

5

Some Special Features of Practical Reason: Strong Altruism as a Logical Requirement

I Reasons for Actions

I have been urging that in the investigation of rationality we should concentrate our attention on reasoning as an activity that actual selves engage in rather than focusing on rationality as an abstract set of logical properties. If we do, then it seems we find in any activity of reasoning a collection of intentional phenomena and a self that tries to organize them so as to produce another intentional state as the end product. In theoretical reason the end product is a belief or acceptance of a proposition; in practical reason it is a prior intention or intention-in-action. A consequence of the analysis of the intentionality of action that I gave in chapter 2 is that actions have intentional contents. So it is not at all mysterious that actions can be the result of a process of reasoning. Just as theoretical reason ends in a belief or an acceptance of a proposition, so practical reason ends in a prior intention to act or an actual action (which has the intentional content of an intention-in-action). Often, but not always, these are preceded by the formation of a secondary desire. For example: I look outside and come to the conclusion that it is going to rain.

Given my primary desire to stay dry and my other beliefs, I form the secondary desire to carry my umbrella, the prior intention to carry my umbrella, and I leave the house carrying an umbrella. Each of the last three steps, including the action itself, has intentional content motivated by the prior steps. I have heard people sneer at Aristotle's apparently quaint claim that an action can be the conclusion of a "practical syllogism." Aristotle was right, the sneerers are wrong.

I have been emphasizing the sense in which theoretical reason is a special case of practical reason: deciding what beliefs to accept and reject is a special case of deciding what to do. Though both theoretical and practical reason lead to a gap where the agent just has to act, reasons for acting are in many respects different from reasons for believing. Reasons for believing allow for conclusive proof, in a way that reasons for acting do not. This is a consequence of the difference in direction of fit. In this section I want to explore some of the special features of reasons for action and their consequences for practical reason. What is special about reasons for action? What are the differences between reasons for doing something and reasons for believing or accepting something? In both cases we have a set of intentional contents with the upward and downward directions of fit. Downward direction of fitters are supposed to be true, so they are responsible to states of affairs in the world. What sort of upward direction of fitters do we have and what are they responsible to? In the case of theoretical reason, the answer is relatively easy. To have a belief is to be committed to its truth, so if I am engaged in theoretical reason on the basis of my beliefs, I am committed to truth. Commitment has the world-to-mind or upward direction of fit

and commitment to truth provides a reason for acceptance of true propositions. To say that something is true implies that you ought to believe it. To spell this out in more detail: suppose I want to know whether to believe that *p*. Suppose I have conclusive proof that *p* is true. Since belief involves a commitment to truth and commitment has the upward direction of fit, I ought to believe (accept, recognize or acknowledge) that *p*.

Both practical and theoretical reason are subject to rational constraints, but reasons for action have some additional special features. First, reasons for action have a kind of first-person status that reasons for believing do not have. Reasons for believing are typically in the form of evidence or proof of the truth of the proposition believed, and truth is impersonal. Truth is a reason for anybody to believe. But where action is concerned, even if the reason is a reason for anybody, reasons for action still must appeal to something inner or first-personal in a way that reasons for believing do not. Once you have established *truth* there isn't any further question about whether you should believe it, because to have the belief that *p* is true is already to have the belief that *p*. But because of the difference in direction of fit between belief and intention, there is nothing analogous to truth where reasons for acting are concerned. In theoretical reason, the right reasons get you to a belief that is true. In practical reason the right reasons get you to an intention that is ... what? There is no *x* such that intention is to *x* as truth is to belief. Everyone has a reason for seeking self-preservation, flourishing, autonomy, and a whole lot of other desirable goals. But none of these stands to action as truth stands to belief, because in every case the goal has to be represented by the agents' intentional contents as a separate goal. In the case

of belief, the goal of truth is built into the belief. No such goal is built into reasons for acting, prior intentions, or intentions-in-action.

Second, reasons for acting have a special relation to time that is unlike that of reasons for believing. Reasons for acting are always forward-looking. And this is true even in cases where we are giving reasons why an agent acted as he or she did in the past. A present reason for acting is always a reason for a self to perform an action either *now* or *later*. A past reason for action was a reason in the past for a self to perform that action then or later.

Related to these two is a third point. Reasons for action must be able to motivate an action. If the reason is given why a past action was performed, then the reason must have functioned causally in the performance of the action, because it must have been the reason that the agent *acted on*. If the reason is for a future action, then it must be a reason that the agent *can act on*. But to say that is to say that the reason is either actually or potentially effective, because the notion of acting on a reason, as we saw, is the notion of making the reason effective in the performance of the action. In the last chapter I called attention to the motivational feature of reasons in order to argue that every total reason must contain at least one motivator.

What sorts of factitives can be motivators? The answer to that question given by the Classical Model is brutally simple: all motivators are desires, where "desire" is broadly construed to include such things as the goals, ends, and objectives of the agent. Reason is and ought to be the slave of the passions. Recent authors are somewhat vague about what the list of motivational entities would include, and they talk generally about "pro-attitudes" (a

term invented, I believe, by Patrick Nowell-Smith)¹ and the “subjective motivational set” (Williams),² but the general idea is clear enough. Without some kind of desire-like internal psychological state, the process of reasoning could never produce an action. Köhler’s chimpanzees are the model. Without desire they would never get off the ground.

