

# Self-Constitution: Action, Identity, and Integrity

## Lecture Three

### Autonomy, Efficacy, and Agency

Christine M. Korsgaard

3.1.1 In my last lecture, I argued that the hypothetical and categorical imperatives are constitutive principles of actions. Unless we are guided by these principles – unless we are in some sense trying to conform to them – we are not acting at all. The conception of action that yields this conclusion is Kant’s conception: that action is determining oneself to be the cause of some end.

3.1.2 The case for the hypothetical imperative is easy. Taking the means to an end, and determining yourself to cause the end, are so closely linked that Kant himself characterized the relationship as “analytic.” While obviously apt in one way, the term is unfortunate in another, for it suggests that if someone fails to take the means to a certain object, we are logically entitled to conclude that that object is not after all his end. If this were right, Kant’s view, like Hume’s, would degenerate into tautology: your end would be whatever you in fact pursue. And in that case, no one would ever be instrumentally irrational, and there would be no hypothetical imperative. What Kant must mean instead – or in any case what I mean – is that the commitment to realize an end binds you, obligates you, to take the means. And this is a commitment that you may fail to meet. Finding the means daunting, frightening, tedious, or painful, you cannot face them and do not go forward. Finding yourself nevertheless unprepared to decide that the end is not worth it, you cannot not change your mind and you cannot not go back. A paralyzed will is not the

same thing as one that has simply failed to operate; an abortive effort at self-determination has taken place. The standard represented by the hypothetical imperative, though constitutive, is normative as well.

3.1.3 The same conception of action, determining yourself to be a cause, provides us with an argument for the categorical imperative. For it is clear that willing an end is more than just being a place through which the causal chain leading to that end happens to run. It is not enough that that the end is caused by some force that is working in you or on you; it must be caused by you: the movements that lead to the realization of the end must be attributable to you. And for this to happen *you* must be something over and above the forces working in you or on you. Action is not just allowing some force that is working in you or on you to operate – or rather, better, we can say that it is, but then we must put the emphasis on what’s involved in that “allowing.” Incentive D invites you to perform act A. Yes, you say to yourself, I will: I will do A for the sake of D. Only then have you determined yourself by the self-given law of your own causality, the law of doing A for the sake of D; and only then have you acted.

3.1.4 The categorical imperative, according to this argument, is also a constitutive principle of action. For just as no one who is not trying to cause some end, some state of affairs in the world, counts as acting, so no one who is not trying to determine his own causality counts as acting. Actually, the two things are so closely bound together that they seem to be inseparable, for you cannot be trying to realize some end without trying to determine yourself to realize that end, and you cannot be trying to determine yourself

to realize the end without trying to determine your own causality. In fact, the two ideas are so closely linked that there is something artificial in the idea that there are *two* imperatives. There is really just one imperative here: act in accordance with a maxim you can will as a universal law. The hypothetical imperative merely specifies the kind of law we are looking for – a causal law. And that thought is already contained in the idea that what we are looking for is a law that governs *action*. It appears that there is only one law of practical reason, and it is the categorical imperative.

3.1.5 Now this seems to be a stronger conclusion than Kant thought he had reached, for at least in the *Groundwork*, Kant appears to claim that there is such a thing as heteronomous action, that action of this kind is governed by hypothetical imperatives but not categorical ones, and furthermore that the categorical imperative, unlike the hypothetical imperative, is “synthetic” rather than “analytic.” But there is something amiss here, at least if I am right in supposing – as I argued in Lecture One – that on Kant’s view the objects of our choices are what I there called “actions” rather than mere “acts.” For if the object of your choice is always a whole action – that is, an act undertaken for the sake of a certain end – then it seems clear that your choice could not be governed by a hypothetical imperative alone. For the hypothetical imperative concerns only the relationship between the act and the end, and has nothing whatever to say about whether the whole package, the act for the sake of the end, is a thing worth doing for its own sake. If that is what we choose, then choice must be governed by a categorical imperative, for only a categorical imperative governs the choice of actions and not mere acts.

3.2.1 I will come back to that point in a moment. First let me remind you of the promise with which I left you at the end of my last lecture. In my first lecture I argued that in order to establish what the constitutive standards of anything are, we must look to its form in the Aristotelian sense – to the teleological organization that makes it the kind of thing that it is. So order to establish that my claims are true, I need to show you what the function or *telos* of an action is.

