

1

The Classical Model of Rationality and Its Weaknesses

I The Problem of Rationality

During the First World War a famous animal psychologist, Wolfgang Köhler, working on the island of Tenerife, showed that apes were capable of rational decision making. In a typical experiment he put an ape in an environment containing a box, a stick, and a bunch of bananas high up out of reach. After a while the ape figured out how to get the bananas. He moved the box under the bananas, got the stick, climbed up on the box, reached up with the stick and brought down the bananas.¹ Köhler was more interested in Gestalt psychology than in rationality, but his apes exemplified a form of rationality that has been paradigmatic in our theories. The idea is that rational decision making is a matter of selecting *means* that will enable us to achieve our *ends*. The ends are entirely a matter of what we desire. We come to the decision making situation with a prior inventory of desired ends, and rationality is entirely a matter of figuring out the means to our ends.

1. Wolfgang Köhler, *The Mentality of Apes*, second edition, London: Routledge and Kegan Paul, 1927. The animals were chimpanzees.

There is no question that the ape exemplifies one type of human rational decision making. But there is a very large number of other types of rational decision making that the ape did not, and presumably could not, engage in. The ape could attempt to figure out how to get bananas now, but he could not attempt to figure out how to get bananas next week. For humans, unlike the ape, much decision making is about the organization of time beyond the immediate present. Furthermore, the ape cannot consider large chunks of time terminating in his own death. Much human decision making, indeed most major decisions, such as where to live, what sort of career to pursue, what kind of family to have, whom to marry, has to do with the allocation of time prior to death. Death, one might say, is the horizon of human rationality; but thoughts about death and the ability to plan with death in mind would seem to be beyond the limitations of the ape's conceptual apparatus. A second difference between human rationality and the ape case is that humans are typically forced to choose between conflicting and incompatible ends. Sometimes that is true of animal decision making—Buridan's ass is a famous hypothetical case—but for Köhler's ape it was the box, stick, and bananas or nothing. The ape's third limitation is that he cannot consider reasons for action that are not dependent on his desires. That is, it seems that his desire to do something with the chair and the stick can be motivated only by a prior desire to eat the bananas. But in the case of human beings, it turns out we have a rather large number of reasons that are not desires. These desire-independent reasons can form the ground for desires, but their being reasons for us does not depend on their being based on desires. This is an interesting and contentious point, and I will return to it in more detail in

subsequent chapters. A fourth point of difference between ourselves and the ape is that it appears that the ape has, if anything, only a very limited conception of himself as a self, that is to say, as a rational agent making decisions and capable of assuming responsibility in the future for decisions taken in the present, or responsibility in the present for decisions taken in the past. And a fifth difference, related to the fourth, is that the chimp, unlike the human, does not see his decisions as in any way expressions of, nor commitments to, general principles that apply equally to himself and to other selves.

It is customary in these discussions to say that what the ape lacks is language. The idea, apparently, is that if only we could succeed in teaching the apes the rudiments of linguistic communication, they would have the full range of rational decision making apparatus and responsibility that humans do. I very much doubt that that is the case. The simple ability to symbolize is not by itself sufficient for the full gamut of rational thought processes. Efforts to teach chimpanzees to use symbols linguistically have had, at best, only ambiguous results. But even if they have succeeded, it seems to me that the types of use of symbols purportedly taught to Washoe, Lana, and other famous experimental chimps are insufficient to account for the range of human rational capacities that come with certain special features of human linguistic abilities. The point is that the mere capacity to symbolize does not by itself yield the full range of human rationality. What is necessary, as we will see in these pages, is the capacity for certain types of linguistic representation, and for those types it seems to me we cannot make a clear distinction between the intellectual capacities that are expressed in the notation and the use of the notation itself. The key is this: animals can

deceive but they cannot lie. The ability to lie is a consequence of the more profound human ability to undertake certain sorts of *commitments*, and those commitments are cases where the human animal intentionally imposes *conditions of satisfaction on conditions of satisfaction*. If you do not understand this point, don't worry; I will explain it in the chapters to come.

Persistent philosophical problems, like the problem of rationality, have a characteristic logical structure: How can it be the case that *p*, given that it appears to be certainly the case that *q*, where *q* apparently makes it impossible that *p*. The classic example of this pattern is, of course, the problem of free will. How can it be the case that we perform free actions, given that every event has a cause, and causal determination makes free actions impossible? The same logical structure pervades a large number of other problems. How can it be the case that we have consciousness, given that we are entirely composed of unconscious bits of matter? The same problem arises about intentionality: how can it be the case that we have intentional states—states that refer to objects and states of affairs in the world beyond themselves—given that we are made entirely of bits of matter that lack intentionality? A similar problem arises in skepticism: how can it be the case that we know anything, given that we can never be sure we are not dreaming, hallucinating, or being deceived by evil demons? In ethics: how can there be any values in the world, given that the world consists entirely of value-neutral facts? A variation on the same question: how can we know what *ought* to be the case given that all knowledge is about what *is in fact* the case, and we can never derive a statement about what ought to be the case from any set of statements about what is in fact the case?

The problem of rationality, a variant of these persistent problems, can be posed as follows. How can there be rational decision making in world where everything that occurs happens as a result of brute, blind, natural causal forces?