Why are the classical theorists so confident about this model? Well, its simplicity is appealing and makes its features nicely formalizable in decision theory. But there are also powerful philosophical reasons in support of it. First, in real life, a lot of cases are like that. The simplest cases are where the reason simply is a desire of a certain sort. “Why are you drinking water?” Because I am thirsty. Another sort of case is where there is some fact that the agent believes will lead to the satisfaction of his desire. “Why are you drinking water?” Because it will cure my headache. Full story: I want to cure my headache, I believe that drinking water will cure my headache, therefore I want to drink water. In such a case the desire to drink water is itself a motivated desire, motivated by another desire together with a belief about how to satisfy that desire.

Another argument for the Classical Model is that in the structure of actual deliberation the conclusion must be some desire-like intentional state such as a secondary desire, a prior intention, or an intention-in-action. And where could that state rationally come from if not from an earlier desire? Without a desire or pro-attitude as a starting

1. Patrick Nowell-Smith, *Ethics*, London: Penguin Books, 1954, p. 112.

2. “External and Internal Reasons,” reprinted in *Moral Luck*, Cambridge: Cambridge University Press, 1981, pp. 101–113.

point, it seems there is no way that deliberation could rationally end in a desire or desire-like intentional state.

The obvious objection to the claim of the Classical Model that only desires can motivate is that there are many motivationally effective reasons for action, such as obligations, that are not desires. "Why are you drinking water?" Because I am under an obligation to do so. I promised my spouse.

To all of these examples the classical theorist gives the same answer. Your obligation, for example, is only a reason for action because you *desire* to fulfill your obligations. One of the central points in dispute between me and the Classical Model is exactly on this issue. On my view the obligation is—or at least can be—the reason for an effective desire (i.e., a desire the agent acts on), rather than a prior desire functioning as a reason for the effectiveness of the obligation. I will come back to this point in the next chapter.

A fourth feature of reasons for acting is that if the reason is taken as a reason for the performance of a free action, it cannot be taken by the agent as causally sufficient. If he thinks of himself as truly compelled, then he cannot think of himself as freely acting on a reason. In the case of human actions, because of the gap, the reason can be a good or adequate reason without providing causally sufficient conditions for the act. And, more important from the agent's point of view, the reason must not be seen as causally sufficient. As I remarked in earlier chapters, the applicability of the concept of rationality in decision making presupposes free choice. Indeed, for rational agents free choice is both necessary and sufficient for the applicability of rationality. Free choice implies that the act is rationally assessable, and rational assessability implies

free choice. It might seem that there are plenty of counterexamples to this claim. "What about the drug addict who cannot help himself but nonetheless is capable of rationality in selecting the rational means, rather than irrational means, to satisfy his craving?" But even this case supports the general point, because we are tacitly supposing that the drug addict has a choice of the means to satisfy his overwhelming desire. That is, to the extent that we regard the agent as acting rationally we are supposing that to that extent he is making free choices, even though the overall project of satisfying his addiction is not a matter of free choice for him and thus falls outside the scope of rationality. The gap is a feature of both reasoning about what to believe, and reasoning about what to do. But it plays a special role in reasoning about what to do, as I have tried to describe.

So, to sum up: in addition to the two general constraints of *rationality*, (together with justification), and *the gap*, which apply to reasons for believing as well as reasons for doing, there are at least three additional special features of reasons for action. They are, in a special sense *first-personal*, they are essentially *future-directed*, and they are essentially *motivational* in the sense that they must be capable of motivating an action. Just to have some grand words, let us call these five the conditions of Rationality, Freedom, Subjectivity, Temporality, and Causation.

Why should all of these hang together in the way that they do? Why are there these connections? At one level, I do not think that is a difficult question. Rationality is a biological phenomenon. Rationality in action is that feature which enables organisms, with brains big and complex enough to have conscious selves, to coordinate their intentional contents, so as to produce better actions than

would be produced by random behavior, instinct, tropism, or acting on impulse. To get the biological advantages of rational behavior, the animal has to have its own conscious motives (Subjectivity), some of these have to be forward-looking (Temporality), they have to be able to motivate real behavior in the form of bodily movements (Causation), and they have to do it under the presupposition of freedom operating in the gap (Freedom). “Practical reason” is the name of that capacity for coordination. Indeed, these features are not logically independent: the first two features, Subjectivity and Temporality, follow from the third feature, motivational Causation. A motive has to be someone’s motive (Subjectivity) to act now or in the future (Temporality).

The connection between rationality and the gap of freedom is this: *rationality applies only where there is free choice, because rationality must be able to make a difference.* If my actions are really completely *caused* by my beliefs and desires, so that I really can’t help myself, then I have no choice and rationality can make no difference at all to my behavior. If I am in the grip of causally sufficient conditions, there is no room for deliberation to operate and my action falls outside the scope of rational assessment. Furthermore a demand for justification makes sense only in cases where alternative possibilities were open to the agent.