3.2.2 My view is that action is self-constitution, so of course I am going to argue that the function of an action is to constitute an agent. More specifically I will argue that the essential characteristics of an agent are *efficacy* and *autonomy*. These terms mean pretty much what you think they mean. Speaking roughly, and putting the point in a way best suited to human agents, an agent is *efficacious* when she succeeds in bringing about whatever state of affairs she intended to bring about through her action. A reminder is important here. I do not mean in saying this to imply an *act* is never done for its own sake; that is, I am not suggesting that an agent must always be trying to accomplish some purpose beyond the act itself. As I mentioned in my first lecture one might, for example, dance for the sheer joy of dancing. But even someone who dances for the sheer joy of dancing is subject to a standard of efficacy, because he may fail. He may, for instance, fall flat on his face. And if his steps are not in any way guided by this standard of efficacy – if he makes *no* effort not to fall flat on his face – then he is not dancing, but merely flailing about. This much normativity – that the agent is guided by some norm of efficacy – is inherent in the very idea of action. And to that extent the very idea of action is a normative idea.

I said a moment ago that the properties of an agent are efficacy and autonomy. An agent is *autonomous* when her movements are in some clear sense self-determined or her own. These two properties, efficacy and autonomy, correspond to Kant's two imperatives, hypothetical and categorical. The hypothetical imperative commands us to be efficacious, and the categorical imperative commands us to be autonomous. Since the function of an action is to render you efficacious and autonomous, your action must, in order to be a good one, conform to these imperatives. If you fail to follow the imperatives you will not be efficacious and autonomous, and then you will not be an agent. An action constitutes an agent by being chosen in a way that renders you, the agent, efficacious and autonomous.

3.2.3 I am fully aware that it sounds backwards. How can the agent perform an action, you will ask, unless she is already autonomous and efficacious? This question, I believe, is based on a false picture of the way agents are related to their actions. It is based on the idea that actions are produced or caused by their agents, and that is not correct. We may indeed say that when an agent acts, her *movements* are produced or caused by her. But we can say that only after we have already identified the case as one of action. My question is about how we do that: in particular, what entitles us to attribute a movement to an agent as her own. For that – the authoredness of it – is the essence of an action.

To break the spell that makes you think the way I am talking is backwards, try it this way. If your action is unsuccessful and you do not bring about the state of affairs that you intended, it is not (or not just) the action that is ineffective. It is *you* that is ineffective. It is not as if you were effective in producing the action, but then the action, once out

there on its own, failed, like a defective machine you have invented and then let loose on the world. The action is not your product: it was *you* that failed. An unsuccessful action renders you ineffective. Therefore a successful action is one that renders you effective. A similar point holds for autonomy, as I will be trying to make clear in the rest of these lectures. In fact it almost has to, given what I have just said. For no question of your efficacy can arise unless the movements through which you are supposed to have been efficacious are *yours*, in the relevant sense of *yours*. If I shove Tom at Bernard, hoping that Bernard will topple over, and he does not, it isn't *Tom* who has been ineffective, it is *me*. It makes sense to evaluate movements as effective or ineffective only if they are self-determined movements, so it is my movements, not Tom's, that are subject to the hypothetical imperative in this case. Therefore it makes sense to evaluate movements as effective or ineffective, that is, as governed by the hypothetical imperative, only if they are also self-determined, that is, governed by the categorical imperative. As I said before, the hypothetical imperative is not really an independent principle. Therefore a successful action, an action that is good as an action, is one that renders its agent both efficacious and autonomous.

3.3.1 Now there's a problem with what I have been saying, which I flagged earlier when I mentioned that my remarks were most appropriate to the case of human agents and human action. It is a problem that emerges very starkly in Kant's own account of the foundation of the categorical imperative, in section three of the *Groundwork*. Kant's account begins with the claim that volition *is* – and I quote – “the causality of living beings

insofar as they are rational” (G 4: 446). The remark harks back to Kant’s initial definition of rationality in section two, in a passage I discussed last time. Kant says:

Everything in nature works in accordance with laws. Only a rational being has the capacity to act *in accordance with the representation* of laws, that is, in accordance with principles, or has a *will*. Since reason is required for the derivation of actions from laws, the will is nothing other than practical reason. (G 4: 412)

But in other works Kant apparently disavows this strong connection between rationality and volition. In several places he mentions the idea of *arbitrium brutum*, an animal choice. And in *The Metaphysics of Morals*, he identifies the capacity for action quite broadly – in fact a little too broadly -- with “life.” He says:

The *faculty of desire* is the faculty to be by means of one’s representations the cause of the objects of these representations. The faculty of a being to act in accordance with its representations is called *life*. (MM 6:211)

The reason is obvious enough. Human beings are, after all, not the only creatures who act. The distinction between actions and events also applies to the other animals. A non-human action, no less than a human one, is in some way ascribed to the acting animal itself. The movements are her own. When a cat chases a mouse, that is not something that happens to the cat, but something that it does. To this extent, we regard the other animals as being the authors of their actions, and as having something like volition.