II The Classical Model of Rationality

In the discussion of ape rationality, I remarked that in our intellectual culture, we have a quite specific tradition of discussing rationality and practical reason, rationality in action. This tradition goes back to Aristotle's claim that deliberation is always about means, never about ends,² and it continues in Hume's famous claim that "Reason is and ought to be the slave of the passions," and in Kant's claim that "He who wills the end wills the means." The tradition receives its most sophisticated formulation in contemporary mathematical decision theory. The tradition is by no means unified, and I would not wish to suggest that Aristotle, Hume, and Kant share the same conception of rationality. On the contrary, there are striking differences between them. But there is a common thread, and I believe that of the classical philosophers, Hume gives the clearest statement of what I will be referring to as "the Classical Model." I have for a long time had doubts about this tradition and I am going to spend most of this first chapter exposing some of its main features and making a preliminary statement of some of my doubts. One way to describe the Classical Model is to say that it represents human rationality as a more complex version of ape rationality.

2. Alan Code has pointed out to me that this standard attribution may be a misunderstanding of Aristotle's actual views.

When I first learned about mathematical decision theory as an undergraduate at Oxford, it seemed to me there was an obvious problem with it: it seems to be a strict consequence of the axioms that if I value my life and I value twenty-five cents (a quarter is not very much money but it is enough to pick up off the sidewalk, for example), there must be some odds at which I would bet my life against a quarter. I thought about it, and I concluded there are no odds at which I would bet my life against a quarter, and if there were, I would not bet my child's life against a quarter. So, over the years, I argued about this with several famous decision theorists, starting with Jimmy Savage in Ann Arbor and including Isaac Levi in New York, and usually, after about half an hour of discussion, they came to the conclusion: "You're just plain irrational." Well, I am not so sure. I think maybe they have a problem with their theory of rationality. Some years later the limitations of this conception of rationality were really brought home to me (and this has some practical importance), during the Vietnam War when I went to visit a friend of mine, who was a high official of the Defense Department, in the Pentagon. I tried to argue him out of the war policy the United States was following, particularly the policy of bombing North Vietnam. He had a Ph.D. in mathematical economics. He went to the blackboard and drew the curves of traditional microeconomic analysis; and then said, "Where these two curves intersect, the marginal utility of resisting is equal to the marginal disutility of being bombed. At that point, they have to give up. All we are assuming is that they are rational. All we are assuming is that the enemy is rational!"

I knew then that we were in serious trouble, not only in our theory of rationality but in its application in practice.

It seems crazy to assume that the decision facing Ho Chih Minh and his colleagues was like a decision to buy a tube of toothpaste, strictly one of maximizing expected utility, but it is not easy to say exactly what is wrong with that assumption, and in the course of this book I want try to say exactly what is wrong with it. As a preliminary intuitive formulation we can say this much. In human rationality, as opposed to ape rationality, there is a distinction between reasons for action which are entirely matters of satisfying some desire or other and reasons which are desire independent. The basic distinction between different sorts of reasons for action is between those reasons which are matters of what you want to do or what you have to do in order to get what you want, on the one hand, and those reasons which are matters of what you have to do regardless of what you want, on the other hand.

Six Assumptions Behind the Classical Model

In this chapter I will state and discuss six assumptions that are largely constitutive of what I have been calling "the Classical Model of Rationality." I do not wish to suggest that the model is unified in the sense that if one accepts one proposition one is committed to all the others. On the contrary, some authors accept some parts and reject other parts. But I do wish to claim that the model forms a coherent whole, and it is one that I find both implicitly and explicitly influential in contemporary writings. Furthermore, the model articulates a conception of rationality that I was brought up on as a student of economics and moral philosophy at Oxford. It did not seem to me satisfactory then, and it does not seem to me satisfactory now.

1. Actions, where rational, are caused by beliefs and desires.

Beliefs and desires function both as causes and as reasons for our actions, and rationality is largely a matter of coordinating beliefs and desires so that they cause actions “in the right way.”

It is important to emphasize that the sense of “cause” here is the common or Aristotelian “efficient cause” sense of the word where a cause of an event is what makes it happen. Such causes, in a particular context, are sufficient conditions for an event to occur. To say that specific beliefs and desires caused a particular action is like saying that the earthquake caused the building to collapse.

2. Rationality is a matter of obeying rules, the special rules that make the distinction between rational and irrational thought and behavior.

Our task as theoreticians is to try to make explicit the inexplicit rules of rationality that fortunately most rational people are able to follow unconsciously. Just as they can speak English without knowing the rules of grammar, or they can speak in prose without knowing that they are speaking in prose, as in the famous example of Monsieur Jourdain, so they can behave rationally without knowing the rules that determine rationality and without even being aware that they are following those rules. But we, as theorists, have as our aim to discover and formulate those rules.

3. Rationality is a separate cognitive faculty.

According to Aristotle and a distinguished tradition that he initiated, the possession of rationality is our defining trait as humans: the human being is a rational animal.