II Constructing a Rational Animal

To illustrate the special role and character of practical reason, I would like to present the following thought experiment. Imagine that you are designing and building a robot that will be a “rational animal.” The point of the

thought experiment is to illustrate the logical relations between certain crucial features of human existence. Whatever else we are, we are the products, at least metaphorically speaking, of a certain sort of engineering. I do not believe it was the divine engineering of the creationist story, but rather as far as we know it was the unintentional, metaphorical, "as-if" engineering of evolutionary processes. But one way or another, we are the result of a certain set of processes that have been guided by certain sorts of design needs. Given that we are the products of engineering, even if only "as-if" engineering, the point of asking the question how rational beings might be designed is to get us to see how much you need to put into your design in order to see how much you can get out as a result of what you put in. What do you require as an actual design feature, and what do you get for free? (Many of the questions in the history of philosophy are contained in this question, by the way.) Because rationality is not a separate faculty or module, but rather a feature internal to other cognitive and volitional capacities, I believe that we will find that we have to put in most, though not all, of human mental faculties in order to have a "machine" capable of rationality.

The first feature you have to put into your robot is consciousness. You have to build a robot brain that has the power of human brains to cause and sustain inner, qualitative, unified, subjective states of awareness and sentience. Without consciousness you cannot get into the game of rationality at all. But passive perceptual consciousness is not enough. You need the active consciousness of agency. That is, you need to build a being that is consciously able to initiate actions. But in order to do that, the robot must have desires as well as intentions. This is

because it must be able to want to do the things it tries to do. So at a bare minimum we have to have a machine capable of perception, action, and desire. Furthermore, if these actions are to be rational actions, the robot has to be able to engage in deliberation. This requirement is a more weighty matter than it might at first seem. I do not see how a robot could engage in deliberation without a very large chunk of the human and animal apparatus of intentionality. First, there must be the capacity to store information in the form of memories, and this memory capacity will be a source of beliefs. Second, it must have the ability to coordinate both the downward direction-of-fitters (beliefs, perceptions, etc.) and upward direction-of-fitters (desires, inclinations, etc.) in a conscious stream of thought. That is, it is not enough to have perceptions, memories, desires, and intentions; the robot also must be able to put all this apparatus to work in a conscious sequence of deliberative thoughts. It has to be able to think that because so and so is the case, and it wants such and such, it should do this act and not that act, even if it can think these thoughts only wordlessly. In order that it should have all this intentionalistic apparatus it must have what (in chapter 2) I call the Background, the set of pre intentionalistic capacities that enable it to interpret and apply its own intentional states. Finally, the robot must be such that the stream of thought is capable of ending in decisions and subsequent actions.

So the additions we had to make to the robot after giving it consciousness were quite substantial: The robot has to have conscious perceptual phenomena, conscious conative phenomena (desires), and conscious volitional phenomena (both prior intentions and intentions-in-actions), and it has to have the capacity for conscious deliberation result-

ing in decisions and actions, with all the apparatus that such a deliberative process involves. The way that I have described the case, we have already built the experiences of the gap into the robot. And because it has all of these features, as I noted in chapter 3, it already has a self in my sense. Selfhood in my sense comes for free once you have a conscious intentional being capable of engaging in free actions on the basis of reasons. Now a crucial question is raised immediately. Once the robot has all of that, does it already have the mechanism necessary for rational decision making of the fully human variety? Well, not quite. So far we have not built a humanoid robot, but, one might say, an artificial chimpanzee. To get to human decision-making powers we need to put in certain other features.

Once you have both conscious and unconscious mental states and processes together with both downward (perceptions, memories, beliefs, etc.) and upward direction-of-fitters (desires, inclinations, intentions, etc.) and you have the capacity to coordinate all this in the stream of conscious thought ending in decision making, the next central element to build into the robot is, without doubt, language. It is important to say exactly what features of language would be required by a rational agent. An animal does not require any language in order to have simple intentional states like hunger and thirst, and it does not even have to have language in order to make simple decisions, nor indeed does it need a language to engage in simple instrumental reasoning of the sort that Köhler's chimpanzees engaged in. But for full-blown rationality, certain very specific features of language are essential. Not all the features of natural human languages are essential to rationality. For example, rational thought processes do

not require color words, the passive voice, or definite articles. But fully human rationality does need certain essential linguistic devices. First, our robot must have the basic speech act forms that relate language to reality with both the word-to-world direction of fit, and the world-to-word direction of fit. It must, at the bare minimum, have the capacity to represent how things are in the world (assertives), as well as the capacity to represent how it tries to get others to act in the world (directives), and how it commits itself to act in the world (commissives). Furthermore, it must have the capacity to communicate all of this to other possessors of language. Language is both to think with and to talk with, but when we are concerned with talking, we have to have a language that is public, that enables the robot to communicate with others. Because we are building this robot in our own image, so to speak, we will build it with the capacity to communicate with us. Furthermore, it seems to me the robot has to have some set of devices for representing temporal relationships. If it is going to be able to plan for the future, which is characteristic of practical reason, it has to be able to represent the future and its relation to the present and the past. What else would it need? Well, it seems to me it would have to have some way of articulating logical relations. It need not have precisely our inventory of logical vocabulary, but it must have some way of marking negation, conjunction, implication, and disjunction. Furthermore, it seems to me it would also need some set, however minimal, of metalinguistic terms for appraising success and failure in achieving direction of fit, and logical coherence. So it needs something in the range that includes "true" and "false," "valid" and "invalid," "accurate" and

“inaccurate,” “relevant” and “irrelevant.” Now that we have given it this much of a language we might as well give it a name. Call it “the Beast.”

