6

Moral Reasoning

GILBERT HARMAN, KELBY MASON, AND WALTER SINNOTT-ARMSTRONG

Jane: "Hi, Kate. Do you want to grab a quick bite? I'm tired, but I feel like eating something before I go to bed."
Kate: "I can't. I'm desperate. You know that big philosophy paper that's due tomorrow? I haven't even started it. I spent all evening talking to Eleanor about breaking up with Matt."
Jane: "Wow, that's too bad. My paper took me a long time. I had to write two versions. The first one wasn't any good, so I started all over."
Kate: "Really? Hmm. Did you show anybody the first one?"
Jane: "No."
Kate: "Was it really bad?"
Jane: "Not that bad, but I want to get an 'A'. It's my major, you know."
Kate: "Well, then, do you think you could email me your first paper? I could polish it up and hand it in. That would really save my life. I don't know how else I could finish the assignment in time. And I'm doing really bad in the course, so I can't afford to mess up this paper. Please."
Jane: "Well, uhh . . . . you've never asked me anything like that before. [Pause] No. I can't do that. Sorry."
Kate: "Why not? Nobody'll find out. You said you didn't show it to anybody. And I'm going to make changes. We are friends, aren't we? Please, please, please."
Jane: "Sorry, but that's cheating. I just can't do it. Good luck, though. I hope you write a good one."
Kate: "Thanks a lot."

In this simple example, Jane forms a moral judgment. Did she engage in moral reasoning? When? What form did it take? That depends partly on which processes get called "moral reasoning."

Jane started with an initial set of moral and non-moral beliefs that were or could become conscious. She ended up with a new set of such beliefs, including the moral belief that she morally ought not to send her paper to Kate, a belief that she had not even thought about before. In addition, Jane started with a set of intentions and added a new intention, namely, an intention to refuse to send her paper to Kate, which she had not formed before. In some contexts, forming new beliefs and intentions in this way is described as moral reasoning.

How did Jane form her new moral belief and new intention? At first, Jane just took in information about the situation, which she then analyzed in the light of her prior beliefs and intentions. Her analysis might have been unconscious in the way that her auditory system unconsciously broke down the verbal stimulus from Kate and then inferred the meaning of Kate's utterances. In this case her analysis may be described as an instance of unconscious moral reasoning, even if it appears to her as a direct *moral intuition* or emotional reaction that seems at a conscious level to involve no inference at all. (On moral intuitions, see Sinnott-Armstrong, Young, & Cushman, Chapter 7 in this volume.)

Later, when Kate asks, ''Why not?'', Jane responds, ''that's cheating. I just can't do it . . . '' This response suggests an argument: ''For me to email you my first paper would be cheating. Cheating is wrong, except in extreme circumstances. But this is not an extreme circumstance. Therefore, it would be wrong for me to email you my first paper. I can't do what I know is wrong. So I can't email you my first paper.'' Jane probably did not go through these steps explicitly before Kate asked her ''Why not?'' However, Jane still might have gone through some of these steps or some steps like these explicitly before she uttered the sentences that suggest this argument. And Jane did utter public sentences that seem to fit into some such form. Conscious thinking as well as public assertions of *arguments* like this are also often called moral reasoning.

Now suppose that Jane begins to doubt that she did the right thing, so she goes and asks her philosophy professor what counts as cheating, whether cheating is wrong, and why. Her professor might argue for a theory like rule utilitarianism, talk about the problems that would arise if cheating were openly and generally permitted, and infer that cheating is morally wrong. Her professor might then define cheating, apply the definition to Jane's case, and conclude that it would have been morally wrong for Jane to email her first paper to Kate. This kind of *reflective* argument may occur relatively rarely outside of settings like the philosophy classroom, but, when it does, it is often included under ''moral reasoning.''

Thus at least three kinds of processes might be called moral reasoning. First, unconscious information processing, possibly including emotions, can lead to a new moral judgment without any consciousness of any steps in any inference. Second, when faced with a moral issue, people might consciously go through steps by thinking of and endorsing thoughts whose contents fit into standard patterns of deductive arguments or inductive inferences. Third, when they have enough time, some people sometimes engage in more extensive reflection and infer moral conclusions, even surprising moral conclusions, from consciously articulated thoughts.

While theories of moral reasoning often address all or some of these kinds of moral reasoning, different theories may emphasize different kinds of reasoning, and theories also differ in their goals. Accounts of reasoning can be either *descriptive* psychological theories, attempting to characterize something about how people actually reason, or *normative* theories, attempting to say something about how people ought to reason or to characterize certain aspects of good or bad reasoning.

Most of this chapter will focus on one popular theory of moral reasoning, which claims that moral reasoning does and should fit the form of deductive arguments. We shall argue that such a deductive model is inadequate in several ways. Later we shall make some tentative suggestions about where to look for a better model. But first we need to clarify what moral reasoning is.

1. Kinds of Moral Reasoning

1.1. *Internal versus External Reasoning*

We distinguish *internal* or *intrapersonal* reasoning—reasoning something out by oneself, inference or personal deliberation—from *external* or *interpersonal* reasoning—bargaining, negotiation, argument, justification (to others), explanation (to others), and other sorts of reasoning done for or together with other people. The two types of reasoning often intersect; for instance, when two people argue, they will also be doing some internal reasoning as they go, deciding what to say next, whether their opponent's conclusions follow from their premises and so on (again, this internal reasoning might well be unconscious). In the other direction, we sometimes use external reasoning to help along our internal reasoning, as when talking through an issue with somebody else helps us see it in a new light. Nonetheless, the two types of reasoning are clearly different, involving different sorts of processes—various

mental operations in the one case, and various public acts in the other. In this chapter, we shall be concerned with moral reasoning of the internal kind.

1.2. *Theoretical versus Practical Reasoning*

It is useful to distinguish "theoretical" reasoning or inference from "practical" reasoning or deliberation in roughly the following way. Internal *practical* reasoning is reasoning that in the first instance is apt to modify one's decisions, plans, or intentions; internal *theoretical* reasoning is reasoning that in the first instance is apt to modify one's beliefs ("apt" because of limiting cases in which reasoning leaves matters as they are, without any effect on one's beliefs or intentions).

One way to distinguish the two types of reasoning is by looking at how their conclusions are expressed. The results of theoretical reasoning are typically expressed in declarative sentences, such as "Albert went to the late show at the Garden Theater." By contrast, the results of practical reasoning are typically expressed with imperatives, such as "Let's go to the late show at the Garden."

Much internal reasoning is a mixture of practical and theoretical reasoning. One reasons about what is the case in order to decide what to do; one's decision to do something can influence what one believes will happen. Moral reasoning can be either theoretical or practical, or a mixture. It is theoretical to the extent that the issue is (what to believe) about what someone morally ought to do, or what is morally good or bad, right or wrong, just or unjust. Moral reasoning is practical to the extent that the issue is what to do when moral considerations are or might be relevant.

What Practical and Theoretical Reasoning have in Common Internal reasoning of both sorts can be goal directed, conservative, and coherence seeking. It can be directed toward responding to particular questions; it can seek to make minimal changes in one's beliefs and decisions; and it can try to avoid inconsistency and other incoherence and attempt to make one's beliefs and intentions fit together better.

So, for example, one might seek to increase the positive coherence of one's moral views by finding acceptable moral principles that fit with one's opinions about particular cases and one might try to avoid accepting moral views that are in conflict with each other, given one's non-moral opinions. We discuss empirical research specifically directed to this model of internal reasoning later in this chapter.

How Internal Practical and Theoretical Reasoning Differ Despite the commonalities between them, and the difficulty of sharply demarcating them, there

are at least three important ways in which internal practical and theoretical reasoning differ, having to do with wishful thinking, arbitrary choices, and "direction of fit." First, the fact that one wants something to occur can provide a reason to decide to make it occur, but not a reason to believe it has occurred. Wishful thinking is to be pursued in internal practical reasoning—in the sense that we can let our desires determine what we decide to do—but avoided in internal theoretical reasoning—in the sense that we typically shouldn't let our desires guide what we believe.

Second, internal practical reasoning can and often must make arbitrary choices, where internal theoretical reasoning should not. Suppose there are several equally good routes to where Mary would like to go. It may well be rational for Mary arbitrarily to choose one route and follow it, and it may be irrational for her not to do so. By contrast, consider Bob, who is trying to determine which route Mary took. It would not be rational for Bob to choose one route arbitrarily and form the belief that Mary took that route instead of any of the others. It would not be irrational for Bob to suspend belief and form neither the belief that Mary took route *A* nor the belief that Mary took route *B*. Bob might be justified in believing that Mary took either route *A* or route *B* without being justified in believing that Mary took route *A* and without being justified in believing that Mary took route *B*, even though Mary is not justified in choosing to take either route *A* or route *B* unless she arbitrarily chooses which to take.

A third difference is somewhat difficult to express, but it has to do with something like the "direction of fit" (Austin, 1953). Internal theoretical reasoning is part of an attempt to fit what one accepts to how things are. Internal practical reasoning is an attempt to accept something that may affect how things are. Roughly speaking, theoretical reasoning is reasoning about how the world already is, and practical reasoning is reasoning about how, if at all, to change the world. Evidence that something is going to happen provides a theoretical reason to believe it will happen, not a practical reason to make it happen (Hampshire, 1959). This way of putting things is inexact, however, because changes in one's beliefs can lead to changes in one's plans. If Mary is intending to meet Bob at his house and then discovers that Bob is not going to be there, she should change her plans (Harman, 1976). Even so, the basic idea is clear enough.

Internal reasoning (theoretical or practical) typically leads to (relatively small) changes in one's propositional attitudes (beliefs, intentions, desires, etc.) by addition and subtraction. Of course, there are other ways to change one's propositional attitudes. One can forget things, or suffer from illness and injuries leading to more drastic changes. These processes don't appear to be instances of

reasoning. However, it is not easy to distinguish processes of internal reasoning from such other processes.

For one thing, there appear to be rational ways of acquiring beliefs as ''direct'' responses to sensation and perception. Are these instances of reasoning? The matter is complicated because unconscious ''computation'' may occur in such cases, and it is difficult to distinguish such computation from unconscious reasoning (as we discuss in Section 1.3). It is not clear whether these and certain other changes in belief should be classified as reasoning.