Nowadays the fashionable term for faculty is “module,” but the general idea is that humans have various special cognitive capacities, one for vision, one for language, etc., and rationality is one of these special faculties, perhaps even the most distinctive of our human capacities. A recent book even speculates on the evolutionary advantages of our having this faculty.³

4. Apparent cases of weakness of will, what the Greeks called *akrasia*, can arise only in cases where there is something wrong with the psychological antecedents of the action.

Because rational actions are caused by beliefs and desires, and the beliefs and desires typically cause the action by first leading to the formation of an intention, apparent cases of weakness of will require a special explanation. How is it at all possible that an agent can have the right beliefs and desires, and form the right sort of intention, and still not perform the action? The standard account is that apparent cases of *akrasia* are all cases where the agent did not in fact have the right kind of antecedents to the action. Because the beliefs and desires, and derivatively the intentions, are causes, then if you stack them up rationally, the action will ensue by causal necessity. So in cases where the action does not ensue, there must have been something wrong with the causes.

Weakness of will has always been a problem for the Classical Model, and there is a lot of literature on the subject,⁴ but weakness of will is always made out to be

3. Robert Nozick, *The Nature of Rationality*, Princeton: Princeton University Press, 1993.

4. For an anthology of earlier work, see *Weakness of Will*, edited by G. W. Mortimore, London: Macmillan St. Martin's Press, 1971.

something very strange and hard to explain, something that could only happen under odd, or bizarre, circumstances. My own view, which I will explain later, is that *akrasia* in rational beings is as common as wine in France. Anybody who has ever tried to stop smoking, lose weight, or drink less at big parties will know what I am talking about.

5. Practical reason has to start with an inventory of the agent's primary ends, including the agent's goals and fundamental desires, objectives, and purposes; and these are not themselves subject to rational constraints.

In order to engage in the activity of practical reasoning, an agent must first have a set of things that he or she wants or values, and then practical reasoning is a matter of figuring out how best to satisfy this set of desires and values. We can state this point by saying that in order for practical reasoning to have any field in which to operate, the agent must begin with a set of primary desires, where desires are construed broadly, so that the agent's evaluations, whether moral, aesthetic, or otherwise, count as desires. But unless you have some such set of desires to start with, there is no scope for reason, because reason is a matter of figuring out what else you ought to desire, given that you already desire something. And those primary desires are not themselves subject to rational constraints.

The model of practical reason is something like the following. Suppose you want to go to Paris, and you reason how best to go. You could take a ship or go by kayak or take an airplane, and finally after the exercise of practical reason, you decide to take the airplane. But if this is the only way that practical reason can operate, by figuring out "means" to "ends," two things follow: first, there can be

no reasons for action that do not arise from desires, broadly construed. That is, there cannot be any desire-independent reasons for action. And second, those initial or primary desires cannot themselves be rationally evaluated. Reason is always about the means, never about the ends.

This claim—that there can be no desire-independent reasons for action—is at the heart of the Classical Model. Hume's statement that "Reason is and ought to be the slave of the passions" is usually interpreted as making this claim; and the same claim is made by many recent authors. For example, Herbert Simon writes, "Reason is wholly instrumental. It cannot tell us where to go; at best it can tell us how to get there. It is a gun for hire that can be employed in the service of any goals that we have, good or bad."⁵ Bertrand Russell is even more succinct: "Reason has a perfectly clear and concise meaning. It signifies the choice of the right means to an end that you wish to achieve. It has nothing whatever to do with the choice of ends."⁶

6. The whole system of rationality works only if the set of primary desires is consistent.

A typical expression of this view is given by Jon Elster: "Beliefs and desires can hardly be reasons for action unless they are consistent. They must not involve logical, conceptual, or pragmatic contradictions."⁷ It is easy to see

5. *Reason in Human Affairs*, Stanford, CA: Stanford University Press, 1983, pp. 7–8.

6. *Human Society in Ethics and Politics*, London: Allen and Unwin, 1954, p. viii.

7. *Sour Grapes: Studies in the Subversion of Rationality*, Cambridge: Cambridge University Press, 1983, p. 4.

why this seems plausible: if rationality is a matter of reasoning logically, there cannot be any inconsistencies or contradictions in the axioms. A contradiction implies anything, so if you had a contradiction in your initial set of desires, anything would follow, or so it seems.

Some Doubts about the Classical Model

I could continue this list, and we will have occasion to enrich the characterization of the Classical Model in the course of this book. But even this short list gives the general flavor of the concept, and I want to open the argument by giving some reasons why I think every one of these claims is false. At best they describe special cases, but they do not give a general theory of the role of rationality in thought and action.

1. Rational actions are not caused by beliefs and desires. In general only irrational and nonrational actions are caused by beliefs and desires.