In the course of constructing all of this representational apparatus, both mental and linguistic representations, we will have had to have given the Beast the apparatus necessary to apply these representations to concrete situations and to interpret the representations that it receives from other sources. These abilities, the abilities to apply and interpret representations, constitute what I have been calling the Background.

Now here is the point of the thought experiment: once the Beast has this much, it already has the apparatus essential for the distinctively human features of rational thought processes and rational behavior. It has a form of rationality that goes far beyond the rational chimpanzees we discussed in chapter 1. Specifically, once the Beast has the ability to perform speech acts, it has the potential for desire-independent reasons for action, indeed it inevitably has the requirement of desire-independent reasons for action, because just about every speech act involves a *commitment* of some kind or other. The famous examples are speech acts like promising, where the speaker is committed to carrying out a future course of action, but asserting commits the speaker to the truth of the proposition asserted, and orders commit the speaker to the belief that the person to whom he or she gives the order is able to do it, to the desire that he or she should do it, and to permitting the hearer to do it. In short, what people have thought of as the distinctive element of promising, namely commitment or obligation, actually pervades just about all speech acts. The only exceptions I can think of would

be simple expressives like "Ouch!" "Damn!" or "Hurrah!" and even they commit the speaker to having certain attitudes.

The bizarre feature of our intellectual tradition, according to which no set of true statements describing how things are in the world can ever logically imply a statement about how they ought to be, is that the very terminology in which the thesis is stated refutes the thesis. Thus, for example, to say that something is true is already to say that you ought to believe it, that other things being equal, you ought not to deny it. The notion of a valid inference is such that, if p can be validly inferred from q , then anyone who asserts p ought not to deny q , that anyone who is committed to p ought to recognize his commitment to q .

The point of the thought experiment can also be put as follows: once you have the apparatus of consciousness, intentionality, and a language rich enough to perform the various types of speech acts and express various logical and temporal relations, then you already have the apparatus necessary for rationality. Rationality is not an extra module or faculty. It is already built into the apparatus that we have described. Furthermore, something much richer than instrumental or ends-means rationality is already built into the apparatus we have described, because we have the potential, indeed the requirement, of desire-independent, or external, reasons for action.

We have included in the Beast the experiences of the gap. But have we given it genuine free will, or only the illusion of free will? There are at least two different possibilities. First we might deceive the poor Beast by making its underlying mechanisms totally deterministic. So it has the illusion of free will, because it experiences the gap, but

in fact its behavior is entirely preprogrammed with fully deterministic mechanisms. Another quite distinct possibility is that its conscious experience of decision making in the gap is matched by an indeterministic element in the hardware implementation that is carried forward through time by the conscious level of decision making. I explore both of these possibilities, as far as actual human beings are concerned, in chapter 9.

III Egoism and Altruism in the Beast

Well, what about the favorite topics of moral philosophers, egoism and altruism? How do they stand with our robot? We have not yet explicitly built either egoism or altruism into the Beast. In our intellectual culture we take egoism and self-interest as unproblematic, and regard altruism and generosity as requiring a special explanation. In one way that is right, in another it is wrong. It is right to suppose that the Beast will prefer the satisfaction of its desires to their frustration, and will prefer the alleviation of its pains to their intensification. Other things being equal, that is part of what is involved in having a desire or a pain. And the concern with its own desires, etc. looks like egoism. But in another sense it is wrong to think of egoism as unproblematic, because satisfaction of the desires does not so far tell us the *content* of the desires and so far we have said nothing about the content of the desires in the Beast. It might well be that the Beast finds altruistic desires as natural as egoistic desires. As far as what we have said goes, the Beast might prefer the prosperity of others to its own prosperity.

So let us add another component to our Beast. Let us suppose that we program it to seek what I will vaguely

call "self-interest." Let us build into our Beast a preference for survival over extinction, and a preference for its self-interest over what is not in its interest, that is, we suppose that the Beast does not wish to become injured, damaged, diseased, deprived, or dead. Once the Beast has a self and self-interest, if it also has a conception of time, as we have stipulated, then it will be able to plan for its subsequent survival and flourishing. That is, if the self has interests, and if the self persists through time, and if the self is the agent that exercises rationality, then it will be rational for the self to make plans now to secure its interest in the future, even though it has no present desire to do the things now that are necessary to secure its interests in the future. So we have now two forms of desire-independent or external reasons for action. Roughly speaking, there are commitments, typically made to others, but they can be made to oneself, as well; and there are prudential reasons.