In any case, inferences are supposed to issue in conclusions. Ordinary talk of what has been ''inferred'' is normally talk of a new conclusion that is the result of inference, or perhaps an old conclusion whose continued acceptance is appropriately reinforced by one's reasoning. We do not normally refer to the discarding of a belief in terms of something ''inferred,'' unless the belief is discarded as a result of accepting its negation or denial. But there are cases in which reasoning results in ceasing to believe something previously believed, without believing its negation. In such a case it is somewhat awkward to describe the result of internal reasoning in terms of what has been ''inferred.'' Similarly, when reasoning leads one to discard something one previously accepted, it may be awkward to talk of the ''conclusion'' of the reasoning.

To be sure, it might be said that the ''conclusion'' of one's reasoning in this case is *to stop believing (or intending) X* or, maybe, *that one is (or ought) to stop believing or intending X*. And, although it is syntactically awkward to say that what is ''inferred'' in such a case is *to stop believing (or intending) X* (because it is syntactically awkward to say that *Jack inferred to stop believing (or intending) X*), it might be said that what is ''inferred'' is *that one is (or ought) to stop believing (or intending) X*.

These ways of talking might also be extended to internal reasoning that leads to the acceptance of new beliefs or intentions. It might be said that the ''conclusion'' of one's reasoning in such a case is *to believe (or decide to) Y* or *that one is (or ought) to believe (or decide to) Y* and that what one ''infers'' is *that one ought to believe (or decide to) Y*.

One of these ways of talking might seem to imply that all internal reasoning is practical reasoning, reasoning about what to do—*to stop believing (or intending) X* or *to believe (or decide to) Y* (cf. Levi, 1967). The other way of talking might seem to imply that all reasoning is theoretical reasoning, reasoning about what is the case—*it is the case that one is (or ought) to stop believing (or intending) X* or *it is the case that one ought to believe (or decide to) Y* (cf. Nagel, 1970).

Such reductive proposals have difficulty accounting for the differences between internal theoretical and practical reasoning. A reduction of internal theoretical reasoning to practical reasoning would seem to entail that

there is nothing wrong with arbitrary choice among equally good theoretical conclusions. A reduction of internal practical reasoning to internal theoretical reasoning simply denies that there is such a thing as internal practical reasoning, in the sense of reasoning that results in decisions to do things and otherwise potentially modifies one's plans and intentions. Since neither reduction seems plausible to us, we continue to suppose that internal theoretical and practical reasoning are different, if related, kinds of reasoning.

1.3. *Conscious and Unconscious Moral Reasoning*

Finally, internal reasoning may be conscious or unconscious. Although some accounts of moral judgment identify reasoning with conscious reasoning (Haidt, 2001), most psychological studies of reasoning have been concerned with unconscious aspects of reasoning. For example, there has been controversy about the extent to which reasoning about deduction makes use of deductive rules (Braine & O'Brien, 1998; Rips, 1994) as compared with mental models (Johnson-Laird & Byrne, 1991; Polk & Newell, 1995). All parties to this controversy routinely suppose that such reasoning is not completely conscious and that clever experiments are required in order to decide among these competing theories.

Similarly, recent studies (Holyoak & Simon, 1999; Simon et al., 2001, Simon, 2004; Thagard, 1989, 2000) investigate ways in which scientists or jurors reason in coming to accept theories or verdicts. These studies assume that the relevant process of reasoning (in contrast with its products) is not available to consciousness, so that evidence for theories of reasoning is necessarily indirect. (We discuss some of this literature below.)

Indeed, it is unclear that one is ever fully conscious of the *activity* of internal reasoning rather than of some of its intermediate and final *upshots*. Perhaps, as Lashley (1958) famously wrote, ''No activity of mind is ever conscious.'' To be sure, people are conscious of (aspects of) the *external* discussions or arguments in which they participate and they can consciously imagine participating in such discussions. But that is not to say that they are conscious of the internal processes that lead them to say what they say in those discussions. Similarly, people can consciously and silently count from 1 to 5 without being able to be conscious of the internal processes that produce this conscious sequence of numbers in the right order.

Since external arguments are expressed in words, imagined arguments will be imagined as expressed in words. This does not imply that internal reasoning is itself ever in words as opposed to being reasoning about something expressed in words (Ryle, 1979) and does not imply that internal reasoning is ever conscious.

When theorists refer to "conscious reasoning," they may be referring either to such externally expressed or imagined arguments or to other upshots of internal reasoning.

Finally, Gibbard (1990) and Scanlon (1998) argue that moral thinking is concerned with finding ways of acting that can or could be justified to others. In that case, internal moral reasoning might always involve thinking about external moral reasoning to others and so might always or typically involve conscious envisioning of external reasoning. But the internal reasoning processes about such external reasoning would not themselves normally be conscious.

1.4. *Section Summary*

To summarize this section: reasoning can be characterized along three dimensions, namely practical/theoretical; internal/external; and, when internal, conscious/unconscious. The sort of moral reasoning we shall discuss in this chapter is internal, and may be practical or theoretical, and conscious or unconscious.

2. The Deductive Model of Moral Reasoning

Assuming this understanding of what moral reasoning is, we can now ask how it works, what form it takes, and when it is good. One popular model of moral reasoning gives distinctive answers to all three questions. We call this *the deductive model* of moral reasoning.

On the deductive model, a person's reasoning might start from a very general moral rule or principle, such as the principle of utility, the categorical imperative, the doctrine of double effect, or some claim about abstract rights or social contracts. However, if the person is not a philosopher, the reasoning is usually thought to start with a mid-level moral principle, such as that it is wrong to kill, lie, steal, or hurt other people. In Jane's case, the argument might be something like this:

(P1) Cheating is always morally wrong except in extreme circumstances.
(P2) This act is cheating.
(P3) This circumstance is not extreme.
Therefore, (C) this act is morally wrong.

Background arguments might be added to support these premises. The moral reasoning is then seen as good only if it is deductively valid, only if the

premises are justified, and only if the argument commits no fallacies, such as equivocation or begging the question.

Proponents of the deductive model usually make or assume several claims about it:

(1) Deductive arguments make people justified in believing the conclusions of those arguments.
(2) A person's beliefs in the premises cause that person to believe the conclusion.
(3) The premises are independent, so the universal premise (P1) contains all of the moral content, and the other premises are morally neutral.
(4) The terms in the argument fit the so-called classical view of concepts (as defined by necessary and sufficient conditions for class inclusion).

Not everyone who defends the deductive model makes all of these claims, but they are all common. (1) is clearly a normative claim about the justification relation between the premises and conclusions of a deductive moral argument. By contrast, (2)–(4) are most naturally interpreted as descriptive claims about the way people actually morally reason. Someone could make them as normative claims—claiming that one's moral reasoning *ought* to conform to them—but we'll discuss them under the more natural descriptive interpretation.

The deductive model can be applied universally or less broadly. Some philosophers seem to claim that all moral reasoning (or at least all good moral reasoning) fits this model. Call this the *universal deduction claim*. Others suggest the weaker claim that much, but not all, moral reasoning fits this deductive model. Call that the *partial deduction claim*. Most philosophers do not commit themselves explicitly to either claim. Instead, they simply talk and write as if moral reasoning fits the deductive model without providing details or evidence for this assumption.

Although it has not often been explicitly stated, this deductive model has been highly influential throughout the history of philosophy. For example, Stich (1993) attributes something like this view (or at least the claim about concepts) to Plato's *Republic* and other dialogues. The Socratic search for necessary and sufficient conditions makes sense only if our moral views really are based on concepts with such conditions, as the classical view proposes. When Kant first introduces the categorical imperative, he says, ''common human reason does not think of it abstractly in such a universal form, but it always has it in view and uses it as the standard of its judgments'' (Kant, 1785: 403–4). To keep it ''in view'' is, presumably, to be conscious of it in some way, and Kant ''uses'' the categorical imperative in deductive arguments when he applies it to examples. Similarly, Mill attributes the deductive model to opponents of utilitarianism

who posit a natural "moral faculty": "our moral faculty . . . supplies us only with the general principles of moral judgments; it is a branch of our reason, not of our sensitive faculty; and must be looked to for the abstract doctrines of morality, not for perception of it in the concrete" (Mill, 1861/2001: 2). Our "moral faculty," on such a view, delivers to us moral principles from which we must deductively argue to the morality of specific cases.

In the twentieth century, Hare says, "the only inferences which take place in [moral reasoning] are deductive" (1963: 88) and "it is most important, in a verbal exposition of an argument about what to do, not to allow value-words in the minor premise" (1952: 57). Rule-utilitarians also often suggest that people make moral judgments about cases by deriving them from rules that are justified by the utility of using those rules in conscious moral reasoning that is deductive in form (cf. Hooker & Little, 2000: 76, on acceptance of rules). When Donagan analyzes common morality in the Hebrew–Christian tradition, he claims, "every derived precept is strictly deduced, by way of some specificatory premise, either from the fundamental principle or from some precept already derived" (1977: 71). He does add that the specificatory premises are established by "unformalized analytical reasoning" (ibid.: 72), but that does not undermine the general picture of a "simple deductive system" (ibid.: 71) behind common people's moral judgments. More recently, McKeever and Ridge argue that "moral thought and judgment presuppose the possibility of our having available to us a set of unhedged moral principles (which go from descriptive antecedents to moral consequents) which codifies all of morality available to us" (2006: 170). The principles are unhedged and their antecedents are descriptive so that moral conclusions can be deduced from the principles plus morally neutral premises, as the deductive model requires. Principles that enable such deductions are claimed to be an intrinsic goal of "our actual moral practice" (ibid.: 179).

Traditional psychologists also sometimes assume a deductive model. The most famous studies of moral reasoning were done by Lawrence Kohlberg (1981). Kohlberg's method was simple. He presented subjects (mainly children and adolescents) with dilemmas where morally relevant factors conflicted. In his most famous example, Heinz could save his dying wife only by stealing a drug. Kohlberg asked subjects what they or the characters would or should do in those dilemmas, and then he asked them why. Kohlberg found that children from many cultures typically move in order through three main levels, each including two main stages of moral belief and reasoning:

Level A: Preconventional
  Stage 1 = Punishment and Obedience
  Stage 2 = Individual Instrumental Purpose

Level B: Conventional
Stage 3 = Mutual Interpersonal Expectations and Conformity
Stage 4 = (Preserving) Social Order
Level C: Postconventional and Principled Level
Stage 5 = Prior Rights and Social Contract or Utility
Stage 6 = Universal Ethical Principles

Kohlberg's theory is impressive and important, but it faces many problems. First, his descriptions of his stages are often imprecise or even incoherent. Stage 5, for example, covers theories that some philosophers see as opposed, including utilitarianism and social contract theory. Second, Kohlberg's Stage 6 is questionable because his only "examples of Stage 6 come either from historical figures or from interviews with people who have extensive philosophic training" (Kohlberg, 1981: 100). Third, even at lower levels, few subjects gave responses completely within a single stage, although the percentage of responses within a single stage varied, and these variations formed patterns. Fourth, psychologists have questioned Kohlberg's evidence that all people move through these same stages. The most famous critique of this sort is by Gilligan (1982), who pointed out that women and girls scored lower in Kohlberg's hierarchy, so his findings suggest that women are somehow deficient in their moral reasoning. Instead, Gilligan claims, women and girls engage in a different kind of moral reasoning that Kohlberg's model misses. (For a recent, critical meta-analysis of studies of Gilligan's claims, see Jaffe & Hyde, 2000.)