Let us start, as an entering wedge, with the idea that rational actions are those that are caused by beliefs and desires. It is important to emphasize that the sense of “cause” is the ordinary “efficient cause” sense, as in: the explosion caused the building to collapse, or the earthquake caused the destruction of the freeway. I want to say that cases of actions for which the antecedent beliefs and desires really are causally sufficient, far from being models of rationality, are in fact bizarre and typically irrational cases. These are the cases where, for example, the agent is in the grip of an obsession or an addiction and cannot do otherwise than to act upon his desire. But in a

typical case of rational decision making where, for example, I am trying to decide which candidate to vote for, I have a choice and I consider various reasons for choosing among the alternatives available to me. But I can only engage in this activity if I assume that my set of beliefs and desires by itself is not causally sufficient to determine my action. The operation of rationality presupposes that there is a gap between the set of intentional states on the basis of which I make my decision, and the actual making of the decision. That is, unless I presuppose that there is a gap, I cannot get started with the process of rational decision making. To see this point you need only consider cases where there is no gap, where the belief and the desire are really causally sufficient. This is the case, for example, where the drug addict has an overpowering urge to take heroin, and he believes that this is heroin; so, compulsively, he takes it. In such a case the belief and the desire are sufficient to determine the action, because the addict cannot help himself. But that is hardly the model of rationality. Such cases seem to be outside the scope of rationality altogether.

In the normal case of rational action, we have to presuppose that the antecedent set of beliefs and desires is not causally sufficient to determine the action. This is a presupposition of the process of deliberation and is absolutely indispensable for the application of rationality. We presuppose that there is a gap between the "causes" of the action in the form of beliefs and desires and the "effect" in the form of the action. This gap has a traditional name. It is called "the freedom of the will." In order to engage in rational decision making we have to presuppose free will. Indeed, as we will see later, we have to presuppose free

will in any rational activity whatever. We cannot avoid the presupposition, because even a refusal to engage in rational decision making is only intelligible to us as a refusal if we take it as an exercise of freedom. To see this, consider examples. Suppose you go into a restaurant, and the waiter brings you the menu. You have a choice between, let's say, veal chops and spaghetti; you cannot say: "Look, I am a determinist, *che sarà, sarà*. I will just wait and see what I order! I will wait to see what my beliefs and desires cause." This refusal to exercise your freedom is itself only intelligible to you as an exercise of freedom. Kant pointed this out a long time ago: There is no way to think away your own freedom in the process of voluntary action because the process of deliberation itself can only proceed on the presupposition of freedom, on the presupposition that there is a gap between the causes in the form of your beliefs, desires, and other reasons, and the actual decision that you make.

If we are going to speak precisely about this, I think we must say that there are (at least) three gaps. First, there is the gap of rational decision making, where you try to make up your mind what you are going to do. Here the gap is between the reasons for making up your mind, and the actual decision that you make. Second, there is a gap between the decision and the action. Just as the reasons for the decision were not causally sufficient to produce the decision, so the decision is not causally sufficient to produce the action. There comes the point, after you have made up your mind, when you actually have to do it. And once again, you cannot sit back and let the decision cause the action, any more than you can sit back and let the reasons cause the decision. For example, let us suppose

you have made up your mind that you are going to vote for candidate Jones. You go into the voting booth with this decision firmly in mind, but once there you still have to do it. And sometimes, because of this second gap, you just do not do it. For a variety of possible reasons—or maybe none—you do not do the thing you have decided to do.

There is a third gap that arises for actions and activities extended in time, a gap between the initiation of the action and its continuation to completion. Suppose, for example, that you have decided to learn Portuguese, swim the English Channel, or write a book about rationality. There is first the gap between the reasons for the decision and the decision, second the gap between the decision and the initiation of the action, and third there is a gap between starting the task and its continuation to completion. Even once you have started you cannot let the causes operate by themselves; you have to make a continuous voluntary effort to keep going with the action or activity to its completion.

At this point of the discussion I want to emphasize two points: the existence of the gap(s) and the centrality of the gap(s) for the topic of rationality.

What is the argument for the existence of the gap(s)? I will develop these arguments in more detail in chapter 3; for present purposes we can say that the simplest arguments are the ones I just gave. Consider any situation of rational decision making and acting and you will see that you have a sense of alternative possibilities open to you and that your acting and deliberating make sense only on the presupposition of those alternative possibilities. Contrast these situations with those where you have no such sense of possibilities. In a situation in which you are in the

grip of an overpowering rage, so that you are, as they say, totally out of control, you have no sense that you could be doing something else.

Another way to see the existence of the gap is to notice that in a decision making situation you often have several different reasons for performing an action, yet you act on one and not the others and you know without observation which one you acted on. This is a remarkable fact, and notice the curious locution we have for describing it: *you acted on* such and such a reason. Suppose for example that you had a whole bunch of reasons both for and against voting for Clinton in the presidential election. You thought he would be a better president for the economy but worse for foreign policy. You liked the fact that he went to your old college but didn't like his personal style. In the end you voted for him because he went to your old college. The reasons did not operate on you. Rather you *chose* one reason and acted on that one. You made that reason effective by *acting on it*.

This is why, incidentally, the explanation of your action and its justification may not be the same. Suppose you are asked to justify voting for Clinton; you might do so by appealing to his superior management of the economy. But it may be the case that the actual reason you acted on was that he went to your old college in Oxford, and you thought, "College loyalty comes first." And the remarkable thing about this phenomenon is: in the normal case you know without observation which reason was effective, because you made it effective. That is to say, a reason for action is an effective reason only if you make it effective.