Rational self-interest in our enlightened robot does not come for free, but it does not require much of a technological investment beyond the bare minimum necessary for consciousness, intentionality, and language. If the Beast has needs and interests and the capacity to recognize these needs and interests, and has a self and an awareness of its self extending into the future, it is not much of an addition to give it a motivation for acting now so as to look out for its interests in the future.

Now we come to a crucial question: does the Beast have any rational basis for caring about the interests of others? What is the relation between the self-interest that we have built in and the altruism that we have neglected? The standard approach to this question by moral philosophers is to try to build altruism out of egoism. There are, if I understand them, at least three ways of doing this. First,

we imagine that we simply do it as an engineering task. We put altruism into our Beast, just as we have already put egoism into the Beast. This is one way of interpreting the sociobiologists. The idea is that we are genetically inclined to at least certain forms of altruism, and we are supposed to be able to account for the genetic basis of altruism through such things as group selection or kin selection. Altruism is just a natural inclination, and insofar as it can be effective at all, it can be just as effective as any other internal reason. Our Beast simply has an inclination to look out for the interests of others. Second, and more interesting, an effort has been made by Thomas Nagel³ to show the formal similarity between prudential reasons and altruistic reasons. To consider the interests of others is just as rationally based as considering one's own future interests. Third and finally, an effort has been made in the Kantian tradition, most notably by Christine Korsgaard,⁴ to derive altruism from autonomy. If, because of my autonomy or freedom, I have to will my own actions; and if the will is subject to constraints of generality such that I am rationally required that each thing I will, I should be able to will as a universal law; then I will be rationally required to treat other people as my equals in the moral realm, because the universal laws that I will apply equally to me and to them.

There is something right about all three of these approaches, but also something unsatisfactory. If I just feel an inclination to altruism, then that is much too fragile to form the basis for practical reason where altruism is concerned. The inclination to altruism has no special

3. *The Possibility of Altruism*, Princeton: Princeton University Press, 1970.

4. *The Sources of Normativity*, Cambridge: Cambridge University Press, 1996.

binding force. Often one does not feel such inclinations, and many people feel counterinclinations, such as an inclination to sadism, cruelty, or indifference. And on this account, altruism would just be one inclination among others. What is special about the inclination to help others? So let's turn to Nagel's analogy between prudence and altruism. The point that is true seems to me to be this: once I have consciousness and the self and am able to use language, I am already committed to the existence of other consciousnesses and selves on a par with my own. How exactly? That there is such a thing as my conscious self makes sense to me only if it is different from other things in the universe. If there is a me, then there must be a not-me. And if the not-me entities in the universe include entities with which I communicate in the performance of speech acts, then some of the not-me's in the universe must be presupposed by me to be conscious agents with a selfhood just like my own. So I am one self among others. But the question still remains, why should I care about the others? There is indeed a formal similarity between caring about my future self and caring about another self: in both cases I have to consider the interests of entities that are not present to my consciousness here and now when I am making the decisions. But there is a drastic asymmetry: in prudential reasoning, the self I care about is me. That is, the self that makes the decisions and carries out the actions is identical with the beneficiary of the decisions and actions. For altruistic reasoning, that identity is lost. I am not attempting here to do full justice to Nagel's subtle argument. I am simply raising a difficulty that I find with it, before going on to discuss another argument for the same conclusion, and then to present my own.

Let us then turn to examine Korsgaard's Kantian account of how autonomy generates universality and

universality generates altruism. Her solution is presented as an interpretation of Kant's views and here is how it goes: Kant argues that (1) we have to act under the presupposition of our own free will. He then continues that (2) free will, if it is to be a will at all, must be determined in accordance with a law. Since, therefore, (3) free will has to be determined under its *own* law (by 1), it turns out that (4) the Categorical Imperative is a law of free will.⁵ The dubious step here is the second step. Why should the exercise of my free will in decision making require any sort of law at all? Why can't I freely decide what to do, just like that? Certainly no argument so far has been presented why there must be a law in order for me to make free rational decisions.

To answer this objection Korsgaard draws an analogy with causation. She says causation has two components, the notion of making something happen, and the notion of a law. We require the second component, a law, because we could not properly *identify* a case of something making something else happen if we could not assume it under a causal law. That is, she thinks regularity is necessary for the identification of causation. Then she claims the causation of the will is exactly analogous to causation in general. For if I am to act of my own free will, then I am the cause of my actions. But if that is the case, I must be able to distinguish between *myself* causing the action, and some *desire* or *impulse* that is in me that causes my body to move. I have to see myself as something distinct from my first-order impulses and desires. But if that is the case, in order that the actions should genuinely be my actions, that is, that they should come from myself rather than just be expressions of my first-order desires, I have to act

5. Korsgaard, *Sources of Normativity*, pp. 221–222.

under some universal principles. So the law that I create for myself is exactly analogous to the laws of causation. We could not *identify* acts as the acts of a self unless they were done under some universal principle. In order that the actions can be truly said to be actions of myself, it turns out that I must be a law-giving agent. Indeed it is only because we impose universal volitional principles on our decisions that we can be said to have a self at all. The self is constituted by these universalized decisions. For Korsgaard the key sentence, I believe, is the following: "For if *all* of my decisions were particular and anomalous, there would be no identifiable difference between *my acting* and an assortment of first-order impulses being causally effective in or through my body. And then there would be no self—no mind—no me—who is the one who does the act" (p. 228).