But Kant never tells us what difference this acknowledgement might make to his deduction of the moral law, which uses the claim that volition is “the causality of living beings insofar as they are rational” as a premise. And more generally his account as it

stands leaves it obscure what non-free, non-rational volition could be. For the stark contrast between being self-determining, or autonomous, and being what Kant calls “heteronomous” or determined by natural laws seems to leave no place for the actions of non-human animals. If the movements of animals are directed from outside, by alien causes, then it is not clear how those movements are different from the movements of objects to which we do not ascribe actions, or from those movements of animals which we do not ascribe to the animals themselves. The antelope perceives the approaching lion, and runs away. The antelope is tackled by the lion, and falls over. Running away is something an antelope does while falling over is something that happens to it. But if both are equally cases of the antelope’s movements being determined by alien causes, where does the difference lie?

This brings us back to the issue I began with, Kant’s apparent conviction that some actions, heteronomous actions, are governed only by hypothetical imperatives. For the problem we are considering now is intimately related to the more notorious problem how on a Kantian conception we are supposed to conceive of bad action. The *Groundwork* portrays bad action as heteronomous action. Commentators often complain if that is supposed to mean action that is caused by external forces, it is impossible to see how people are ever responsible for bad action. But of course the problem is much worse than that, for if a person’s movements are caused by external forces, it is not clear why we should call them actions at all. And the same would be true of the actions of non-human animals.

3.3.2 A little surprisingly, there is a version of this problem even in Aristotle's apparently more naturalistic account of action. Aristotle distinguishes three kinds of action: the involuntary, the voluntary, and the chosen. The latter is a division of the former, for all chosen action is also voluntary. Small children and non-human animals generally act voluntarily, while only adult human beings act from choice. For the voluntary, Aristotle tells us, it is sufficient that "the moving principle is in the agent himself, he being aware of the particular circumstances of the action."

Now there are well-known problems about interpreting this criterion of the moving principle being "in" the agent. The moving principles of respiration, circulation, and digestion seem in some sense to be in us, but these things are not voluntary. Aristotle opposes the voluntary to the compulsory, defined as "that of which the moving principle is outside, being a principle in which nothing is contributed by the person who acts or is acted upon." So the "internal" character of the moving principle must have something to do with the agent's contribution. The difficulty arises when Aristotle tries to explain the compulsory in more detail. He gives us three examples of the compulsory, which seem intended to be increasingly close to being voluntary. The first is concerns a man carried somewhere by the wind, or by other men who have him in their power. This does not appear to be an example of compulsory action, for someone carried somewhere by a wind doesn't do anything at all. We might do better to imagine that the man in the power of others is, say, tied to the back of a wagon, so that he is forced to walk rather than being carried. The second example concerns coercion: a tyrant who has your loved ones in his power orders you to do something, threatening to harm them if you do not. The third example is one of unfortunate circumstances: it concerns some sailors who must throw

their cargo overboard during a storm in order to save their own lives. These last two cases are actually mixed, Aristotle tells us, because “they are worthy of choice at the time when they are done.” We could also say this about the man who walks behind the cart, for if he does not move his feet he will be dragged and it will be worse for him. Yet unless he chooses to walk, he performs no action, not even an compulsory one. And therein lies the problem. Action requires some contribution from the agent, and in all these cases, the agent’s contribution appears to be that, given the circumstances, he chose to do the action. But if the agent’s contribution rests in his choice, however constrained that choice might be, what becomes of the category of the merely voluntary? And if we lose the category of the merely voluntary, what becomes of the actions of non-human animals, who can never contribute their choice?

3.3.3 Obviously I inherit this problem. I have claimed that that we cannot recognize someone as acting unless he is at least in some degree governed by the hypothetical and categorical imperative. An agent must at least be trying to render himself efficacious and autonomous if we are to recognize what he is doing as acting at all. So how can the other animals possibly act? I surely do not want to claim they try to obey the Kantian imperatives. I seemed to be faced with a choice – either give up the idea that the Kantian imperatives are constitutive standards of action, or give up the idea that the other animals act.