An understanding of the gap is essential for the topic of rationality because rationality can operate only in the gap. Though the concept of freedom and the concept of ratio-

nality are quite different, the extension of rationality is exactly that of freedom. The simplest argument for this point is that rationality is possible only where irrationality is possible, and that requirement entails the possibility of choosing between various rational options as well as irrational options. The scope of that choice is the gap in question. The claim that rationality can operate only in the gap is as much true of theoretical reason as it is of practical reason, but for theoretical reason it is a more subtle point to make, so I will save it for later and concentrate on practical reason now.

I will have a great deal more to say about the gap in the course of this book, and in a sense the book is about the gap, because the problem of rationality is a problem about the gap. At this stage just two more points:

First: what fills the gap? Nothing. Nothing fills the gap: you make up your mind to do something, or you just haul off and do what you are going to do, or you carry out the decision you previously made, or you keep going, or fail to keep going, in some project that you have undertaken.

Second: even though we have all these experiences, could not the whole thing be an illusion? Yes it could. Our gappy experiences are not self-validating. On the basis of what I have said so far, free will could still be a massive illusion. The *psychological* reality of the gap does not guarantee a corresponding *neurobiological* reality. I will explore these issues in chapter 9.

2. Rationality is not entirely or even largely a matter of following rules of rationality.

Let us turn to the second claim of the Classical Model, that rationality is a matter of rules, that we think and behave rationally only to the extent that we think and act

according to these rules. When asked to justify this claim, I think most traditional theorists would simply appeal to the rules of logic. An obvious kind of case that a defender of the Classical Model might present would be, let's say, a simple modus ponens argument:

If it rains tonight, the ground will be wet.

It will rain tonight.

Therefore, the ground will be wet.

Now, if you are asked to justify this inference, the temptation is to appeal to the rule of modus ponens: p , and if p then q , together imply q .

$$(p \& (p \rightarrow q)) \rightarrow q$$

But that is a fatal mistake. When you say that, you are in the grip of the Lewis Carroll paradox.⁸ I will now remind you how it goes: Achilles and the tortoise are having an argument, and Achilles says (this is not his example but it makes the same point), "If it rains tonight, the ground will be wet, it will rain tonight, therefore the ground will be wet," and the tortoise says, "Fine, write that down, write all that stuff down," And when Achilles had done so he says, "I don't see how you get from the stuff before the 'therefore' to the stuff after. What forces you to to make or even justifies you in making that move?" Achilles says, "Well that move rests on the rule of modus ponens, the rule that p , and if p then q , together imply q ." "Fine," says the tortoise, "So write that down, write that down with all the rest." And when Achilles had done so the tortoise says, "Well we have all that written down, but I still don't see how you get to the conclusion, that the ground will be

8. Lewis Carroll, "What Achilles Said to the Tortoise," *Mind* 4:278–280, April 1895.

wet." "Well don't you see?" says Achilles, "Whenever you have p , and if p then q , and you have the rule of modus ponens that says whenever you have p , and if p then q , you can infer q , then you can infer q ." "Fine," says the tortoise, "now just write all that down." And you see where this is going. We are off and running with an infinite regress.

The way to avoid an infinite regress is to refuse to make the first fatal move of supposing that the rule of modus ponens plays *any role whatever* in the validity of the inference. The derivation does not get its validity from the rule of modus ponens; rather, the inference is perfectly valid as it stands without any outside help. It would be more accurate to say that the rule of modus ponens gets its validity from the fact that it expresses a pattern of an infinite number of inferences that are independently valid. The actual argument does not get its validity from any external source: if it is valid, it can be valid only because the premises entail the conclusion. Because the meanings of the words themselves are sufficient to guarantee the validity of the inference, we can formalize a pattern that describes an infinite number of such inferences. But the inference does not derive its validity from the pattern. The so-called rule of modus ponens is just a statement of a pattern of an infinite number of such independently valid inferences. Remember: *If you think that you need a rule to infer q from p and (if p then q), then you would also need a rule to infer p from p .*

What goes for this argument goes for any valid deductive argument. Logical validity does not derive from the rules of logic.

It is important to understand this point precisely. It is usually said that the mistake of Achilles was to treat modus ponens as another premise and not as a rule. But

that is wrong. Even if he writes it down as a rule and not a premise, there would still be an infinite regress. It is equally wrong (indeed it is the same mistake) to say that the derivation derives its validity from both the premises and the rule of inference.⁹ The correct thing is to say that the rules of logic play no role whatever in the validity of valid inferences. The arguments, if valid, have to be valid as they stand.

We are actually blinded to this point by our very sophistication, because the achievements of proof theory have been so great, and have had such important payoffs in fields like computer science, that we think that the syntactical analogue of modus ponens is really the same thing as the “rule” of logic. But they are quite different. If you have an actual rule that says whenever you see, or your computer “sees,” a symbol with this shape:

p

followed by one with this shape:

$p \rightarrow q,$

you or it writes down one with this shape:

$q,$

you have an actual rule that you can follow and that you can program into the machine so as to causally affect its operations. This is a proof-theoretical analogue of the rule of modus ponens, and it really is substantive, because the marks that this rule operates over are just meaningless

9. For an example of this claim see Peter Railton, “On the Hypothetical and the Non-Hypothetical in Reasoning about Belief and Action,” pp. 53–79 in G. Cullity and B. Gaut, *Ethics and Practical Reason*, Oxford: Oxford University Press, 1997, esp. pp. 76–79.

symbols. The rule operates over otherwise uninterpreted formal elements.