I believe this argument does not work. The basic notion of causation is, indeed, the notion of making something happen. And it is true that in order to *identify* such cases, we have to presuppose regularity. But that requirement is an *epistemic* requirement, not an *ontological* requirement on the very existence of causation. There is no self-contradiction in imagining causes that occur without instantiating any universal regularities. We might not be able to establish with certainty that such and such an event was really the cause of such and such other event unless the experiment were repeatable, unless we could test the individual case by seeing if it instantiated a regularity. But that is a matter of finding out for sure; it is not a matter of the very existence of the relation whereby one thing made another thing happen. Real-life examples make clear the distinction between causation and regularity. When, for example, we investigate the causes of the

First World War, we are trying to explain why it happened. We are not seeking universal regularities. We have to make a Background presupposition of at least some degree of regularity in order to conduct the investigation at all, and without the possibility of causally sufficient conditions and repeatable experiments we may never be completely sure of our answer. But the requirement of regularity is an epistemic requirement for the *identification* of causes; it is not an ontological requirement on the very existence of the relation by which one event makes another happen.

Indeed, the requirement of regularity is an epistemic requirement on just about any notion that has application to the real world. In order to identify something as a chair or a table or a mountain or a tree, we have to presuppose some kind of regularity in its characteristics or uses. Regularity is essential for the identification of an object as a chair, but we should not on that ground say that the notion of chair really contains two components, an object that functions for people to sit in, and a regular principle. Rather we should say a chair is an object that people use to sit in, and like other notions referring to objects, causes, etc., the concept of a chair requires a Background presupposition of regularity.

If we extend the relation of regularity to causation in the case of human beings, we can say that from the third-person point of view it is indeed an epistemic requirement on my *recognizing* somebody's decisions as truly his considered decisions, as opposed to his capricious and whimsical behavior, that they have some sort of order and regularity. But it does not follow, that in order to *be* his decisions, they have to proceed from a universal law that he makes for himself. That is to say, the passage that

I quoted makes a false dichotomy between acting on impulse, which is supposed to be not free, and acting on a universal law, which is free. But acting on impulse can be as much free as acting on a universal law. Korsgaard says that there would be no *identifiable* difference between an unfree act and a capricious act, if all of a person's acts were capricious. But if this point is true, it is still only a third-person epistemic point. From the outside, someone looking at me might not be able to tell which of my actions were truly free if I always acted on impulse. But from the inside, from the first-person point of view, acting on impulse can be as much a free act as acting on sober reflection. Some very cautious persons restrain themselves from ever acting on impulse, whereas free spirits often allow their impulses to move them. The experience of the gap can be the same in both cases. And the one is as much or as little constitutive of the self as the other, because in both cases a self is required to make the decision what to do.

Korsgaard's argument presupposes (1) that in order for the self to make decisions at all, it must make them in accord with a universal principle; and that presupposition itself presupposes (2) that acting on principle is somehow constitutive of the self. I am rejecting both of these claims. Kant was wrong: free action does not require acting according to a self-created law. And the self that engages in free action does not require universal principles in order to be a self. On the contrary, both consistent and capricious behavior in the gap, as I argued in chapter 3, require a *preexisting* self. In short there is no logical requirement whatever that in order for my acts to be free acts, and freely chosen by myself, that they have to

exemplify universal principles. My acts can be absolutely capricious and still be free acts.

This is not the place to try to give a full diagnosis of Korsgaard's powerful philosophical argument, but—all too briefly—I think the source of her mistake is that she wants a gap filler. She wants the self to be the cause of free actions. If you accept that requirement, then on certain natural assumptions, the rest follows. The steps are these: (1) Free actions are caused by the self. (2) But the self in causing must instantiate a law, and the only laws that it could instantiate are self-created. (3) In creating a law the self creates itself as a self.

I am rejecting all of these. If by "cause" we imply "causally sufficient conditions," then free actions are not caused by anything. That is what makes them free. To put this point more precisely: What makes an action free at the psychological level is that it does not have antecedently sufficient psychological causal conditions (see chapter 3 for the argument). The self *performs* the act, but it does not *cause* the act. Nothing fills the gap.