Whether or not Kohlberg's theory is defensible, the main point here is that he distinguishes levels of moral reasoning in terms of principles that could be presented as premises in a deductive structure. People at Stage 1 are supposed to reason like this: "Acts like this are punished. I ought not to do what will get me punished. Therefore, I ought not to do this." People at Stage 2 are supposed to reason like this: "Acts like this will defeat or not serve my purposes. I ought to do only what will serve my purposes. Therefore, I ought not to do this." And so on. The last two stages clearly refer to principles that, along with facts, are supposed to entail moral judgments as conclusions in deductive arguments. This general approach, then, suggests that all moral reasoning, or at least the highest moral reasoning, fits some kind of deductive model.

However, Kohlberg's subjects almost never spelled out such deductive structures (and the only ones who did were trained into the deductive model). Instead, the deductive gloss is added by coders and commentators. Moreover, Kohlberg and his colleagues explicitly asked subjects for reasons. Responses to such questions do not show either that the subjects thought

of those reasons back when they made their judgments or that those reasons caused the judgments or that people use such reasons outside of such artificial circumstances.

More generally, it is not clear that Kohlberg's method can really show much, if anything, about internal moral reasoning. Even if people in a certain group tend to cite reasons of certain sorts when trying to explain and justify their moral beliefs, the context of being asked by an experimenter might distort their reports and even their self-understanding. Consequently, Kohlberg's method cannot show why people hold the moral beliefs they do.

Hence Kohlberg's research cannot support the universal deduction claim. It cannot even support a partial deduction claim about normal circumstances. The most that this method can reveal is that people, when prompted, come up with different kinds of reasons at different stages of development. That is interesting as a study in the forms of public rhetoric that people use at different points in their lives, but it shows nothing about the internal processes that led to their moral judgments.

As far as we know, there is no other empirical evidence that people always or often form moral judgments in the way suggested by the deductive model. Yet it seems obvious to many theorists that people form moral judgments this way. In fact, there is some evidence that children are sometimes capable of reasoning in accord with the deductive model (e.g. Cummins, 1996; Harris & Núñez, 1996). What these studies show is that, given an artificial rule, children are capable of identifying violations; and a natural explanation of their performance here is that they are reasoning in accord with the deductive model, i.e. using the rule as a premise to infer what the person in a scenario ought to be doing and concluding that the person is breaking the rule (or not). But this does not show that children normally accord with the deductive model when they are morally reasoning in real life, and not considering artificial rules contrived by an experimenter. *A fortiori*, this does not show that adults generally reason in accord with the deductive model when they reason morally. Thus there is so far little or no empirical evidence that the deductive model captures most real situated moral reasoning in adults.

Moreover, there are many reasons to doubt various aspects of the deductive model. We shall present four. First, the deductive model seems to conflate inferences with arguments. Second, experimental results show that the premises in the deductive model are not independent in the way deductivists usually suppose. Third, other experimental results suggest that moral beliefs are often not based on moral principles, even when they seem to be. Fourth, the deductive model depends on a classical view of concepts that is questionable. The next four sections will discuss these problems in turn.

3. Do Deductive Arguments Justify Conclusions?

Defenders of the deductive model often claim or assume that good arguments must conform to standards embodied in formal logic, probability theory, and decision theory. This claim depends on a confusion between formal theories of validity and accounts of good reasoning.

3.1. *Logic, Probability, and Decision Theory*

To explain this claim, we need to say something about the relevance of logic, probability, and decision theory to reasoning by oneself—inference and deliberation—and to reasoning with others—public discussion and argument. The issue is somewhat complicated, because the terms ''logic,'' ''theory of probability,'' and ''decision theory'' are used sometimes to refer to formal mathematical theories of implication and consistency, sometimes to refer to theories of method or methodologies, and sometimes to refer to a mixture of theories of these two sorts.

On the formal mathematical side, there is formal or mathematical logic, the mathematical theory of probability, and mathematical formulations of decision theory in terms of maximizing expected utility. An obvious point, but one that is often neglected, is that such formal theories are by themselves neither descriptive theories about what people do nor normative theories about what people ought to do. So they are not theories of reasoning in the sense in which we are here using the term ''reasoning.''

Although accounts of formal logic (e.g. Goldfarb, 2003) sometimes refer to ''valid arguments'' or examples of ''reasoning'' with steps that are supposed to be in accord with certain ''rules of inference,'' the terms ''reasoning'' and ''argument'' are then being used to refer to certain abstract structures of propositions and not for something that people do, not for any concrete process of inference or deliberation one engages in by oneself or any discussion among two or more people. The logical rules in question have neither a psychological, nor a social, nor a normative subject matter. They are rules of implication or rules that have to be satisfied for a structure to count as a valid formal argument, not rules of inference in the sense in which we are here using the term ''inference.''

The logical rule of *modus ponens* says that a conditional and its antecedent jointly imply the consequent of the conditional. The rule does not say that, if one believes or otherwise accepts the conditional and its antecedent, one must or may also believe or accept the consequent. The rule says nothing about

belief in the consequent, and nothing about what may or may not be asserted in an argument in our sense.

There may be corresponding principles about what people do or can or should rationally believe or assert, but such principles would go beyond anything in formal logic. Indeed, it is nontrivial to find corresponding principles that are at all plausible (Harman, 1986: ch. 2). It is certainly not the case that, whenever one believes a conditional and also believes its antecedent, one must or may rationally believe its consequent. It may be that one also already believes the negation of the consequent and should then either stop believing the conditional or stop believing its antecedent.

A further point is that inference takes time and uses limited resources. Given that any particular set of beliefs has infinitely many logical consequences, it is simply not true that one rationally should waste time and resources cluttering one's mind with logical implications of what one believes.

Similar remarks apply to consistency and coherence. Formal logic, probability theory, and decision theory characterize consistency of propositions and coherence of assignments of probability and utility. Such formal theories do not say anything about what combinations of propositions people should or should not assert or believe, or what assignments of probability and utility they should accept. There may be corresponding principles connecting consistency and coherence with what people should not rationally believe or assert, but again those principles go beyond anything in formal logic, probability theory, and decision theory, and again it is nontrivial to find such principles that are at all plausible.

Given limited resources, it is not normally rational to devote significant resources to the computationally intractable task of checking one's beliefs and probability assignments for consistency and coherence. Furthermore, having discovered inconsistency or incoherence in one's beliefs and assignments, it is not necessarily rational to drop everything else one is doing to try to figure out the best way to eliminate it. The question of what to do after having discovered inconsistency or incoherence is a practical or methodological issue that can be addressed only by a normative theory of reasoning. The answer is not automatically provided by formal logic, probability theory, and decision theory.

As we mentioned earlier, the terms "logic," "probability theory," and "decision theory" can be used not only for purely formal theories but also for methodological accounts of how such formal theories might be relevant to rational reasoning and argument (Mill, 1846: Dewey, 1938). Our point is that these methodological proposals are additions to the purely formal theories and do not follow directly from them.

3.2. *Application to Moral Reasoning*

These general points have devastating implications for normative uses of the deductive model of moral reasoning. The deductive model suggests that a believer becomes justified in believing a moral claim when that person formulates an argument with that moral claim as a conclusion and other beliefs as premises. The argument supposedly works if and only if it is deductively valid.

This picture runs into several problems. First, suppose Jane believes that (C) it is morally wrong to send her old paper to Kate, she also believes that (P1) cheating is always morally wrong and that (P2) sending her old paper to Kate would be cheating, and she knows that the latter two beliefs entail the former. Add also that her belief in the conclusion is causally based on her belief in the premises. According to the deductive model, the fact that she formulates this argument and bases her belief on it makes her justified in believing her moral conclusion.

But this can't be right. When Jane believes the premises, formulates the argument, and recognizes its validity, she still has several options. The argument shows that Jane cannot consistently (a) believe the premises and deny the conclusion. If that were the only other alternative, then Jane would have to (b) believe the premises and believe the conclusion. Still, Jane could instead deny a premise. She could (c) give up her belief that all cheating is morally wrong and replace it with the vague belief that cheating is usually wrong or with the qualified belief that all cheating is wrong except when it is the only way to help a friend in need. Alternatively, Jane could (d) give up her belief that sending her paper to Kate would be cheating, if she redefines cheating so that it does not include sharing paper drafts that will later be modified.

How can Jane decide among options (b)–(d)? The deductively valid argument cannot help her decide, since all that argument does is rule out option (a) as inconsistent. Thus, if reasoning is change of belief or intention, as discussed above, then the deductive argument cannot tell Jane whether to change her beliefs by adding a belief in the conclusion, as in (b), or instead to change her beliefs by removing a belief in some premise, as in (c) and (d). Since moral reasoning in this case involves deciding whether or not to believe the conclusion, this moral reasoning cannot be modeled by the deductive argument, because that deductive argument by itself cannot help us make that decision.

This point is about epistemology: formulating and basing one's belief on a deductively valid argument cannot be sufficient to justify belief in its conclusion. Nonetheless, as a matter of descriptive psychology, it still might be

true that people sometimes believe premises like P1 and P2, then notice that the premises entail a conclusion like C, so they form a new belief in C as a result. The facts that they could give up one of the premises in order to avoid the conclusion and that they might be justified in doing so do not undermine the claim that people often do in fact accept conclusions of deductive arguments in order to maintain consistency without giving up their beliefs in the premises. Some of the psychological pressure to add the conclusion to one's stock of beliefs might come from a general tendency not to give up beliefs that one already holds (a kind of doxastic conservatism). The remaining question, of how often moral beliefs are actually formed in this way, is a topic for further study in empirical moral psychology.

4. Are the Premises Independent?

Many philosophers who use the deductive model seem to assume that some of its premises are morally loaded and others are morally neutral. Recall, again,

(P1) Cheating is always morally wrong except in extreme circumstances.
(P2) This act is cheating.
(P3) This circumstance is not extreme.
Therefore, (C) this act is morally wrong.