Thus are we blinded to the fact that in real-life reasoning, the rule of modus ponens plays no justificatory role at all. We can make proof-theoretical or syntactical models, where the model exactly mirrors the substantive or contentful processes of actual human reasoning. And of course, as we all know, you can do a lot with the models. If you get the syntax right, then you can plug in the semantics at the beginning and it will go along for a free ride, and you get the right semantics out at the end because you have the right syntactical transformations.

There are certain famous problems, most famously Gödel's Theorem, but if we leave them to one side, the sophistication of our simulations in machine models of reasoning makes us forget the semantic content. But in real-life reasoning it is the semantic content that guarantees the validity of the inference, *not the syntactical rule*.

There are two important philosophical points to be made about the Lewis Carroll paradox. The first, which I have been belaboring, is that the rule plays no role whatever in the validity of the inference. The second is about the gap. *We need to distinguish between entailment and validity as logical relations on the one hand, and inferring as a voluntary human activity on the other.* In the case we considered, the premises entail the conclusion, so the inference is valid. But there is nothing that forces any actual human being to make that inference. You have the same gap for the human activity of inferring as you do for any other voluntary activity. Even if we convinced both Achilles and the Tortoise that the inference was valid as it stands and that the rule of modus ponens does not lend any validity to the inference, all the same, the tortoise

might still, irrationally, refuse to make the inference. The gap applies even to logical inferences.

I am not saying that there could not be any rules to help us in rational decision making. On the contrary there are many famous such rules and even maxims. Here are some of them: "A stitch in time saves nine." "Look before you leap." "He who laughs last laughs best." And my favorite, "Le coeur a ses raisons que la raison ne connaît pas." What I am saying is that rationality is not constituted as a set of rules, and rationality in thought as well as in action is not defined by any set of rules. The structure of intentional states and the constitutive rules of speech acts already contain constraints of rationality.

3. There is no separate faculty of rationality.

It should be implicit in what I have said that there cannot be a separate faculty of rationality distinct from such capacities as those for language, thought, perception, and the various forms of intentionality, because rational constraints are already built into, they are internal to, the structure of intentionality in general and language in particular. Once you have intentional states, once you have beliefs and desires and hopes and fears, and, especially, once you have language, then you already have the constraints of rationality. That is, if you have a beast that has the capacity for forming beliefs on the basis of its perceptions, and has the capacity for forming desires in addition to beliefs, and also has the capacity to express all this in a language, then it already has the constraints of rationality built into those structures. To make this clear with an example: there is no way you can make a statement without committing yourself regarding such questions as, "Is it true or false?" "Is it consistent, or inconsistent with other

things I have said?" So, the constraints of rationality are not an extra faculty in addition to intentionality and language. Once you have intentionality and language, you already have phenomena that internally and constitutively possess the constraints of rationality.

I like to think of it this way: The constraints of rationality ought to be thought of adverbially. They are a matter of the way in which we coordinate our intentionality. They are a matter of the way in which we coordinate the relations between our beliefs, desires, hopes, fears, and perceptions, and other intentional phenomena.

That coordination presupposes the existence of the gap. It presupposes that the phenomena at any given point are not causally sufficient to fix the rational solution to a problem. And I think we can now see why the same point operates for theoretical as well as for practical reason. If I hold up my hand in front of my face, there is no gap involved in seeing my hand, because I cannot help seeing my hand in front of my face if there is sufficient light and my eyesight is good. It is not up to me. So there is no question of such a perception being either rational or irrational. But now, suppose I refuse to believe that there is a hand in front of my face, even in this situation where I cannot help seeing it. Suppose I just refuse to accept it: "You say there's a hand there but I damn well refuse to accept that claim." Now the question of rationality does arise, and I think we would say that I am being irrational in such a situation.

I want to emphasize a point I made earlier. You can only have rationality where you have the possibility of irrationality. And with just sheer, raw perceptions, you do not get rationality or irrationality. They only come into play where you have a gap, where the existence of the

intentional phenomena by themselves is not sufficient to cause the outcome, and these are cases where you have to decide what you are going to do or think.

This is why people whose behavior is determined by sufficient causal conditions are removed from the scope of rational assessment. For example, not long ago I was in a committee meeting, and a person whom I had previously respected voted in the stupidest possible way. I said to him afterwards, "How could you have voted that way on that issue?" And he said, "Well, I'm just incurably politically correct. I just can't help myself." His claim amounts to saying that his decision making in this case was outside the scope of rational assessment, because the apparent irrationality was a result of the fact that he had no choice at all, that the causes were causally sufficient.

4. Weakness of will is a common, natural form of irrationality. It is a natural consequence of the gap.

On the Classical Model, cases of weakness of will are strictly speaking impossible. If the antecedents of the action are both rational and causal, and the causes set sufficient conditions, then the action has to ensue. It follows that if you did not do the thing you set out to do, then that can only be because there was something wrong with the way you set up the antecedents of the action. Your intention was not the right kind of intention,¹⁰ or you were not fully morally committed to the course you claimed to be committed to.¹¹

10. Donald Davidson, "How Is Weakness of the Will Possible?" *Essays on Actions and Events*, Oxford: Clarendon Press, Oxford University Press, New York, 1980.

