terms with Rawls' somewhat different, and more complex, approach to the relationship between empirical and normative adequacy in *A Theory of Justice,* as captured by his notion of reflective equilibrium.

In Chapter 8, I examine Thomas Nagel's brief objections to Rawls' linguistic analogy in his early review of *A Theory of Justice* (Nagel 1973). In particular, I use Nagel's objections as an opportunity to explore some of the implications of the competence–performance distinction for moral theory. I argue that Nagel's arguments are untenable as they stand and that, even if one charitably reconstructs them, they fail to constitute a compelling criticism of the conception of moral theory Rawls describes in *A Theory of Justice.* The main reason is that Nagel fails to acknowledge Rawls' reliance on the competence–performance distinction and to recognize the theory dependence of the corresponding distinction in linguistics. Once these points are properly understood, Nagel's criticisms do not seem persuasive. I defend a similar argument with respect to those philosophers, such as Norman Daniels (1979, 1980) and Richard Brandt (1979, 1990), who criticize Rawls by relying on what I call *the objection from prejudice,* according to which Rawls' conception of moral theory is flawed because it consists of a mere reshuffling of our prejudices. In response, I argue that this objection fails to acknowledge Rawls' legitimate and indeed indispensable use of the competence–performance distinction, in the form of his concept of a *considered judgment.* Rawls uses this concept to define the difference between a prejudice and a judgment in which moral capacities are likely to be displayed without distortion. Since, as a question of method, Rawls is entitled to draw this distinction and then seek empirical support for it within the framework of his theory, the objection from prejudice appears to be without force.

In Chapter 9, I examine Ronald Dworkin's discussion of Rawls' linguistic analogy in his influential book *Taking Rights Seriously.* Among other things, I contend that Dworkin's naturalist and constructivist interpretations of Rawls represent a false antithesis; neither is an accurate model of the conception of moral theory Rawls actually describes in *A Theory of Justice.* I argue that Dworkin's main error in this regard appears to be his mistaken assumption that a naturalistic approach to moral theory must be "realist" rather than "mentalist" (or what I describe in Chapters 3 and 8 as an "E-morality" conception of moral theory rather than an I-morality conception). In short, there is a credible, mentalistic alternative to Dworkin's untenable version of naturalism, which researchers interested in the idea of Universal Moral Grammar can seek to develop. Furthermore, this genuinely naturalistic alternative offers a sound basis on which to support and defend a robust conception of universal human rights.

In Part Four, I summarize the main points of the book and attempt to place Rawls' linguistic analogy within a broader historical and philosophical context. Finally, I conclude by arguing that the theory of moral cognition might be able to vindicate Rawls' guiding conviction that humankind possesses a shared moral nature if philosophers, linguists, cognitive scientists, and legal scholars would join forces and pursue the research program outlined in this book.