Premise P1 is definitely a moral principle, but premise P2 might seem to be a morally neutral classification of a kind of act. However, it is difficult to imagine how to define "cheating" in a morally neutral way; "cheating" seems to be a morally loaded concept, albeit a ""thick" one (Williams, 1985). One of us has a golfing buddy who regularly kicks his ball off of tree roots so that he won't hurt his hands by hitting a root. He has done this for years. Is it cheating? It violates the normal rules of golf, but people who play with him know he does it, and he knows that they know and that they will allow it. Groups of golfers often make special allowances like this, but he and his friends have never made this one explicit, and it sometimes does affect the outcomes of bets. Although it is not clear, it seems likely that people who think that he should not kick his ball will call this act cheating, and people who think there is nothing wrong with what he does will not call it cheating. If so, the notion of cheating is morally loaded, and so is premise P2.

This point applies also to lying: even if a person already accepts a principle, such as "Never lie," she or he still needs to classify acts as lying. Are white lies lies? You ask me whether I like your new book. I say "Yes," when I really

think it is mediocre or even bad. Is that a lie? Those who think it is dishonest and immoral will be more likely to call it a lie. It is not clear whether people call an act immoral because they classify it as a lie or, instead, call it a lie because they see it as immoral. Which comes first, the classification or the judgment? Maybe sometimes a person considers an act, realizes that it is (or is not) a lie on some morally neutral definition, applies the principle that lying is always wrong, and concludes that the act is wrong. But it is not clear how often that kind of moral reasoning actually occurs.

One case is particularly important and, perhaps, surprising. A common moral rule or principle says that it is morally wrong to kill intentionally without an adequate reason. Many philosophers seem to assume that we can classify an act as killing (or not) merely by asking whether the act causes a death, and that we can determine whether an act causes a death independently of any moral judgment of the act. (Causation is a scientific notion, isn't it?) Many philosophers also assume that to say an act was done intentionally is merely to describe the act or its agent's mental state. (Intentions are psychological states, aren't they?)

These assumptions have been questioned, however, by recent empirical results. First, Alicke (1992) has shown that whether an act is picked out as the cause of a harm depends in at least some cases on whether the act is judged to be morally wrong. Second, Knobe (2003) reported results that are often interpreted as suggesting that whether a person is seen as causing a harm intentionally as opposed to unintentionally also depends on background beliefs about the moral status of the act or the value of its effects. More recently, Cushman, Knobe, and Sinnott-Armstrong (2008) found that whether an act is classified as killing as opposed to merely letting die also depends on subjects' moral judgments of the act. All of these experiments suggest that people do not classify an act as intentional killing independently of their moral judgments of that act.

Defenders of the deductive model could respond that, even if classifications like cheating, lying, and killing as well as cause and intention are not morally neutral, other classifications of acts still might be morally neutral. The problem is that neutral classifications would be difficult to build into unhedged moral principles that people could apply deductively. Even if this difficulty could be overcome in theory, there is no evidence that people actually deduce moral judgments from such neutral classifications. Instead, moral reasoners usually refer to the kinds of classifications that seem to be already morally loaded—like cheating, lying, and killing.

These results undermine the deductive model's assumption that common moral reasoning starts from premises that classify acts independently of moral

judgments. Without that assumption, the deductive arguments postulated by the deductive model cannot really be what lead people to form their moral judgments. We do not classify acts as causing harm, as intentional, or as killing and then reach a moral conclusion later only by means of applying a moral principle. Instead, we form some moral judgment of the act before we classify the act or accept the minor premise that classifies the act. The argument comes after the moral judgment. Moreover, anyone who denies the conclusion will or could automatically deny one of the premises, because that premise depends on a moral judgment of the act, perhaps the very moral judgment in the conclusion. This kind of circularity undermines the attempt to model real moral reasoning as a deductive structure.

5. Are Moral Judgments based on Principles?

Sometimes it seems obvious to us that we judge an act as morally wrong because we have already classified that act as killing. However, this appearance might well be an illusion. One recent study suggests that it is an illusion in some central cases.

Sinnott-Armstrong, Mallon, McCoy, and Hull (2008) collected moral judgments about three "trolley problems," a familiar form of philosophical thought experiment that has lately been exploited in empirical work. In the *side track* case (see Figure 6.1), a runaway trolley will kill five people on a main track unless Peter pulls a lever that will deflect the trolley onto a side track where it will kill only one and then go off into a field.

In the *loop track* case (see Figure 6.2), a runaway trolley will again kill five people on a main track unless Peter pulls a lever that will deflect the trolley

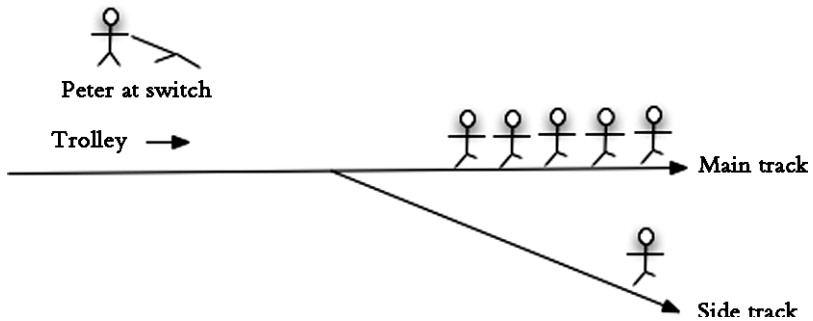

Figure 6.1. The trolley problem: side track case

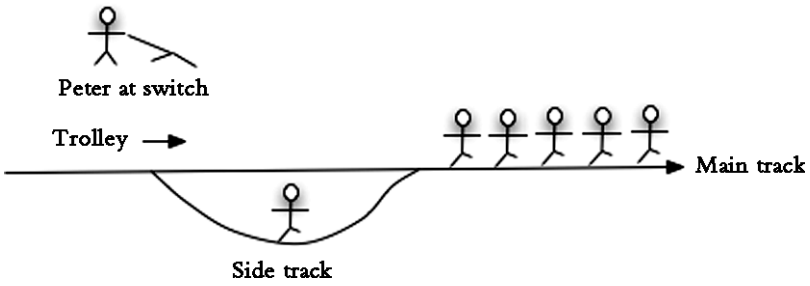

Figure 6.2. The trolley problem: loop track case

onto a side track, but this time the side track loops back and rejoins the main track so that the trolley would still kill the five if not for the fact that, if it goes onto the side track, it will hit and kill one person on the loop track, and that person's body will slow the trolley to a stop before it returns to the main track and hits the five.

The third case is *combination track* (see Figure 6.3). Here Peter can save the five on the main track only by turning the runaway trolley onto a side track that loops back onto the main track just before the five people (as in the loop track case). This time, before this loop track gets back to the main track, a second side track splits off from the loop track into an empty field (as in the side track case). Before the trolley gets to this second side track, Peter will be able to pull a lever that will divert the trolley onto the second side track and into the field. Unfortunately, before the trolley gets to the second side track, it will hit and kill an innocent person on the looped side track. Hitting this person is not enough to stop the trolley or save the five unless the trolley is redirected onto the second side track.

The two factors that matter here are intention and timing. In the loop track case, subjects judge that Peter intentionally kills the person on the side track,

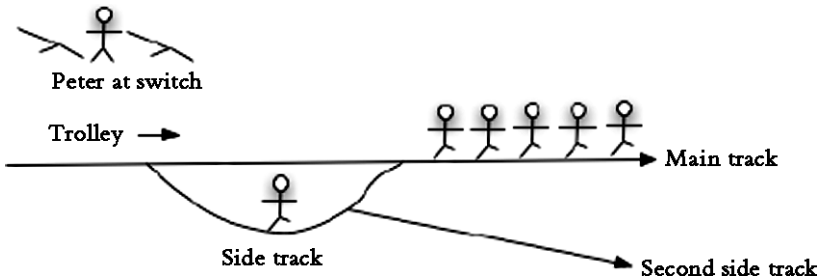

Figure 6.3. The trolley problem: combination track case

because Peter's plan for ending the threat to the five will fail unless the trolley hits the person on the side track. In contrast, in the side track and combination track cases, Peter's plan to save the five will work even if the trolley misses the person on the side track, so subjects judge that Peter does not intentionally kill the person on the side track in either the side track case or the combination track case.

Next consider timing. In the side track case, the bad effect of the trolley hitting the lone person occurs *after* the good effect of the five being saved, because the five are safe as soon as the trolley enters the side track. In contrast, in the loop track and combination track cases, the bad effect of the trolley hitting the person occurs *before* the good effect of the five being saved, since that good effect occurs only when the trolley slows down in the loop track case and only when the trolley goes off onto the second side track in combination track.

Comparing these three cases allows us to separate effects of intention and timing. Sinnott-Armstrong et al. asked subjects to judge not only whether Peter's act was morally wrong but also whether Peter killed the person on the side track. They found that subjects' moral judgments of whether Peter's act was wrong *did* depend on *intention*, because these moral judgments were different in the case where the death was intended (the loop track case) than in the cases where the death was not intended (the side track and combination track cases). However, these moral judgments did *not* depend on *timing*, because they were statistically indistinguishable in the case where the good effect occurred first (side track) and the cases where the bad effect occurred first (the loop track and combination track cases). The opposite was found for classifications of Peter's act as killing. Whether subjects classified Peter's act as killing did *not* depend on *intention*, because subjects' classifications were statistically indistinguishable in the case where the death was intended (the loop track case) and the cases where the death was not intended (the side track and combination track cases). However, subjects' classifications as killing did depend on *timing*, because these classifications were statistically different in the case where the good effect occurred first (the side track case) than in the cases where the bad effect occurred first (the loop track and combination track cases). In short, intention but not timing affects judgments of moral wrongness, whereas timing but not intention affects classifications as killing.

This comparison suggests that these subjects did not judge the act to be morally wrong simply because it is killing (or intentional killing). If moral judgments really were based on such a simple rule about killing, then what affects classifications as killing would also affect moral judgments of wrongness. Since temporal order affects what subjects classify as killing, temporal order

would also indirectly affect subjects' moral judgments. However, temporal order *did* affect classifications as killing, but temporal order did *not* affect subjects' moral judgments. Thus their moral judgments in such cases must not be based on deduction from any simple rule about killing.

These subjects did, however, often externally justify their moral judgments by saying that it is morally wrong to kill. They seemed to think that they were deducing their moral judgments from a moral principle against killing. This shows that, at least in some central cases, when moral judgments seem to be based on deductive arguments, this appearance is an illusion. It is not clear whether this result generalizes to other cases, but, if it does, then this result would be a serious problem for the deductive model as a description of actual moral reasoning.