11. R. M. Hare, *The Language of Morals*, Oxford: Oxford University Press, 1952.

I want to say, on the contrary, that no matter how perfectly you structure the antecedents of your action, weakness of will is always possible. Here is how: at any given point in our waking lives, we are confronted with an indefinitely large range of possibilities. I can raise my right arm, or I can raise my left arm; I can put my hat on top of my head, or I can wave it around. I can drink water or not drink water. More radically, I can walk out of the room and go to Timbuktu, or join a monastery, or do any number of other things. I have an open-ended sense of possibilities. Now, of course, in real life there will be restrictions set by my Background, by my biological limitations and by the culture that I have been brought up in. The Background restricts my sense of the possibilities that are open to me at any given time. I cannot, for example, in real life, imagine doing what St. Simeon Stylites did. He spent thirty five years on top of a pillar, just sitting there on a tiny platform, all for the glory of God. That is not an option that I could seriously consider. But I still have an indefinite range of real options that I am capable of perceiving as options. Weakness of will arises simply from the fact that at any point the gap provides an indefinitely large range of choices open to me and some of them will seem attractive even if I have already made up my mind to refuse them. It does not matter how you structure the causes of the action in the form of antecedent intentional states—beliefs, desires, choices, decisions, intentions—in the case of voluntary actions, the causes still do not set sufficient conditions, and this opens the way for weakness of will.

It is an unfortunate feature of our philosophical tradition that we make weakness of will out to be something really strange, really bizarre, whereas, I have to say, I

think it is very common in real life. I devote chapter 7 to this issue, so I will not say any more about it now.

5. Contrary to the Classical Model there are desire-independent reasons for action.

The fifth thesis of the Classical Model that I want to challenge has a very long history in our philosophical tradition. The idea is this: a rational act can only be motivated by a desire, where “desire” is construed broadly to include moral values that one has accepted and various sorts of evaluations that one has made. Desires need not be all egotistical, but for any rational process of deliberation there must be some desire that the agent had prior to the process, otherwise there would be nothing to reason from. There would not be any basis on which you could do your reasoning, if you did not have a set of desires in advance. Thus there can be no reasoning about ends, only about means. A sophisticated contemporary version of this view is found in the work of Bernard Williams,¹² who claims that there cannot be any “external” reasons for an agent to act. Any reason that is a reason for the agent must appeal to something “internal” to his “motivational set.” This, in my terminology, amounts to saying that there cannot be any desire-independent reasons for action.

I am going to criticize this view in great detail later, but at this point I want to make only one criticism. This view

12. “External and Internal Reasons,” reprinted in his *Moral Luck: Philosophical Papers 1973–1980* Cambridge: Cambridge University Press, 1981, pp. 101–113. Williams denies that his model is restricted to ends-means reasoning, but the other sorts of cases he considers, such as inventing alternative courses of action, do not seem to me to alter the basic ends-means structure of his model. See his “Internal Reasons and the Obscurity of Blame” reprinted in his *Making Sense of Humanity and Other Philosophical Papers*, Cambridge University Press, 1998, pp. 38–45.

has the following absurd consequence: at any given point in one's life no matter what the facts are, and no matter what one has done in the past or knows about one's future, no one can have any reason to do anything unless right then and there, there is an element of that person's motivational set, a desire broadly construed, to do that thing, or a desire for which doing that thing would be a "means" to that "end," that is, a means to satisfying that desire.

Now why is that absurd? Well, try to apply it to real-life examples. Suppose you go into a bar and order a beer. The waiter brings the beer and you drink it. Then the waiter brings you the bill and you say to him, "I have looked at my motivational set and I find no internal reason for paying for this beer. None at all. Ordering and drinking the beer is one thing, finding something in my motivational set is something else. The two are logically independent. Paying for the beer is not something I desire for its own sake, nor is it a means to an end or constitutive of some end that is represented in my motivational set. I have read Professor Williams, and I have also read Hume on this subject, and I looked carefully at my motivational set, and I cannot find any desire there to pay this bill! I just can't! And therefore, according to all the standard accounts of reasoning, I have no reason whatever to pay for this beer. It is not just that I don't have a strong enough reason, or that I have other conflicting reasons, but I have zero reason. I looked at my motivational set, I went through the entire inventory, and I found no desire that would lead by a sound deliberative route to the action of my paying for the beer."

We find this speech absurd because we understand that when you ordered the beer and drank it, if you are a

sane and rational person, you were intentionally *creating* a desire-independent reason, a reason for doing something regardless of what was in your motivational set when the time came to do it. The absurdity lies in the fact that on the Classical Model the existence of a reason for an agent to act depends on the existence of a certain sort of psychological element in his motivational set, it depends on the existence of a desire, broadly construed, then and there; and in the absence of that desire the agent has no reason, regardless of all the other facts about him and his history, and regardless of what he knows. But in real life the sheer knowledge of external facts in the world, such as the fact that you ordered the beer and drank it, can be a rationally compelling reason to pay for it.