But I don’t intend to do either of those things. In what follows I will explain what I think action is. On the basis of that explanation, I will also explain how autonomy and efficacy are constitutive standards for actions, even though the other animals do not try to

conform to them. Apart from saving the phenomena – for I think it is clear that the other animals do act – the account will throw light on what is distinctive about human action, and why it is that human actions alone are governed by imperatives, and can be morally good or bad.

3.4.1 Recall from Lecture One that according to Aristotle, a living thing is a thing with a special kind of form, a self-maintaining form. It is designed so as to maintain and reproduce itself, that is to say, to maintain and reproduce its own form. So it is its own end; its telos or function is just to be - and to continue being - what it is. And its organs, instincts, and natural activities are all arranged to that end. That much applies as much to plants as to animals: a plant also has a self-maintaining form.

But Aristotle tells us that animals are distinguished from plants in being alive in a further sense, given by a functionally related set of powers that plants lack. Aristotle emphasizes perception and sensation, but notes that these are accompanied by imagination, pleasure and pain, desire (*orexis*), and usually, local movement (OS II, 413b23). What is distinctive of animals, in other words, is that they carry out a part of their self-maintaining activities through action. They are alive in a further sense than plants, for they spend their lives *doing* things. This brings us to our question: what is action?

3.4.2 First, an action is an intelligent movement, in a simple descriptive sense: the animal's movement is responsive to a representation or conception that it forms of the world, or of its environment. That is why Aristotle associates action essentially with perception.

When I say that action is intelligent, of course, I'm not using "intelligent" in a laudatory sense. In the sense I am using the term, a spider crawling towards the moth caught in its web or a cockroach running underneath the toaster as you try to swat it with a newspaper, exhibit intelligence, for they respond to representations or conceptions of their environment. A perception of something as *dinner*, or *danger* – that is to-be-eaten or to-be-avoided - determines the course of the animal's movements. Kant, in a passage I quoted earlier, puts it this way:

The *faculty of desire* is the faculty to be by means of one's representations the cause of the objects of these representations. (MM 6:211)

The spider, for example, represents the moth to itself as dinner, and that is the cause of the moth's being dinner.

In the cases I have mentioned, an object desired or feared is represented as actually existing in the environment, but of course action may also begin from a conception of something that could be there, as when an animal goes looking for food or a mate. Even in that case, however, its movements are guided by a representation of its environment, for the shape or course of its movement – where it looks, how it goes about looking – is determined by its conception of the world it is moving through.

3.4.3 Before I go on I should clarify something about the way I am using "movement."

When talking about spiders and cockroaches, it is natural enough to identify actions with physical movements, or at least to suppose that actions are a type of, or perhaps supervene on, physical movements. But we do not want to write this thought into our conception of action, for a variety of reasons. For one thing, not every action is physical: making a

promise to someone, for instance, is not, nor are performatives generally. Of course one may insist that making a promise must have some physical manifestation, at least an utterance. But what about making a promise to oneself? Perhaps some physicalists will insist that even *that* must have some sort of physical manifestation or leave some sort of trace in the brain, but I do not think this thesis in the philosophy of mind should be made part of our conception of action. Not all agents are physical entities either: corporations and governments are not. Again, it might be argued that they and their actions must supervene on some sort of physical events, if only the recording of their decisions on paper. But what about God? Isn't the idea of divine action coherent?

In fact this question helps us to get at the relevant conception of movement. To act is to render a change in the world (or in the limiting case, to prevent or forestall one). When an agent acts, something must happen as a result of the action. Now we may be tempted at first to think that there is this difference between the actions of God and finite creatures. When a finite creature acts, or produces a change in the world, there has to be *a way* the creature does it, in the sense of a method, or even a means. Whereas when God acts, He needs no method: His very thought of what He would effect effects it. But I think we should resist this idea. Although many things we do involve a means or a method, not all of them do. For instance, suppose I decide to raise my arm. There is no way I do that, no method, no means – I just do it. To be sure there is a way it *happens* – nerve signals run down from my brain to my arm, or whatever – and because there is a way it happens I am constrained in a way that God could not be. If my nervous system isn't working properly, I may not be able to raise my arm. But that “way it happens” isn't a “way I do it” for I do not send nerve signals off to my arms, like a mother sending her

children off to school. So when I say that action necessarily involves a movement, the movement I am talking about is the effecting of a change in the world itself, not a *method* of effecting the change in the world.