The lesson is not that moral reasoning *never* works by classification and deduction. Moral reasoning still might fit the deductive model in other cases. That seems plausible, for example, in cases when someone reasons that an otherwise neutral act is illegal and, hence, is morally wrong (although the source of moral judgment is not clear even in that case). At least we can reach this conclusion: moral reasoning sometimes does not work by deducing moral judgments from intuitive, consciously accessible moral principles, even when it seems to. This weak conclusion might be all that we can say, because there is no reason to think that all actual moral reasoning must fit into any specific form or model.

6. Against the Classical View of Concepts

Moral reasoning seems to involve, at the very least, some manipulation of one's prior moral beliefs. To understand the nature of moral reasoning, then, we must understand how the mind stores and uses these prior moral beliefs, including, among other things, beliefs about moral norms and about the rightness or wrongness of particular actions or types of action.

According to the deductive model, our moral beliefs are stored in the form of general principles, and we reason about particular cases by applying these general principles to specific cases. The general principles might be very complex, but they are nonetheless genuine principles, as opposed to, say, a disjunctive list of specific instances.

In this section, following Stich's (1993) lead, we shall discuss several different theories of how mental content is stored and manipulated, and how these theories relate to the deductive model and broader debates in moral philosophy.

We start with computationalism and the so-called classical view of concepts, and then move on to views that are progressively less friendly to the deductive model.

6.1. *What are Concepts?*

The dominant account of mental processes in cognitive science is *computationalism*, according to which the brain is a sort of computer, and mental processes are computations (e.g. Haugeland, 1978; Newell & Simon, 1976). In its orthodox version, computationalism is typically combined with the *representational theory of mind* (RTM). According to RTM, mental computations involve the manipulation of stored mental representations, which are like sentences in natural language. Just as sentences are composed of simpler elements (words) that contribute their semantic value to the overall semantic value of the sentence, so thoughts are composed of simpler representations that, likewise, contribute their semantic value to the overall semantic value of thoughts. These simpler mental representations are *concepts*. Concepts are the constituents of thought, and thoughts inherit their content from the concepts they contain. Like words, concepts can represent kinds, properties, sets, individuals, relations, and so on.

According to orthodox computationalism, then, mental information is generally stored as strings of mental representations—sentences in Mentalese, so to speak, composed from concepts. When Jane makes the new moral judgment that it would be wrong to email Kate her first paper, she comes to stand in a new relationship to a complex representation with the content "it would be wrong to email Kate my first paper." That representation is itself composed of simpler representations, concepts like KATE, EMAIL and WRONG ACTION (we follow the convention of representing concepts in small caps). Jane's moral knowledge consists in her having representations like these, and their standing in certain relationships to one another. To understand the nature of moral knowledge, then, we further need to understand the nature of concepts.[1]

FN:1

6.2. *The Classical View of Concepts*

One theory of concepts has been so influential that it is called the *classical view* (Laurence & Margolis, 1999). According to the classical view, all or most concepts have definitions in terms of necessary and sufficient conditions.

[1] For far more expansive reviews of the literature on concepts and many of the theoretical details we omit here, see Margolis & Laurence (1999) and Murphy (2002).

UNCLE, for example, might be defined as BROTHER OF A PARENT. The concepts contained in the definitions are themselves defined in terms of simpler concepts (MALE and SIBLING), which, in turn, are defined in still simpler concepts, and so on, down to a bedrock of basic, undefined elements.

In one particularly common formulation of the view, which Laurence and Margolis (1999) call the *containment model*, complex concepts are literally *composed* of the elements in their definitions. In that case, UNCLE would be literally made up of the "simpler" concepts BROTHER and PARENT, which themselves would be made up of still simpler concepts. On an alternative formulation, the *inferential* model, complex concepts are not *literally* composed of the concepts in their definitions, but rather stand in an inferential relation to them. UNCLE would then be inferentially related to the concepts BROTHER and PARENT in a way that represents the belief that uncles are (necessarily) brothers of parents.

The classical view of concepts is so called because, in either the containment or inferential version, it was predominant in philosophy until the mid-twentieth century. As already noted, when Plato presents Socrates' search for necessary and sufficient conditions for justice and right action, for instance, he seems to be assuming the classical view. Similarly, Locke fairly explicitly endorses the containment model when he claims that the "*Idea* of the *Sun*" is "an aggregate of those several simple *Ideas*, Bright, Hot, Roundish, having a constant regular motion, at a certain distance from us, and, perhaps, some other" (Locke, 1700/1975: 298–299). And Kant endorses a containment version of the classical view when he characterizes analytic judgments as those where we judge the predicate to be "(covertly) contained" in the subject (Kant, 1787/1929: 48).

The classical view gives rise to a very clear picture of how categorization works. First we categorize an object or event as falling under various basic concepts, and then simply build up more complex concepts from those. So Locke might claim that, when we look up into the sky and see the sun, we first recognize it as BRIGHT, HOT, ROUNDISH, and so on. On the containment model, we've thereby categorized it as the sun, for the concept SUN just *is* the aggregate of those simpler concepts. On the inferential model, we may have to engage in some further inference from those basic concepts to the complex concept of which they are a definition.

6.3. *Challenges and Revisions to the Classical View*

The classical view was long the dominant theory of concepts in philosophy and psychology. But in the 1970s, a series of empirical problems emerged, the

most important of which were the so-called *typicality effects* (e.g. Rips et al., 1973; Rosch & Mervis, 1975). For any given concept, people can judge how typical an item is of that concept, and there is high inter-subject agreement on these judgments. For instance, American subjects consider a robin a more typical instance of the concept BIRD than a chicken is. More typical items are categorized faster than less typical items and are more often named when subjects are asked for instances of the concept. Finally, typical items are not necessarily the most frequent or familiar items in the list. A robin is judged to be a more typical bird than a chicken, even though many people have fewer encounters with robins than with chickens (albeit dead ones). What makes an instance a typical item for a concept is, rather, that it has lots of features in common with other instances of the same concept and few features in common with non-instances. Thus robins fly, sing, lay eggs, are small, nest in trees and eat insects—common features in birds. Chickens, by contrast, don't fly, don't sing, don't nest in trees, and aren't small—and therefore chickens aren't typical birds.

Strictly speaking, these empirical effects aren't incompatible with the classical view, but *prima facie* it is hard to see why classically defined concepts would produce them. If BIRD has a definition in terms of necessary and sufficient conditions, then presumably chickens and robins meet those conditions equally well. And since robins aren't more frequent or familiar than chickens, there is nothing in the classical theory to explain why robins are categorized as birds more quickly than chickens, why they are judged more typical, and so on.

There were two main responses to these results: the first was to retain the basic idea of the classical view that concepts have definitions, but add a separate *identification procedure* that could account for the typicality effects; second, to account for the typicality effects directly, by building typicality into the very structure of concepts and therefore rejecting the classical view entirely. On the first response, sometimes called a *dual theory* of concepts, concepts still have a definitional *core* of necessary and sufficient conditions that an item must meet to count as an instance of that concept. But in addition to their core, concepts also have an associated set of non-definitional features that can be used in quick-and-dirty categorization; these are the identification procedure.[2] For the concept BIRD, these features might include features like flies, sings, etc. Since these features are used in quick identification, robins will be categorized as birds more quickly than chickens will. But since chickens meet the core conditions for the concept BIRD, subjects will also agree that chickens are nonetheless birds as well.

FN:2

[2] For examples of dual theories, see Armstrong et al. (1983); Osherson & Smith (1981).

Few psychologists, however, chose to pursue dual theories in the 1980s and 1990s. The more common response was, rather, to develop new theories of concepts, of which two are particularly relevant here: prototype and exemplar theories.

6.4. *Alternatives to the Classical View*

Prototype Theories In the *Philosophical Investigations*, Wittgenstein (1953) introduced the idea that many concepts could not be defined in terms of necessary and sufficient conditions. He illustrated this with the concept GAME. Rather than any shared set of features, what makes board-games, card-games, ball-games etc. all instances of GAME are *family resemblances* between the instances. Some board-games have some things in common with some card-games, other board-games have other things in common with ball-games, and so on. What makes these all instances of GAME is that each instance has enough features in common with enough other instances. Concepts like GAME are also sometimes called *cluster concepts* (Gasking, 1960) since they have a cluster of common features, none of which is necessary to be an instance of the concept.

In response to the typicality effects they had discovered, Rosch and Mervis (1975) developed Wittgenstein's ideas into a psychological theory of concepts. According to prototype theories, a concept is a *prototype*—a summary representation of information about a category, something like a weighted list of (representations of) features generally shared by instances of the concept. The weights for each feature determine how important that feature is to the concept but, as with cluster concepts, in general no single feature is necessary. An item is an instance of the concept if it shares enough of the weighted features.

Thus, for instance, the concept BIRD might contain the following features (among others): flies, sings, lays eggs, is small, nests in trees, eats insects (Smith, 1995). And these features will each be weighted; perhaps "lays eggs" has high weight, "flies" and "eats insects" slightly lower, and "is small" not very much weight at all.

On a prototype theory, categorization is a comparison of an item to various feature lists. We can think of the comparison as calculating how many "points" the item scores on different lists. For every feature on a given list that the item shares, it gains "points"; for every feature that the item lacks, it loses points. The number of points lost and gained for each feature depends on that feature's weight. The item is categorized as an instance of whichever concept gives it the most points. There are various models for the underlying calculations, which give rise to various more specific prototype theories.

According to prototype theorists, this calculation of "points" is really a calculation of the "similarity" between two types of mental representation. On the one hand, there is the representation of the item to be categorized. On the other hand are the various concepts under which it could be categorized. Categorization is a comparison of the first sort of representation with the second sort; the item is categorized as an instance of whichever concept its representation is most similar to.

Exemplar Theories There are also exemplar theories of concepts (e.g. Medin & Schaffer, 1978). Like prototype theories, exemplar theories claim that concepts don't have definitional cores. But exemplar theories go one step further than prototype theories, denying that concepts contain any kind of summary representation about instances of the concept. Instead, a concept just *is* a set of stored (representations of) instances. The concept BIRD just is a set of representations of birds.

As with prototype theories, categorization in exemplar theories is a measurement of similarity. To categorize an item, agents measure the similarity of (their representation of) that item to the various representations constituting different concepts, using some similarity metric or other. (As with prototype theories, various exemplar theories differ in the details of these similarity metrics.) An item is categorized as an instance of BIRD just in case it is similar enough to one's representations or *exemplars* of previously encountered birds.