The question, how is it possible that there can be desire-independent reasons for action, is an interesting and non-trivial question. I think most of the standard accounts are mistaken. I intend to devote extensive discussion to this issue later in this book, in chapter 6, so I will not discuss it further here.

There are really two strands to this aspect of the Classical Model. First we are supposed to think that all reasoning is about means not about ends, that there are no external reasons for action. And second, we are to believe a corollary, that the primary ends in the motivational set are outside the scope of reason. Remember that Hume also says, "'Tis not contrary to the dictates of reason, to prefer the destruction of the whole world to the scratching of my little finger." The way to assess any such claim is always to bring it down to real-life cases. Suppose the president of the United States went on television and said, "I have consulted with the Cabinet and the leaders of Congress,

and I have decided that there's no reason why I should prefer the scratching of my little finger to the destruction of the whole world." If he did this in real life we would feel he had, to use the terminology of Hume's era, "lost his reason." There is something fishy about Hume's claim and about the general thesis that one's fundamental ends can be anything whatever, and are totally outside the scope of rationality, that where primary desires are concerned, everything has equal status and is equally arbitrary. I think that cannot be the right way to look at these matters.

The thesis that there are no desire-independent reasons for action, that there are no external reasons, is logically closely related to Hume's doctrine that one cannot derive an "ought" from an "is." Here is the connection. "Ought" statements express reasons for action. To say that someone ought to do something is to imply that there is a reason for him to do it. So Hume's claim amounts to the claim that statements asserting the existence of reasons for action cannot be derived from statements about how things are. But how things are is a matter of how things are in the world as it exists independent of the agent's motivational set. So on this interpretation, the claim that how things are in the world cannot imply the existence of any reasons in an agent's motivational set (one cannot derive "ought" from "is") is closely related to the claim that there are no facts in the world, independent of the agent, that by themselves constitute reasons for action (there are no external reasons). Hume says, in effect, we cannot get values from facts; Williams says we cannot get motivations from external facts by themselves. The point of connection lies in the fact that the acceptance of a value is the acceptance of a motivation. However we interpret

both claims, I think they are both demonstrably false, and I intend to discuss this issue in some detail in the course of this book.

6. Inconsistent reasons for action are common and indeed inevitable. there is no rational requirement that rational decision making must start with a consistent set of desires or other primary reasons for acting.

The last point I want to take up is the question of consistency. As with the argument about weakness of will, this part of the Classical Model—the claim that the set of primary desires from which one reasons must be consistent—does not seem to me just a little bit false, but radically mistaken. It seems to me that most practical reasoning is typically about adjudicating between conflicting, inconsistent desires and other sorts of reasons. Right now, today, I very much want to be in Paris but I also want very much to be in Berkeley. And this is not a bizarre situation; rather it seems to me typical that we have an inconsistent set of ends. Given the extra premise that I know I cannot be both in Berkeley and in Paris at the same time, I have an inconsistent set of desires; and the task of rationality, the task of practical reason, is to try to find some way to adjudicate between these various inconsistent aims. Typically in practical reasoning you have to figure out how to give up on satisfying some desires in order to satisfy others. The standard way out of this problem in the literature is to say that rationality is not about desires as such but about *preferences*. Rational deliberation must begin with a well-ordered preference schedule. The problem with that answer is that in real life deliberation is largely about forming a set of preferences. A well-ordered set of preferences is typically the *result* of

successful deliberation, and is not its *precondition*. Which do I prefer, to be in Berkeley or Paris? Well, I would have to think about it.

And even after you have made up your mind, you decide "OK, I'm going to Paris," that decision itself introduces all sorts of other conflicts. You want to go to Paris, but you do not want to stand in line at airports, you do not want to eat airplane food, you do not want to sit next to people who are trying to put their elbow where you are trying to put your elbow. And so on. There are just all kinds of things that you do not want to happen, which you know are going to happen once you try to carry out your decision to go to Paris and to go by plane. By satisfying one desire you frustrate other desires. The point I want to emphasize is that there is a long tradition associated with the Classical Model, whereby inconsistent reasons for action, such as inconsistent obligations, are supposed to be philosophically odd or unusual. Often people in the tradition try to fudge the inconsistencies by saying that some of the apparently inconsistent obligations are not real honest-to-john obligations, but mere "prima facie" obligations. But rational decision making is typically about choosing between conflicting reasons for action, and you only have a genuine conflict of obligations where they are all genuine obligations. There is a serious question as to how there can be logically inconsistent but equally valid reasons for action, and why practical reason must involve conflicts between such valid but logically inconsistent reasons. I will take up this issue in more detail in subsequent chapters.

The aim of this chapter has been to introduce the subject matter of this book by laying bare some of the constitutive principles of the tradition I wish to overcome, and by

stating, in a preliminary way, some of my objections to the tradition. We began the chapter with Köhler's apes, so let's end with them. On the Classical Model human rationality is an extension of chimpanzee rationality. We are extremely clever, talking chimps. But I think there are some fundamental differences between human rationality and the instrumental reasoning of the chimpanzees. The greatest single difference between humans and the rest of the animal kingdom as far as rationality is concerned is our ability to create, recognize, and act on desire-independent reasons for action. I will explore this and other features of human rationality in the rest of this book.