Sonderdruck *Nicht im Handel*

Perspektiven der Analytischen Philosophie
Perspectives in Analytical Philosophy

Herausgegeben von
Georg Meggle und Julian Nida-Rümelin

Band 19

Preferences

Edited by
Christoph Fehige and Ulla Wessels

Walter de Gruyter · Berlin · New York
1998

Marcus Willaschek

Agency, Autonomy, and Moral Obligation*

Abstract: The paper proposes and, in part, defends an understanding of human agency, autonomy, and moral obligation as integral parts of our concept of a person. Specifically, the first part (sects. 1–12) argues for a causal theory of action in which the acting person plays a central role in the causal history of her actions. The person exercises her causal influence according to normative principles of rationality. That presupposes some independence from her own motivation including the ability to acknowledge or reject parts of it as a basis of her rational decisions. This ability is constitutive of the autonomy of the person. The second part (sects. 13–29) presents an argument to the effect that the concept of autonomy presupposes a general universalist principle of morality. Autonomy involves a distinction between motives that are 'authentic' and motives that are not. This distinction does not rest on a substantive idea of what autonomous action is, but rather on a formal or procedural notion. Nevertheless, it presupposes a normative standard which is different from and largely independent of the motives a person in fact has. This standard can be found in the ideas of impartial benevolence and universal rational consent which inform universalist conceptions of morality.

1. One of the most specific properties of human beings, lying at the core of such attributes as their being 'rational' or 'political' animals, is their ability to be a subject of normative standards. To be sure, not all humans can be addressees of normative demands. Moreover, conceptually speaking, the domain of normative standards is not confined to the human race. Nevertheless, the only apt subjects of norms and rules we know are human beings. And it seems to be a decisive step in the typical development of a human being to grow up from being a mere object of physical and psychological manipulation to being subject of normative demands. Such demands are often seen as mere restrictions on our choices and decisions. But, as will be argued in what is essentially a Kantian way, some of these normative demands in fact are necessary preconditions for any choice or decision, for the ascription of motivation, the ascription of actions, and the 'practical' identity of the autonomous person. Since I will mainly be interested in the interrelation between these different topics, the discussion of any of them must leave important questions unaddressed. I hope, though, that the overall view will make up for the incompleteness of its parts. I will begin by sketching the relation between the person and her actions.

* I would like to thank Alyssa Bernstein, Tamar Gendler, Ulrike Heuer, Wilfried Hinsch, David MacArthur, Richard Moran, Liam Murphy, Michael Quante, Candace Vogler, and the participants at the *Preferences* conference for many helpful comments on various earlier versions of this paper.

2. *Actions* are pieces of behavior which are, under some description, intentional. Behavior is intentional only if it can be rationally justified, from the agent's own point of view, by citing the attitudes of the agent (typically a desire and a belief) which causally explain the behavior. An action is behavior caused by a reason. This is the central dogma of the causal theory of action, which receives overwhelming warrant from the way we actually explain and justify our actions.[1] To see this, one only has to think of cases where causal explanation and rational justification do not coincide: If the referee's blowing her whistle was caused by an insuperable urge to cough, it would be misdescribed as her ending the game and is not attributable to her as her action, even if, at the relevant time, she had a perfectly good reason to blow her whistle. It is essential for the working of our moral and legal notions of responsibility that for most kinds of behavior, like writing a letter or driving a car, this kind of independence of the causes from the reasons for acting is highly improbable.

True as this may be, the causal role of reasons cannot be regarded as sufficient for agency. This becomes obvious, for example, in cases of so-called 'deviant' causal chains: If the referee, in her first international appearance, wants to open the game and believes that she has to blow the whistle in her mouth in order to do so, her having these attitudes, under the circumstances, might cause a nervous cough which in turn causes the whistle to sound. Although her behavior in fact has been caused by her having attitudes which justify the behavior, the causation has gone the wrong way. It seems natural to say that, because of the deviance of the causal chain, the agent is involved in her behavior not *actively*, but only *passively* – it is something that happens to her, as opposed to something she does. This suggests that in standard cases of intentional action, the behavior has to be regarded as brought about by, or causally dependent on, *the agent herself.* Thus, the causal theory of action has often been criticized for not being able to accommodate the active and causal role attributed to the agent by our ordinary notion of intentional action.[2]

On the other hand, though, the idea of the agent as a causal factor in the history of an action may seem deeply obscure and at odds with well-established scientific approaches to a natural phenomenon like human behavior. According to Chisholm's theory of "agent-causality", for instance, the agent is a causal primitive, bringing about actions by way of a special kind of causation dif-

[1] For a classic statement of this view, cf. Davidson (1963) and the other articles collected in the first part of Davidson (1980); for a more recent account, cf. Brand (1984).

[2] E.g. Chisholm (1966), Kaulbach (1974), Rohs (1986), Velleman (1992).

ferent from ordinary event-causation.[3] But as David Velleman has recently stressed, reservation with respect to such a "thick" picture of agent-causation does not preclude the employment of a broader notion that leaves open the question whether or not the causal role of the agent is reducible to instances of event causation[4] (or perhaps to other relations). Thus, independently of the success of reductive attempts, we may assume that if a piece of behavior qualifies as an intentional action, it has been causally brought about by the agent.