Nearest Neighbor Pattern Classification Issues of categorization are also addressed under the heading of "pattern recognition" or ""pattern classification" in statistical learning theory and machine learning (e.g. Duda et al., 2001; Harman & Kulkarni, 2007). In the basic framework, the task is to use certain features of an item to assign a category or label to it. Features can have many possible values and are taken to be real valued. If there are D features, there is a D-dimensional *feature space* with a dimension for each feature. Each point in the feature space represents an assignment of values to each of the D features.

In the basic framework, it is assumed that there is an unknown background statistical probability distribution determining probabilistic relations between features and labels and the probabilities that items with various features and labels will be encountered. Data for classification are represented as a set of labeled points in feature space.

A 1-nearest neighbor classification assumes a metric on the feature space and assigns to any unlabeled point the same label as its nearest labeled point. A k-nearest neighbor classification assigns to any unlabeled point the label of a majority of its k nearest points. A k_n-nearest neighbor classification assigns

to an unlabeled point the label of a majority of its k_n nearest points, where k_n is a function of n such that as $n \to \infty$: $k_n \to \infty$ but $\frac{k_n}{n} \to 0$. Clearly, there is a resemblance between such nearest neighbor classifications as studied in statistical learning theory and exemplar theories as studied in psychology.

Statistical learning theory is concerned with finding classification rules that minimize expected error and not with finding rules that model human psychology. But the rules that statistical learning theory endorses may provide suggestive models for psychology.

6.5. *Categorization and Moral Principles*

The Classical View, the Deductive Model, and Moral Generalism Of all the theories of mental representation, the theory most amenable to the deductive model is the classical view. If the classical view is correct, then moral concepts have definitions in terms of necessary and sufficient conditions, and so moral categorization consists of testing cases against those definitions. It seems fair enough to construe this process as the application of moral principles, the principles being embodied by the definitions of moral concepts.

To make things a bit more concrete: if moral concepts have classical structure, then a concept such as RIGHT ACTION has a definition. Suppose (what is probably false, but it doesn't matter) that the definition is ACTION THAT MAXIMIZES EXPECTED UTILITY. When we decide whether or not an action is right, we test it against this definition. And this looks very much like the application of the principle "an action is right if and only if it maximizes expected utility." So the classical view of concepts looks like a straightforward implementation of the deductive model.

We need not suppose that a "definition" in this sense captures the *sense* of the concept. The "definition" does not have to be *a priori*. It is enough that it captures one's current *conception*, one's current rule for determining whether the concept applies in a given case. Hare (1952) supposes that although one has such a rule at any given moment for *morally ought*, an account of the meaning of *morally ought* is not provided by that rule but rather by the theory that *morally ought* is used to express universalizable prescriptions, and Hare emphasizes that one may change one's rule without changing what one means by *morally ought*.

What would it mean if the deductive model did correctly describe our moral reasoning? Well, one natural thought might be that this gives some support to moral generalism in its debate against moral particularism.[3] First, however,

FN:3

[3] For examples of this debate, see Dancy (1993), Hooker & Little, (2000), McKeever and Ridge (2006), Sinnott-Armstrong (1999) and Väyrynen (2004).

some crucial terminology. We should distinguish between particularism and generalism as claims about moral *psychology* and as claims about moral *right and wrong*. As a claim about moral psychology, particularism is the view that human moral judgment does not involve the application of moral principles. Call this view *psychological particularism*. As a claim about what is morally right or wrong, particularism is the view that there really aren't any general principles that capture the difference between right and wrong. Call this view *normative particularism*.[4] By contrast, generalism is simply the view denied by particularism. Hence psychological generalism: moral judgment *does* involve the application of moral principles. And normative generalism: there really are general moral principles governing right and wrong.

If the deductive model is correct, then *psychological* generalism seems to be correct. There is a very natural sense in which categorization, in the classical view of concepts, involves the application of principles embodied in the definitions of concepts. But note that the principles might not resemble anything like the principles sought by most generalists. The classical view makes no specific predictions about the definitions for any particular concept, or even about which concepts an agent will have. For all the classical view claims, an agent's concept of RIGHT ACTION *might* have the utilitarian definition just given. Or it might have a massively disjunctive definition consisting of a list of particular right actions: an action is right if and only if it helps the poor in such-and-such a way and it's Tuesday, or it fulfills a promise to call home on Wednesday without hurting anyone, or . . . . Or an agent might not have a single concept of RIGHT ACTION, but instead a long list of different concepts for different types of right action: an action is RIGHT_1 if and only if it maximizes utility; an action is RIGHT_2 if and only if it is performed out of duty; and so on.[5]

In either of these two latter cases, there is a sense in which moral reasoning would be the application of moral principles (namely, the definitions of the moral concepts), and so the deductive model would be correct. It is not clear whether this would vindicate the sort of moral generalism theorists like Rawls defend, since the ''principles'' embodied in our moral concepts would be highly specific. On the other hand, Hare (1952: 56–78) expressly allows

[4] This distinction is related to, but different from, Sinnott-Armstrong's (1999) distinction between *analytic* and *metaphysical* particularism. Analytic particularism is a claim about the *content* of moral judgments, i.e. that they are about particular cases and not principles. By contrast, psychological particularism is a claim about the *mental processes* that produce moral judgments. These two claims could come apart. Moral judgments could be *about* particular cases (and so analytic particularism true) but produced by the application of moral principles (and so psychological particularism false).

[5] Of course, to motivate the claim that such an agent had several different concepts, rather than one disjunctive concept, we would need a good theory of concept individuation. And that's a whole other can of worms.

for implicitly accepted general moral principles that one cannot formulate in words. So perhaps some versions of generalism have room for even such highly specific moral concepts.

Of course even if *psychological* generalism is true, this tells us very little about the truth or falsity of *normative* generalism. It might be that our moral reasoning works by applying general principles, but that it is mistaken to do so. It is well known that our intuitive reasoning about probability and statistics often goes awry, leading us to make mathematical judgments that conflict with our best mathematical theories (Kahneman et al., 1982). So too may our moral psychology conflict with our best meta-ethical theory, if our moral psychology happens to work by applying moral principles but we come to decide (for whatever reason) that the best meta-ethical theory is particularism.

Similarity-Based Theories and Moral Particularism Conversely, it has been argued that, if the classical view is false and one of the similarity-based theories correct, then psychological particularism must be true (Churchland, 1996, 2000; Casebeer, 2003). We just discussed how testing items against definitions (as in the classical view) could be construed as a type of application of principles, but can the same move be made for categorization in similarity-based theories? Recall how categorization works in these theories. In prototype theories, we categorize by counting weighted features and judging whether an item has enough of the important features in common with other instances of the concept. In exemplar theories, we categorize by calculating the similarity between (our representation of) an item and stored representations of instances of the concept. In either case, it may not seem felicitous to describe the process as the application of principles.

Such a description may be more felicitous under *some* specifications of prototype theories, for prototypes are still summary representations of instances of the concept. If the features constituting the concept are few enough, and sufficiently recognizable as moral, then there is a sense in which we can describe categorization as the application of principles—they just happen not to be the sort of the principles we were expecting. Suppose, for instance, the concept of MURDER had three features, all of them weighted heavily: intentional killing; against the victim's will; without official legal sanction. In that case, there is a sense in which the agent follows a principle in categorizing an action as murder, the principle being something like "an action is a murder if and only if it is calculated as passing a certain threshold with respect to those three features." Such a principle, although different from what moral principles are generally taken to be, is still not that far off from a more familiar

principle like "an action is a murder if and only if it is an intentional killing against the victim's will without official legal sanction."[6] If moral concepts were prototypes, but were simple in this way, then there's a reasonable sense in which psychological particularism would be false and the deductive model correct.

Notice that Churchland supposes that the *moral principles* referred to in the issue between particularism and generalism are limited to principles that resemble the kinds of principles that people invoke in *external* discussion and argument with others, "the idea that our internal representations and cognitive activities are essentially just hidden, silent versions of the external statements, arguments, dialogues, and chains of reasoning that appear in our overt speech and print" (Churchland, 2000:293). This is actually an odd restriction when the topic is whether *internal* moral reasoning depends on principles. It would rule out the complex implicit moral principles that Hare (1952: 56–78) alludes to, and it is difficult to take similar restrictions seriously in other intellectual disciplines. For example, it would be bizarre to limit linguistic principles to those that ordinary people appeal to in ordinary discussions about language. Drawing an analogy between moral and linguistic knowledge, Rawls made just this point: linguistic principles are "known to require theoretical constructions that far outrun the ad hoc precepts of our explicit grammatical knowledge" and a "similar situation presumably holds in moral philosophy" (Rawls, 1971: 47). But, for the sake of unpacking Churchland's view, let us continue our discussion of generalism and particularism on the odd assumption that moral principles have to be simple and relatively familiar.

The point then is that prototype concepts need not be simple. The feature list, set of weights, and calculation of similarity might be extremely complex, so complex that they baffle easy characterization. This is the possibility that Casebeer and Churchland have in mind when they suggest that prototype theory entails particularism—the possibility that the "principles" embodied by prototypes are so utterly unlike familiar moral principles that they don't deserve to be called "principles." Thus, it is not so much their endorsement of prototype theory that leads them to psychological particularism, but their endorsement of prototype theory along with some specific claims about what moral prototypes are like.[7]

[6] Needless to say (but we'll say it anyway), we're not actually endorsing this as the correct characterization of MURDER.

[7] To his credit, Casebeer is much more explicit than Churchland about the need for this extra premise and is open to the possibility that the premise might be false (cf. Casebeer, 2003: 114–115).

Andy Clark describes this view of moral prototypes as follows: ''our moral knowledge [as represented by our moral prototypes] may quite spectacularly outrun anything that could be expressed by simple maxims or moral rules'' (Clark, 2000a: 271). At first blush, Clark's (2000a, 2000b) responses to Churchland seem to deny that this view of moral prototypes entails psychological particularism. But Clark seems to agree that the entailment holds; what he actually denies is that people's moral knowledge *is* entirely embodied in these kinds of prototypes.[8] Instead, Clark claims, people also encode moral principles—and if they encode and reason with moral principles, then psychological particularism is false.

Similarly, if moral concepts have exemplar structure, then we lose any hope of giving an informative characterization to the processes underlying moral categorization in terms of moral principles. For exemplar theories do away with summary representation altogether and construe concepts as just a set of representations of instances. The only principle that an agent could be applying in categorization, therefore, would be an extremely uninformative one: ''an action is morally right if and only if it sufficiently resembles my representations of other morally right actions.''