Nevertheless, the example I just used to illustrate action without a method brings out something else important about the relevant notion of movement. Ordinarily, the way animals change the world is most immediately by moving their limbs. One of the few things that we can do that we don't do *by way of* moving our limbs is *move our limbs*. This gives rise to a certain tendency to confuse action with bodily movement, so that what we call an "action film" for instance, is one in which the protagonists jump around a lot. And there is a deep reason for this confusion, for it is not accidental that the only thing I can do without a method or a means is to effect a change in myself. It does not have to be in my body, but it does have to be in myself. Even when we imagine cases in which someone produces an effect magically, we imagine the agent *disposing* himself somehow: chanting, or concentrating, or staring. For that matter, even when we imagine God creating the world, we imagine God most immediately effecting a change in himself: thinking a thought, or perhaps even uttering one, like "let there be light." So although the movement I am talking about is the effecting of a change in the world, it is essential to the idea of action that the agent produces the change in the world by producing a change in himself. This is another way of saying that action essentially involves self-determination. The cases where there is no method are limited to the cases where *all* that the agent is doing is producing a change in himself. So when an animal acts, it effects a certain change in the world by effecting a change in itself, in a way that is responsive to its

conception of its environment. That is the sort of intelligent movement that is in question.

3.4.4 Now this sense of intelligent movement already implies that an action has intentional content. We say that spider is “crawling to the center of its web to eat the moth that is trapped there” or that the cockroach is “running under the toaster to avoid being swatted.” Those phrases specify purposes, but for the kind of intentional content that characterizes action, we do not need to specify further or ulterior purposes. The important thing about action is only that it is done, as we say, “on purpose.” To assign intentional content to a movement in this sense is to make it subject to a normative standard of efficacy, to a standard of success and failure. And being subject to such a standard, as I mentioned before, is essential to the idea of action. Suppose all we say about the cockroach is that “it is running under the toaster.” This is still different from saying of a rock that it is rolling down the hill. If the rock runs into an obstacle and stops rolling before it gets to the bottom of the hill it has not failed. But if the cockroach does not make it under the toaster it *has* failed. What licenses us to talk in this way? Obviously, we do not want to say that the cockroach has formed the intention of getting under the toaster. This is really just the problem we started out from. Since the cockroach itself is not guided by any hypothetical imperative, why do we think of its movements as subject to the norms of efficacy that make those movements count as actions?

3.4.5 According to Aristotle, to assign intentional content to an object’s movement we do not need to suppose that there is some thought process that accompanies the movement.

What licenses intentional description is not the presence of accompanying thought, but rather appeal to the object's form and function. Intentional descriptions apply even to the movements of artifacts. A clock, for example, is an object functionally constructed so as to tell the time. And that is why when we say things like "this clock chimes out the hour" or "The alarm clock will wake me up at eight," we imply the existence of criteria of success and failure. If the clock chimes eleven times when both hands point to twelve, or if the alarm does not go off when eight comes round, then the clock has failed. It is because a clock is organized so as to tell the time that we can assign intentional content to its movements. And in the same way, it is because an animal has a self-maintaining form that we can assign intentional content to its movements. It is because its function is self-maintenance that we describe even a very primitive animal as "looking for something to eat" or "trying to escape the danger," in a way that implies criteria of success and failure.

But of course a clock's chiming is not an action. We ascribe intentional movements to plants and machines, but they are not actions. The clock chimes out the hour, or it wakes you up. The plant turns towards the sun, or its roots grow down through the dry soil to where there is more water. As this last example suggests, these movements may even be reactions to events or conditions in the environment. But these reactions to the environment are not intelligent movements, in the sense I described earlier. They are not the result of the plant or machine forming a *conception* or *representation* of the environment. So to get at the idea of action, we must put these two elements – a movement with intentional content, and responsiveness to a conception of the environment – together. An action is an intentional movement of an animal that is *guided* by a representation or conception that the animal forms of its environment.