IV The Universality of Language and Strong Altruism

Well, let's take stock of where we are. We were trying to answer this question: given that the Beast has been programmed to look out for its own self-interests, is there any logical requirement on it to pay any attention to the interests and needs of other people at all? The words "altruist" and "egoist" get bandied about without much clear definition, so let's try to define them for this discussion. In one sense an egoist is someone who cares only about his own interests and an altruist is someone who cares about the

interests of others. But that definition obscures a crucial distinction. An altruist might be someone who is naturally inclined to care about the interests of others, but for such an altruist acting altruistically is just acting on one inclination among others. He likes to help others the way he likes to drink beer, for example. Let us call this the weak sense of "altruism." But there is another stronger sense of "altruism" that we are trying to get at. An altruist in this sense is someone who recognizes the interest of others as a valid reason for acting even in cases where he has no such inclination. The question is: are there rationally binding *desire-independent* altruistic reasons for action? An altruist in the strong sense is someone who recognizes that there are rationally binding desire-independent reasons for him to act in the interests of others. Both Nagel and Kant-Korsgaard gave arguments to support the rational requirement of altruism in this strong sense. The sociobiologists only answer the question for the weak sense. I have rejected both the Nagel and the Kant-Korsgaard arguments. But I think their conclusion is right, and I think Kant-Korsgaard is right to see that the issue is one of generality. Granted that the Beast and ourselves have reasons to behave egoistically, is there a generality requirement that would extend those reasons to other people in a way that binds our behavior? I think there is.

The generality required to support strong altruism is already built into the structure of language. How exactly? Let us go through the steps to see how language introduces rationally required forms of generality. Both my dog and I can see that a man is at the door, that is, we can both have a visual experience that I describe in words as "seeing that a man is at the door." But there is a big difference in that if I *say* I see a man at the door in language I

am committed to a kind of semantic categorical imperative that has no analogue in the dog. When I say, "That is a man," I am committed to the claim that any entity exactly like that in the relevant respects is also correctly describable as "a man." To put it in Kantian jargon: assertions are bound by the semantic categorical imperative: so assert that the maxim of your assertion can be willed by you as a universal law binding on all speakers. And the maxim is provided by the truth conditions of the proposition asserted. In this case: an object that has those features satisfies the truth conditions for "man."

When you make an assertion of the form *a* is *F*, rationality requires that you be able to will that everyone in a similar situation should assert that *a* is *F*. That is, because the predicate is general, its application requires that any user recognize its generality. Any user of language, in the Kantian formulation, has to be able to will a universal law of its application to relevantly similar cases.⁶

Furthermore, this imperative, unlike some of Kant's, actually meets Kant's condition that the insincere or dishonest person is involved in some kind of self-contradiction when he attempts to will his maxim as a universal law. Thus, suppose I am lying when I say, "That is a man," then I cannot will a universal law that everybody in a similar situation should say, "That is a man," for if they did, the word "man" would cease to have the meaning it does. That is, I cannot consistently conjoin my will

6. Of course, neither in my case nor in Kant's does the ability to will a universal law require that the agent think that it would be a good thing if everybody behaved the way he did. That is not the point at all. It would be at the very least boring and tiresome if everybody in my situation were to say "that is a man." The point of the categorical imperative is logical; there is no logical absurdity in my willing the maxim of the action as a universal law binding on all speakers.

that my utterance be a lie together with my will that the semantic content apply universally according to the semantic categorical imperative.

To put this point without the Kantian apparatus, we can say that any assertion by a speaker *S* of the form *a* is *F* commits *S* to a universal generalization: for any *x*, if *x* is relevantly type-identical to *a*, then *x* is correctly described as "*F*." We are here not talking about entailment relations between propositions, but rather about what a speaker is committed to when he performs a speech act.

Furthermore, the generality requirement applies to other people. For if I am committed to recognizing similar instances as also cases of men, my commitment in a public language requires that I think other people ought also to recognize this and similar cases as cases of men. That is, the generality is built into the structure of language itself, and indeed when it comes to the application of language, it looks as if we get ought's from is's wherever we turn. From the fact that an object is truly described as "a man," it follows that you ought to accept relevantly similar objects also as men, and that other people ought both to accept this as a man and other relevantly similar objects as men. It is impossible to use language without these commitments. I have put this in a grand-sounding terminology, but it is a trivial consequence of the nature of language and speech acts.

The way we get generality into reasons for action in the form of strong altruism is by simply noticing that the generality requirement that works for such predicates as "man," "dog," "tree," and "mountain" also works for "has a reason for action" and other such motivators. I will show this with an example. Suppose I have a pain, and I seek to alleviate my pain. There is a difference between me seeking to alleviate my pain, and my dog's alleviating his

pain by licking his wound. What is the difference? Well, at least this much: I can bring my pain under certain universal generalizations, simply by characterizing it with a word such as "pain." That is, the same feature we found in the discussion of the word "man" will also apply to the word "pain." If I assert "This is a pain" I am committed to the claim, "For all x , if x is relevantly like this, x is a pain."

The generality of language, given certain commonsense assumptions about my own self-interests, will generate strong altruism. I will first put the point in intuitive form and then recast it in a semantic form. Intuitively it seems reasonable to suppose that if I am in pain I have a reason for wanting to alleviate my pain. My feeling this degree of pain involves feeling a need for its alleviation. My need for pain alleviation is for me a reason to alleviate my pain and I even believe that others, where they have the ability and the opportunity, have a reason to help alleviate my pain. But I cannot believe that they have a reason for helping me, without committing myself to believing that in the same situation where the pronouns are reversed, I am bound to recognize that I have a reason for helping them. It is rational of me to want them to help me, for the reason that I am now in need of help. But then in consistency when they are in need of help I am committed to recognizing the existence of their need as a reason for my helping them.