So it seems basically right to claim that psychological particularism is entailed by exemplar theories, or by prototype theories along with certain assumptions about the structure of the prototype.[9] Some versions of similarity-based theories do entail psychological particularism. But, as with the argument from classical theories to psychological generalism, it is important to see that this conclusion does not appear to have direct implications for ethics. The conclusion is psychological, not normative, particularism. The truth of the former (supposing it is true) is no argument for the latter, for the same reason that the analogous argument wouldn't hold in other domains like mathematics. Suppose, for example, that a view analogous to particularism were true for human mathematical reasoning—suppose, that is to say, that most of our judgments of numerosity, comparative size and so on didn't involve the application of simple and familiar rules or principles. Clearly we shouldn't infer from this fact about human psychology any claims about mathematics.

Someone might argue, however, that the relationship between moral psychology and moral right and wrong is distinctive, and that therefore an

[8] E.g. ''the marginalization of summary linguistic formulations and sentential reason [by Churchland] is a mistake. Summary linguistic formulations are not . . . mere tools for the novice. Rather, they are . . . crucial and (as far as we know) irreplaceable elements of genuinely moral reason'' (Clark, 2000a: 274).

[9] Our points in this and the preceding subsection are very similar to those made in Stich (1993).

argument from psychological to normative particularism would go through in the moral case. We think such an argument faces problems, but it might be developed as follows:

> Suppose that psychological particularism really is true. This would mean that, when we make moral judgments or engage in other moral reasoning, we aren't applying moral principles. Perhaps this is because our moral concepts are exemplars, or prototypes of a certain kind.
>
> Now, how do we discover what is right or wrong? Through our moral judgments and other moral reasoning. But if we don't use moral principles in our actual moral judgment, then we can never discover general principles of right and wrong through moral judgment. Contrast this with the case of mathematics. What makes us justified in thinking there is a gap between mathematical reasoning and mathematical reality is that the former is not our only guide to the latter. We can discover mathematical reality through more general processes of reasoning—logic, formal proofs, etc.
>
> But our only guide to moral right and wrong is our moral psychology itself. So even if there really are moral principles, we are epistemically blocked from them. Moreover, since we are thus forced to build an account of right and wrong which doesn't include moral principles, it would seem otiose to postulate that there really are moral principles all the same, lurking mysteriously in the background, ever out of our epistemic reach. Given our epistemic situation, metaphysical moral particularism is more parsimonious than metaphysical moral generalism.

This argument relies on two problematic premises. First, it assumes that if moral judgment doesn't involve the application of moral principles, then moral judgment can't be *described* by moral principles. But in general a process can be governed by principles even if it doesn't involve the principles represented anywhere in the process. The behavior of a billiard-ball physical system might be governed by a simple set of physical laws, even though the billiard balls themselves don't "apply" those laws. Similarly, human moral judgment might be following moral principles that aren't explicitly represented in human minds.

Second, the argument assumes that our intuitive moral psychology is our only guide to moral right and wrong. And there seems ample room to deny that, by claiming that we can also use more general processes of reasoning to discover moral right and wrong. We certainly use such general processes in other cases, such as mathematics, and there's no obvious reason to deny that we can use them in the moral case as well. We discuss reasoning toward narrow or wide reflective equilibrium below. In the present case we might use such reasoning to come up with a set of moral principles that captures most of our judgments. We might also expand the input into this equilibrium beyond moral judgment to more general considerations, such as theoretical simplicity and so on. Such a wide reflective equilibrium gives even more opportunity for

rejecting particular moral judgments, and therefore even greater space between moral reality and moral psychology.

In short, just as we saw in the previous section that psychological generalism does not entail normative generalism, so do we now see that psychological particularism does not entail normative particularism. It would only do so on the further assumptions that (1) if moral agents don't represent and apply moral principles, then their moral judgments aren't governed by moral principles, and (2) our only guide to the correct moral view is our intuitive moral judgments. The first assumption is unmotivated, and the second relies on a contentious moral epistemology. Thus, even if our moral concepts are exemplars or (the relevant kind of) prototype, there is still some hope for generalism as a claim about right and wrong.

7. Reflective Equilibrium

There are, therefore, several good reasons to suppose that the deductive model is not an adequate model. All of the four claims made by the deductive model appear to be false or, at least, to have little evidential support for most moral reasoning. We now briefly discuss an alternative model, already mentioned. On this alternative model, internal reasoning takes the form of making mutual adjustments to one's beliefs and plans, in the light of one's goals, in pursuit of what Rawls (1971) calls a "reflective equilibrium."

Thagard (1989, 2000) develops models of this process using connectionist "constraint satisfaction." The models contain networks of nodes representing particular propositions. Nodes can receive some degree of positive or negative excitation. There are two sorts of links among nodes, mutually reinforcing and mutually inhibiting. Positive links connect nodes with others such that as one of the nodes becomes more excited, the node's excitation increases the excitation of the other nodes and, as one such node becomes less excited or receives negative excitation, excitation of the other nodes is reduced. Negative links connect nodes such that as one node receives more excitation, the others receive less and vice versa. Excitation, positive and negative, cycles round and round the network until it eventually settles into a relatively steady state. Nodes in the steady state that have a positive excitation above a certain threshold represent beliefs. Nodes in the final state that have a negative excitation beyond a certain threshold represent things that are disbelieved. Nodes in the final state with intermediate excitation values represent things that are neither believed nor disbelieved. The resulting state of the network represents a system of beliefs in some sort of equilibrium.

It has often been noted that a connectionist network provides a possible model of certain sorts of Gestalt perception, for example, of a Necker cube (Feldman, 1981). A given vertex might be perceived either as part of the near surface or part of the back surface. This can be modeled by using nodes in a connectionist network to represent vertices and by setting up positive links among the vertices connected by horizontal or vertical lines and negative links between vertices connected by diagonal lines, where the degree of excitation of a vertex is used to represent how near it seems to the perceiver. As excitation on a given vertex is increased, this increases the excitation on the three other vertices of that face and drives down the excitation of the vertices on the other face. The result is that one tends to see the figure with one or the other face in front and the other in back. One tends not to see the figure as some sort of mixture or as indeterminate as to which face is in front.

Thagard (1989) uses his constraint satisfaction connectionist network to model the reasoning of jurors trying to assess the guilt of someone in a trial. The model makes certain predictions. For example, a juror might begin with a view about the reliability of a certain sort of eye-witness identification, a view about whether posting a message on a computer bulletin board is more like writing something in a newspaper or more like saying something in a telephone conversation, and so forth. Suppose the case being decided depends in part on an assessment of such matters. Then Thagard's model predicts that a juror's general confidence in this type of eye-witness identification should increase if the juror judges that in this case the testimony was correct and should decrease if the juror judges that in this case the testimony was not correct. The model predicts a similar effect on the juror's judgment about what posting on a computer network is more similar to, and so forth. The model also predicts that, because of these effects, the juror's resulting reflective equilibrium will lead to the juror's being quite confident in the verdict he or she reaches.

Experiments involving simulated trials confirm this prediction of Thagard's model (Simon, 2004). In these experiments, subjects are first asked their opinions about certain principles of evidence about certain sorts of eyewitness identifications, resemblances, etc. Then they are given material about difficult cases involving such considerations to think about. The subjects' final verdicts and their confidence in their verdicts and in the various principles of evidence are recorded.

One result is that, as predicted, although subjects may divide in their judgment of guilt at the end, with some saying the defendant is guilty and others denying this, subjects are very confident in their judgments and in the considerations that support them. Furthermore, also as predicted, there are

also changes in subjects' judgments about the value of that sort of eye-witness identification, about whether posting on a computer bulletin board is more like writing in a newspaper or having a private conversation, and so forth.

The model implies that judgments in hard cases are sometimes fragile and unreliable under certain conditions. When there is conflicting evidence, there is considerable tension among relevant considerations, just as there is a certain sort of tension among the nodes representing vertices in the Necker cube problem. If some nodes acquire even slightly increased or decreased excitation, the relevant inhibitory and excitatory connections can lead to changes in the excitation of other nodes in a kind of chain reaction or snowballing of considerations leading to a clear verdict, one way or the other, depending on the initial slight push, just as happens in one's perception of a Necker cube. After the Gestalt shift has occurred, however, the case may seem quite clear to the juror because of ways the juror's confidence has shifted in response to the positive and negative connections between nodes.

One upshot of this is that the slight errors in a trial that look like ''harmless errors'' can have a profound effect that cannot be corrected later by telling jurors to ignore something. By then the ignored evidence may have affected the excitation of various other items in such a way that the damage cannot be undone. Similarly, the fact that the prosecution goes first may make a difference by affecting how later material is evaluated.

This fragility of reflective equilibrium casts doubt on using the method of reflective equilibrium to arrive at reliable opinions. This sort of problem has been noted in discussions of Rawls's claim that justification of views about justice consists in getting one's judgments into reflective equilibrium. It is sometimes suggested that the problem might be solved by trying to find a ''wide'' rather than a ''narrow'' reflective equilibrium, where that involves not only seeing how one's current views fit together, but also considering various other views and the arguments that might be given for them, and trying to try to avoid the sorts of effects that arise from the order in which one gets evidence or thinks about an issue (Daniels, 1979). One needs to consider how things would have appeared to one if one had gotten evidence and thought about issues in a different order, for example. In this way one tries to find a *robust* reflective equilibrium that is not sensitive to small changes in one's starting point or the order in which one considers various issues.

Experimenters have shown that if subjects acting as jurors are instructed to try for this sort of wide, robust reflective equilibrium, they are less subject to the sorts of effects that occur when they are not so instructed (Simon, 2004). But the effects do not completely go away. So it is still not clear whether

this model succeeds as a normative theory of reasoning by jurors or by people forming moral judgments.

8. Conclusion

We have made a number of distinctions, between theoretical and practical reasoning; between internal personal reasoning, leading to changes in one's beliefs and intentions, and external social reasoning or argument with others; and between conscious and unconscious reasoning. Where philosophers tend to suppose that reasoning is a conscious process and that theories of reasoning can be assessed against introspection, most psychological studies of reasoning treat it as a largely unconscious process whose principles can be studied only indirectly with clever experiments.

We then described what we take to have been a particularly influential model of internal moral reasoning throughout the history of philosophy (and, to a lesser extent, of psychology)—namely, the deductive model. This model can be found, either in part or wholly, in Plato, Kant, Mill, Hare, rule-utilitarians (like Hooker), natural-law theorists (like Donagan), Kohlberg, and others. The deductive model is characterized by four claims: (1) deductive arguments make people justified in believing the conclusions of those arguments; (2) people's conscious belief in the premises of arguments makes them believe the conclusions of those arguments; (3) the premises in the arguments are independent; and (4) the terms in the arguments are classically defined.