3.4.6 Now as I've just suggested, there are intentional movements, which are not actions, and yet are reactions to environmental cues. Consider for example, the phototropic response of a plant, almost irresistibly describable in action-language as the plant *turning towards* the sun. Why isn't it an action? According to the account I just gave, only because it is not governed by perception, by a representation of the environment: if the plant saw the sun or felt it then it would be an action. Does the concept of an action then require us to make a hard and fast distinction between mere causal reactions to environmental cues on the one hand and perception or representation on the other? Only if we think that the concept of action itself should be hard and fast, and in the case of animals there is no reason to think this. There will not be a hard and fast line in nature between action and other forms of intentionally-describable responses because there is not a hard and fast line in nature between mere reaction and perceptual representation. But this does not threaten what I am saying about action. It only suggests that there is not a hard and fast line in nature between what is action and what is not. And we knew that anyway. There are lots of things that linger on the vague conceptual edges of action, for lots of different reasons: drumming your fingers (which is somewhere between mere expressiveness and action), breathing (for after all, you can hold your breath), omission in cases where it matters morally versus omission in cases where it does not (the first is a bad action and the second is nothing at all). There are many cases in which we need a hard and fast concept for the purposes of philosophical understanding and indeed for ethical and political life, even where there is not a hard and fast line in nature. This is hardly surprising, since there are no hard and fast lines in nature.

3.4.7 The concept of action requires both an intentional movement and a representation or conception of the world. These together are what allow the agent to guide itself through the world. It is because both of these elements are needed for the idea of action, I believe, that philosophers fall too readily into an overly intellectualized conception of action. We tend to think that an agent must be guided by a conception of what he is doing, or even that he must actively entertain a purpose and a method of achieving that purpose – he must form an intention. This is an absurd thing to say about a cockroach running under a toaster; insofar as the idea of knowing what you are doing invokes a concept of the self, it is an absurd thing to say about any creature without self-consciousness. The animal must indeed be guided by a conception, and its movements must have a purpose, but it need not have a conception of its purpose. Self-conscious action – that is to say human action – does require a conception of what you are doing and why. That much is true by definition. But action in the wider sense does not. To put it in Kant's language: animal action is movement guided by one's representations of the world; human action is movement guided in addition by one's representation or conception of a law.

3.4.8 One more point about the phototropic responses of plants. The sun hits the plant, and the plant turns towards the sun. If the plant saw the sun, I said a few minutes ago, then that would be a case of action. There will be resistance to that idea. Some of you will be thinking that *that* cannot possibly be the difference that makes action action. For the case to be one of action, it must involve self-determination: the plant must move its

own leaves. So the plant's turning its leaves toward the sun would only be action if the plant had a *will*, and when it saw the sun, it formed a volition that caused its leaves to turn towards the sun. The plant's will must issue its leaves a sort of order to turn towards the sun, just like all those orders we are always issuing to our arms and legs, and that must be what causes its movement. Right?

Wrong. That is a sort of homunculus or pineal gland theory of the will. We've got a philosophical problem here, so we invent or point to an organ or faculty and say "There! That is the faculty that solves the problem!" How is the will posited as a faculty supposed to solve the problem of making volition possible? Essentially, by being capable of volition.

You can't solve a philosophical problem by giving it a name. The will is not an faculty that makes self-determination possible; the will is the capacity for self-determination. That is what we need to understand in order to understand animal agency: why we attribute a capacity for self-determination to animals.

3.5.1 But this way of putting it makes it clear that we have not yet got to the heart of the matter. When an animal acts, it is supposed to determine itself to movement; the movement is supposed to be its own. On what grounds do we ascribe the movement *to* the animal? This brings us to the third and most important feature of action. So far I have not focused on the causes of an animal's movements, only on the causation the animal exercises *through* its movements, on its efficacy in achieving its ends. But to many philosophers it seems the most essential thing about actions that they are caused in some particular way. And in the case of human actions, this point has been associated with a

moral issue, namely the fact that we hold people responsible for their actions. We could not do this, it is sometimes said – Hume says this, for instance - unless people were the causes of their actions.

Now this cannot be quite right, for to say that an agent is the cause of an action suggests that an action could be caused by something other than an agent. But I take it to be essential to the notion of action that it is attributable to an agent. The question is not why we assign an action to an agent, but rather why we assign certain movements to an agent as her own, thus interpreting them as an action. One traditional response to this problem has been to identify action with movement produced by a particular sort of a causal route through the person, say, a route through the person's psychology. The thought seems to be that a person is more essentially identified with her psychology than with her body, say. So a causal route through her psychology seems well suited to making the movement her own.

But I believe this gets the story almost exactly backwards. The intimate connection between person and action does not rest in the fact that action is caused by the most essential part of the person, but rather in the fact that the most essential part of the person is *constituted* by her actions. This is not to deny, of course, that the explanation of action must involve psychological factors. But a causal path through, say, desire, is not enough to make a movement an action.