The way the generality of language works to produce strong altruism is as follows:

1. I am in pain, so I say "I am in pain." Because I said "I am in pain" I am committed by the generality requirement to recognize that in a similar situation you would be in pain. Because "pain" is a general term in the language, the truth conditions apply indifferently to you and me. I am

committed to applying the open sentence, "X is in pain" to any object that satisfies exactly these conditions.

2. My pain creates a need. Because I am in pain I need help. I am aware of both my pain and my need. So I say, "I need help because I am in pain." Now notice that this is not to be interpreted as a plea for help. It is not an indirect speech act; rather it is a statement made by me about me. The same generality requirement applies again. I am now committed to recognizing that in a similar situation with a reversal of the pronouns if you are in pain, you would need help. I am committed to applying the open sentence "X needs help because X is in pain" in any type-identical situation.

3. I am in pain and need help, and I believe that my need for help is reason for you to help me. So suppose I say, "Because I am in pain and need help, you have a reason to help me." The same generality requirement is in force. I am committed to the universal, for any situation that is relevantly type-identical to this one:

For all x and for all y , if x is in pain and x needs help because x is in pain, y has a reason to help x .

But that commits me to recognize that when you are in pain I have a reason to help you. Notice that we are talking here about the speakers' commitments in the performance of speech acts. We are not at this point concerned with truth or with entailment relations between propositions; rather we are worried about what the speaker is committed to when he or she makes an assertion of this form.

The point for the present discussion is that once we have programmed the Beast in the way that I described,

that is, in addition to basic mental capacities we give it the gap, self-interest, and language, then we have already given it a sufficient logical ground for strong altruism. Notice further that we require no heavy-duty metaphysics. No noumenal world or Kantian Categorical Imperative is necessary. All this argument requires is that we, other people, and the Beast can speak English or some other language, and that we make reasonable self-interested claims. We claim, for example, that our needs are sometimes a reason for someone else to help us.

But why couldn't we block the argument by saying, for example, that my case is special. I deserve special treatment, not accorded to others. One can always make such a claim but to do so goes beyond the semantics of the indexicals. There is nothing in the semantics of "I," "you," "he," etc. that blocks the commonality of truth conditions for "pain," "need," "reason," etc. I am not here trying to eliminate the possibility of special pleading or bad faith. The history of the world is full of people, tribes, classes, nations, etc. who cheat by claiming a right to special privilege, and nothing I say will stop such people from cheating. My point is rather that the universality constraint that gets us from egoism to strong altruism is already built into the universality of language. All we have to assume is that the Beast has certain reasonable self-interested attitudes about its relations with other conscious beings and that it is prepared to state them in language. Once the Beast or anyone is prepared to say "You have a reason to help me because I am in pain and need help," then it is committed, in type-identical situations, to applying universal quantifiers to the open sentence " y has a reason to help x because x is in pain and needs help," because the use of the general terms commits the speaker

to the application of those terms to situations that share the general features that the initial situation had. Language is by its very nature general.

To the extent that one resists this conclusion, I think the resistance comes from another pervasive mistake in our culture: the idea that language cannot be all that important, because it is mere words. How can the mere utterance of words commit me to anything? I encountered this same resistance a generation ago when I showed how to derive "ought" from "is."⁷ Many commentators felt the mere fact that I uttered words can't commit me to anything. There must be some extra moral principle involved or some endorsement of the institutions of language. Or something!

I will have more to say about these issues in the next chapter, but for the moment, we can say the problem is not to see how the utterance of words can commit me, but rather to see how anything *other than* the utterance of words could commit me. The paradigm forms of commitment to courses of action are in the performance of speech acts.

V Conclusion

I have had three main aims in this chapter. I have tried to describe some special features of reasons for action; I have tried to describe what features are necessary for a self-agent to be capable of rationality; and I have tried to derive the principles of strong altruism from the universality of language, together with commonsense assumptions about self-interest.

7. Searle, John R., "How to Derive 'Ought' from 'Is,'" *Philosophical Review*, 73, January 1964, pp. 43–58.

What implications do these arguments and those of the preceding chapters have for the Classical Model of rationality? The Classical Model, we might say, is designed for extremely clever chimpanzees. It does not deal with certain special features of human rationality, especially those special features that are made possible and indeed are required by the institution of language. So far I have discussed three ways in which the Classical Model simply fails to account for certain pervasive features of rational decision making.

1. The Classical Model cannot account for long-term prudential reasoning, where the prudential considerations are not represented in the current motivational set of the self in question. The example of the smoker in Denmark was designed to illustrate this point.
2. The Classical Model cannot account for recognitional rationality where the conscious self recognizes a desire-independent motivator as providing a reason for action. The chimpanzee can presumably recognize immediate sources of danger or desirable objects such as food, but the chimpanzee cannot recognize in that way such factitive entities as obligations, commitments, and long-term needs.
3. The Classical Model cannot account for the implications of the universality of language. Given this universality together with certain natural assumptions about the sorts of reasons one accepts for oneself, strong altruism follows.

In the next chapter we will turn to:

4. The intentional creation of desire-independent reasons by the conscious intentional actions of the self.