What we have argued, over the second half of this chapter, is that all four of these premises are confused, false, or seriously in need of empirical support as claims about most moral reasoning. Against the first claim, it is no trivial matter to derive normative claims about how one *should* reason from formal theories of validity, so it cannot be assumed that deductive arguments produce justification for their conclusions. We also cannot assume that moral judgments are causally produced by beliefs in the premises of such arguments. Third, we presented evidence that people's moral reasoning isn't produced by arguments with independent premises, at least in some cases. We also presented some evidence that people's moral judgments can't be based on the principles that they actually adduce in their support. Finally, we discussed the substantial empirical problems with the classical view of concepts, described three alternatives, and discussed the implications of those alternatives.

Although many options remain available that we haven't ruled out, nonetheless it seems to us that the deductive model is implausible as a model for most

moral reasoning, in light of the evidence. So what other games are there in town? We discussed only one alternative model here, reflective equilibrium. And, as we saw, even if they turn out to be descriptively accurate models of actual moral reasoning, reflective equilibrium models have potentially worrying normative implications due to their fragility (at least in some models). This is obviously just the tip of the iceberg. Much more could be said, for instance, about the potential psychological and philosophical relevance of developments in machine learning (Harman & Kulkarni, 2007), or alternatives to classical computationalism such as mental models (Johnson-Laird, 2006) or connectionism. It would thus be an understatement to conclude that more work remains to be done in order to understand moral reasoning. The truth is, a *lot* more work remains to be done.

References

Alicke, M. D. 1992. "Culpable Causation," *Journal of Personality and Social Psychology* 63: 368–378.

Armstrong, S., Gleitman, L. & Gleitman, H. 1983. "What some concepts might not be," *Cognition* 13: 263–308.

Austin, J. L. 1953. *How to Do Things with Words*. Oxford: Oxford University Press.

Braine, M. D. S., and O'Brien, D. P. (eds.). 1998. *Mental Logic*. Mahwah, NJ: Lawrence Erlbaum Associates.

Casebeer, W. 2003. *Natural Ethical Facts: Evolution, Connectionism, and Moral Cognition*. Cambridge, MA: MIT Press.

Churchland, P. M. 1996. "The Neural Representation of the Social World." In L. May, M. Friedman & A. Clark (eds.), *Mind and Morals*. Cambridge, MA: MIT Press, 91–108.

——2000. "Rules, Know-How and the Future of Moral Cognition," *Canadian Journal of Philosophy* supplementary vol. 26: 291–306.

Clark, A. 2000a. "Word and Action: Reconciling Rules and Know-How in Moral Cognition," *Canadian Journal of Philosophy* supplementary vol. 26: 267–289.

——2000b. "Making Moral Space: A Reply to Churchland," *Canadian Journal of Philosophy* supplementary vol. 26: 307–312.

Cummins, D. 1996. "Evidence for the innateness of deontic reasoning," *Mind & Language* 11: 160–190.

Cushman, F. A., Knobe, J., & Sinnott-Armstrong, W. A. 2008. "Moral Judgments Affect Doing/Allowing Judgments," *Cognition* 108: 281–289.

Daniels, N., 1979. "Wide Reflective Equilibrium and Theory Acceptance in Ethics," *Journal of Philosophy* 76: 256–282.

Dancy, J. 1993. *Moral Reasons*. Oxford, Blackwell.

Dewey, J. 1938. *Logic: The Theory of Inquiry*. New York: Holt, Rinehart and Winston.

Donagan, A. 1977. *The Theory of Morality*. Chicago, IL: University of Chicago Press.

Duda, R. O., Hart, P. E., & Stork, D. G. 2001. *Pattern Classification*. New York: Wiley.

Feldman, J. A. 1981. "A Connectionist Model of Visual Memory." In G. E. Hinton & J. A. Anderson (eds.). *Parallel Models of Associative Memory*, Hillsdale, NJ: Erlbaum, 49–81.

Gasking, D. 1960. "Clusters," *Australasian Journal of Philosophy* 38: 1–36.

Gibbard, A., 1990. *Wise Choices, Apt Feelings: A Theory of Normative Judgment*. Cambridge, MA: Harvard University Press.

Gilligan, C., 1982. *In a Different Voice*. Cambridge, MA: Harvard University Press.

Goldfarb, W. 2003. *Deductive Logic*. Indianapolis, IN: Hackett.

Haidt, J. 2001. "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment," *Psychological Review* 108: 814–834.

Hampshire, S., 1959. *Thought and Action*. London: Chatto and Windus.

Hare, R. M., 1952. *Language of Morals*. Oxford: Clarendon Press.

——1963. *Freedom and Reason*. Oxford: Clarendon Press.

Harman, G., 1976. "Practical Reasoning," *Review of Metaphysics* 29: 431–463.

——1986. *Change in View*. Cambridge, MA: MIT Press.

Harman, G., & Kulkarni, S. 2007. *Reliable Reasoning: Induction and Statistical Learning Theory*. Cambridge, MA: MIT Press.

Harris, P., & Núñez, M. 1996. "Understanding of Permission Rules by Preschool Children," *Child Development* 67: 1572–1591.

Haugeland, J. 1978. "The Nature and Plausibility of Cognitivism," *Behavioral and Brain Sciences* 2: 215–226.

Holyoak, K. J., & Simon, D. 1999. "Bidirectional Reasoning in Decision Making by Constraint Satisfaction," *Journal of Experimental Psychology: General* 128: 3–31.

Hooker, B. & Little, M. 2000. *Moral Particularism*. New York: Oxford University Press.

Jaffe, S., & Hyde, J. S. 2000. "Gender Differences in Moral Orientation: A Meta-Analysis," *Psychological Bulletin* 126: 703–726.

Johnson-Laird, P. N. 2006. *How We Reason*. Oxford: Oxford University Press.

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kahneman, D., Slovic, P. & Tversky, A (eds) (1982) *Judgement Under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.

Kant, I. 1785/1993. *Grounding for the Metaphysics of Morals* (trans. J. Ellington, 3rd ed.). Indianapolis, IN: Hackett.

——1787/1929. *Critique of Pure Reason* (trans. N. K. Smith). London: Macmillan Press.

Knobe, J. 2003. "Intentional Action and Side Effects in Ordinary Language," *Analysis*, 63: 190–193.

Kohlberg, L. 1981. *Essays on Moral Development, Volume 1: The Philosophy of Moral Development*. New York: Harper & Row.

Lashley, K.S. 1958. "Cerebral Organization and Behavior." In H. C. Solomon, S. Cobb, & W. Penfield (eds.), *The Brain and Human Behavior*, Vol. 36. Association for Research in Nervous and Mental Disorders, Research Publications. Baltimore, MD: Williams and Wilkin, 1–18.

Laurence, S. & Margolis, E. 1999. "Concepts and Cognitive Science." In Margolis and Laurence (eds.), *Concepts: Core Readings*, Cambridge, MA: MIT Press, 3–81.

Levi, I. 1967. *Gambling with Truth*. New York: Knopf.

Locke, J. 1700/1975. *An Essay Concerning Human Understanding*. Oxford: Oxford University Press.

Margolis, E. & Laurence, S. 1999. *Concepts: Core Readings*. Cambridge, MA: MIT Press.

Medin, D. & Schaffer, M. 1978. "Context theory of classification learning." *Psychological Review* 85: 207–238.

McKeever, S. & Ridge, M. 2006. *Principled Ethics: Generalism As a Regulative Ideal*. Oxford: Oxford University Press.

Mill, J. S. 1846. *A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation*. New York: Harper.

——1861/2001. *Utilitarianism*. Indianapolis, IN: Hackett.

Murphy, G. L. 2002. *The Big Book of Concepts*. Cambridge, MA: MIT Press.

Nagel, T. 1970. *The Possibility of Altruism*. Oxford: Oxford University Press.

Newell, A., & Simon, H. 1976. "Computer Science As Empirical Inquiry: Symbols and Search." Reprinted in J. Haugeland (ed.), *Mind Design II*, Cambridge, MA: MIT Press, 1997, 81–110.

Osherson, D., & Smith, E. 1981. "On the Adequacy of Prototype Theory as a Theory of Concepts." *Cognition* 9: 35–58.

Polk, T. A., & Newell, A. 1995. "Deduction as Verbal Reasoning," *Psychological Review*, 102: 533–566.

Rawls, J. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Rips, L. J. 1994. *The Psychology of Proof*. Cambridge, MA: MIT Press.

Rips, L., Shoben, E., & Smith, E. 1973. "Semantic Distance and the Verification of Semantic Relations," *Journal of Verbal Learning and Verbal Behavior* 12: 1–20.

Rosch, E. & Mervis, C. 1975. "Family resemblances: studies in the internal structure of categories." *Cognitive Psychology* 7: 573–605.

Ryle, G. 1979. *On Thinking*, Totowa, NJ: Rowman and Littlefield.

Scanlon, T. M. 1998. *What We Owe To Each Other*. Cambridge, MA: Harvard University Press.

Simon, D., et al. 2001. "The Emergence of Coherence over the Course of Decision Making," *Journal of Experimental Psychology: Learning, Memory, and Cognition* 27: 1250–1260.

Simon, D. 2004. "A Third View of the Black Box: Cognitive Coherence in Legal Decision Making," *University of Chicago Law Review* 71: 511–586.

Sinnott-Armstrong, W. 1999. "Varieties of Particularism," *Metaphilosophy* 30: 1–12.

Sinnott-Armstrong, W., Mallon, R., McCoy, T. & Hull, J. (2008). "Intention, Temporal Order, and Moral Judgments," *Mind & Language* 23: 90–106.

Smith, E. 1995. "Concepts and Categorization." In E. Smith and D. Osherson (eds), *Thinking: An Invitation to Cognitive Science* vol. 3, 2nd ed. Cambridge, MA: MIT Press, 3–33.

Stich, S. 1993. "Moral Philosophy and Mental Representation," in M. Hechter, L. Nadel & R. Michod (eds.), *The Origin of Values*,New York: Aldine de Gruyter, 215–228.

Thagard, P. 1989. "Explanatory Coherence," *Behavioral and Brain Science* 12: 435–467.

——2000. *Coherence in Thought and Action*. Cambridge, MA: MIT Press.

Väyrynen, P., 2004. "Particularism and Default Reasons," *Ethical Theory and Moral Practice* 7: 53–79.

Williams, B. 1985. *Ethics and the Limits of Philosophy*. London: Fontana.

Wittgenstein, L. 1953. *Philosophical Investigations*,trans. G. E. M. Anscombe. Oxford: Blackwell.