3.5.2 Earlier I mentioned Aristotle's view that an intentional movement's vulnerability to standards of success and failure, its intentionality in the sense we want here, derives from the fact that the moving object has a certain form or functional organization. To see a

movement as an action, as subject to standards of success and failure, we must see it as a movement attributable to the animal's form. Since the animal's form is what unifies it into an individual object, its form is not merely something within the animal. So when the animal's movement can be attributed to its form, it is the animal itself, the animal *as a whole*, that moves. And when I say that the movement is attributable to the animal's form, I don't mean merely that the animal's form contributes importantly to the cause of the movement. I mean rather that the animal is formed *so as* to produce a movement of that kind.

3.5.3 Let me put this another way. When an animal is guided by its perceptions through its environment, its movements are subject to a standard of efficacy, a standard of success and failure. It is subject to a standard of success and failure, because there is something it is trying to do – it is trying to be itself and continue being itself. It is trying to do this not in the sense that it forms the intention of doing it, but in the sense that that is its nature: it has a self-maintaining form. The principles that govern an animal's movements as it guides itself through its environment – the principles that govern its reactions to its perceptions – are what we may call its instincts. An animal's movements are self-determined when they are governed by its instincts, for when they are governed by its instincts, they spring from its own nature. An animal's instincts then are its will, the laws of its own causality. They determine what it does in response to what, what it does for the sake of what. When it acts from its instincts then, the animal's movements are its own. It acts according to its own laws, and therefore autonomously.

It is even tempting to say that the animal's instincts are imperative for it. When you see an animal acting under the influence of a powerful instinct, it does have that look. But of course I am not saying non-human animals are therefore subject to imperatives, as human beings are. So what makes the difference? To answer this question we must look at action from the inside, at this psychology of action. And for this part of the argument I turn to Kant.

3.5.4 In Kantian moral psychology, the starting point for action is what Kant calls an incentive (*Triebfeder*). An incentive is a motivationally loaded representation of an object. I am using the term "object" broadly here to include not only substances but also states of affairs and activities. The object may be actually perceived, or conceived as a possible item in the environment, a way that things might be. You are subject to an incentive when you are aware of the features of some object that make the object attractive or appealing to you. Perhaps the object satisfies one of your needs; or perhaps because of the nature of your species or your own particular nature the object is one you are capable of enjoying. It interests you, it arouses the exercise of your faculties, it excites your natural curiosity, or it provides some sort of emotional comfort or satisfaction. It doesn't matter what – something about you makes you conceive this object as appealing or welcome in a particular way. The object answers to something in you or to the condition you are in. Incentives can also be negative. You may represent an object to yourself as painful or threatening or disgusting, or in some other way unwelcome.

Incentives operate on animals causally but they do not directly cause the animal's movements. If an incentive directly caused the animal's movement, it would be something

within the animal, not the animal as a whole, that determined the movement, and then as we have seen it would not be a case of action. A desire for food, after all, can cause you to salivate. If it also could cause you to go to the refrigerator, then salivating and going to the refrigerator would equally be actions. If we are to count a movement as an action, the movement must be caused by the animal itself, not by its representations or perceptions. Instead, an incentive works on an animal by making some movement or response seem appropriate to it, by presenting it as a thing to do. According to Kant, incentives work in conjunction with principles, which determine (or perhaps I should say describe) the agent's responses to those incentives, responses which are guided by the agent's conception of the world. The principle represents what Aristotle calls the agent's "contribution" to the action, the thing needed to make it voluntary. Every action must involve both an incentive and a principle: that is, something is presented to the animal's consciousness, *on* which it then acts.

In the human case, in the case of a person, it is easy to say what makes action different from mere response, for human beings act on reasons. A person's principles determine what the person counts as a reason. To the extent that the person *determines himself* to intentional movement, he *takes* his desire for food to provide him with a reason for going to the refrigerator; and that is not the same as its directly causing him to go to the refrigerator. We may represent this fact – his own causality or self-determination – by saying that it is his *principle* to get something to eat when he feels hungry, at least absent some reason why not.

3.5.5 In a non-rational animal, the principles in question are the animal's instincts. The role of both instincts and rational principles in the model is to capture the element of self-determination that is essential to action. As I said before, the animal's instincts are the laws of its causality, definitive of its will. They determine the sort of thing the animal does when faced with a certain stimulus. An animal's instincts determine it to hunt when it is hungry, flee when it is afraid, fight when it is threatened, and so on. Instinctive action is autonomous in the sense that the animal's movements are not directed by alien causes, but rather by the laws of its own nature.

You can see from this description that incentives and principles exist in natural pairs. The fact that an animal has certain instincts explains why it is subject to the associated incentives. In this sense the animal's instincts play a double role in the account of its actions. They both explain why the animal is subject to certain incentives in the first place, and what it does in response to those incentives once they are present. A motive, one might say, is an incentive operating under a certain principle or instinct, or viewed from the standpoint provided by that principle or instinct. For example, for a person whose principle is to help those in need, the fact that another is suffering appears as an incentive to help, an occasion for action. For an animal with the instincts of a cat, a small scurrying rodent is an occasion to give chase. By putting it this way, I mean to convey the fact that it is in a certain way artificial to separate the work of incentive and the work of principle, or at least, to separate them in the non-human animal's experience. In the human case it will vary. The experience of acting from instinct is obviously not, phenomenologically, like the experience of applying a rational principle to a case. But for that matter, acting on a rational principle need not involve any step-by-step process of

reasoning, for when a principle is deeply internalized we may simply *recognize* the case as one falling under the principle, where that is a single experience. Principles and instincts play a role in structuring our perceived environment. Yet the two aspects must be separated in our analysis of action in order to capture the difference between being *motivated*, which requires self-determination, and merely being *caused*, which does not.

3.5.6 When an animal acts, it is determined by its form, by its instincts, to produce a change in the world, guided by its conception or representation of the world. But an animal's form is what gives it its identity, what makes it the animal it is. So to say that an animal's form determines it to cause a certain effect is to say that the animal determines itself to be the cause of that effect. Action is self-determination, and, to that extent, it is autonomous. And as I have said before, it is only because action is autonomous that the question of its efficacy can come up. If one thing causes another, there is no room for success or failure. But if an animal determines itself to be the cause of something, and yet does not bring that thing about, then *it* has failed. Autonomy and efficacy are the properties of agents – all agents, not just human agents.

3.6.1 But we are subject to imperatives and non-human animals are not. So then what makes the difference? I think by now it should be clear. In one sense an animal constitutes its own will. It constitutes itself and its will is itself. It makes itself the kind of agent that does what it does by doing what it does. But an animal does not choose the principles of its own causality – it does not choose the content of its instincts. We human beings on the other hand do choose the principles of our own causality – we choose our

maxims. And the categorical and hypothetical imperatives are rules for doing this – rules for the construction of maxims. It is because we, unlike the other animals, must choose the laws of our own causality that we are subject to imperatives.

What this shows is that there are actually two senses of autonomy or self-determination. In one sense, to be autonomous or self-determined is to be governed by the principles of your own causality, principles that are definitive of your will. In another, deeper, sense to be autonomous or self-determined is to *choose* the principles that are definitive of your will. This is the kind of self-determination that Kant called “spontaneity.” Every agent, even an animal agent, is autonomous and self-determined in the first sense, or it would make no sense to attribute its movements to it. Only responsible agents, human agents, are autonomous in the second and deeper sense.

3.6.2 That is where the difference lies, but I have not said yet about why it exists. So let me now conclude by explaining what will happen in the rest of these lectures. In my next lecture I will argue that when the capacity for action becomes self-conscious – that is to say, when agents become conscious of their capacity for action - then those agents, human agents, find it both necessary and possible to choose the principles that determine their will. It is self-consciousness that produces autonomy in the second and deeper sense that I just identified. At the same time, we become subject to the Kantian imperatives, which govern this choice. I will also show you why at the very same moment, we begin to constitute a new form of agency, personal or practical identity, for which we are responsible.

In the lecture that follows, I will explain how bad action is possible. It should be clear enough from the account I just gave that bad action will when an agent chooses the wrong law for the law of her causality. And this is the view to which Kant himself, after the *Groundwork*, eventually came around. But if the standard of goodness and badness is to remain a constitutive standard, we must understand this in a certain way, for the wrong law must not be wrong in a merely external sense. The wrong law must be one that fails to unify and so to constitute the person's agency. And it is worth noting that if we can reach this conclusion – if we can show that bad action involves a failure to constitute oneself the agent of the action – Kant's original view will be true after all – for the bad person will be heteronomous. Good action, by contrast, will be what Plato said it was right from the start: the manifestation of a truly unified will. That idea will be my subject in the last lecture.