3. This leaves us with two further questions: What is an agent? And how does the causal influence of the agent square with the causal role of the agent's beliefs and desires which serve to justify the action? A safe answer to the first question is that the agent, the being that acts, is a *person.* Typically, a person has or, perhaps, is an organized body, has intentional attitudes like beliefs and desires, and entertains states of consciousness. In fact, we attribute actions to the same entity to which we attribute beliefs, desires, pain, or visual experiences.[5] It is only persons who are responsible for what they do.

Now, it might seem that this answer makes the problem of agent causation even worse. Though we often regard persons as active and causally relevant parts of the world, we usually do not think of them as the causes of their own actions. While Paul may cause the ball to move by kicking it, he does not seem to cause his kicking. Rather than *causing* his kicking, the kicking is what he *does.* To see how a person nevertheless may be regarded as a causal factor in action, we have to turn to the second question. What in fact is supposed to cause Paul's kicking are certain of his beliefs and desires. Obviously, these attitudes are not causally sufficient for his kicking. This may suggest that the person enters the history of his actions as a *further* necessary condition, over and above the relevant beliefs and desires. But in fact, our common picture of agency strongly favors another possibility: The role of the person is to *make* some of her beliefs and desires causes of her actions.

This is most obvious in cases of action after conscious deliberation and decision. Martha wonders whether to go out tonight or to stay at home. While she at first is undecided, at some point she makes up her mind to go out.

[3] Chisholm (1966), (1976).

[4] Cf. Velleman (1992), who cites Chisholm (1978), p. 622, for a statement of this view: "The issues about 'agent-causation' [...] have been misplaced. The philosophical question is not – or at least it shouldn't be – the question whether or not there is 'agent-causation'. The philosophical question should be, rather, the question whether 'agent-causation' is reducible to 'event-causation' ".

[5] This tells strongly against attempts like Velleman's (1992) to reduce agent causation to event causation by way of identifying the agent with a special kind of motive.

Whatever else one may want to say about this, when Martha actually goes out, we do not assume that her reasons in favor of going out have caused her decision, but rather that they have been made causes of her behavior by her decision to go out.

Of course, only very few actions are the result of explicit decisions. In most cases, we simply act on one reason or another. We don't have to decide, because there was no indecision. Nevertheless, we assume some causal influence of the person, over and above the fact that it is this person's reasons that (among other things) cause her actions. Even if in fact no decision was involved, we suppose that the person *could have decided* against acting as she did. In these cases, the causal influence of the person seems to be only 'negative': In not deciding against the action the person *allows* her relevant attitudes a causal impact on her behavior. Thus, both in cases of conscious deliberation and decision and in cases where what to do is a matter of course, the causal role of the person consists in her influence on which of her beliefs and desires will be causally effective.[6]

4. There is a famous attempt, though, to get rid of the person's role in the causal history of the person's actions: the hypothetical analysis of 'could have done otherwise'.[7] For our purposes, this analysis can be seen as carried out in two steps. First, "Martha could have refrained from beating Paul" is rendered as "Martha would have refrained from beating Paul if she had decided so". But could Martha have decided otherwise? A second step is called for: "Martha would have decided to refrain from beating Paul if she had had other beliefs and desires". Taking the two steps together, the supposed possibility that a person might have acted otherwise no longer involves a causal influence of the person, but only the possibility of the person's having motives different from

[6] Above I distinguished between the intentional blowing of a whistle and an unintentional cough on the grounds that in the former the agent plays an active role, while in the latter she is involved only passively. This now needs some qualification, since an active-passive-distinction is working also within the realm of intentional action, between actions one decides to do and those one simply does. These two distinctions differ from one another. There is a sense in which being passive is to *suspend* one's activity ("He remained completely silent") which is appropriate for the 'passive' involvement of the agent in many of her intentional actions. In another sense, though, someone is passive with respect to those of her movements which are mere reactions to external influences. In this 'Newtonian' sense, activity and passivity are relative notions. (The first billiard ball is passive with respect to the impact of the cue, but active with respect to the movement of the second ball.) The agent is never passive in this sense with respect to her own intentional action. One way to incorporate agent-causality into our general picture of nature could involve regarding the latter notion of activity as mere abstraction from the former one.

[7] E.g. Moore (1911), Aune (1967).

the ones she actually has. Since it was exactly the concept of decision and the possibility of acting otherwise which seemed to involve the assumption of a causal role of the person, the hypothetical analysis would make that assumption altogether superfluous.

But there are reasons to believe that the hypothetical analysis fails. It cannot account for a large number of cases where the possibility of acting otherwise is of great importance for the way we attribute responsibility, most notably cases of weakness of the will. Martha believes that it would be better to stay at home, but nevertheless goes out. If her going out is an action at all, and not a case of compulsive behavior, Martha must have had at least *some* reasons for going out, but since she believes that she has *better* reasons to stay at home, her action is weak-willed.[8] Nevertheless, we suppose that Martha could have acted otherwise – without her reasons and motives being any different from what they actually are. What we are inclined to say is rather that, her motivation being as it is, she should have, and could have, done otherwise. (Notice that if she could not possibly have refrained from going out, even though she believed it to be better to stay at home, her behavior would have to be regarded not as weak-willed, but as compulsive, and she would not be held responsible for it.) It was only up to Martha, the *person*, to engage in a more rational course of action.

5. Weakness and strength of the will are not bizarre phenomena which we could simply ignore. The possibility of exercising 'will-power', the difficulty of doing what one knows (or believes) to be right, the ability to stick to one's decisions, and the possibility of failing to do so and being responsible for it form central elements of our conception of agency and the background for our attributions of responsibility. Perhaps, the psychological energy or force necessary to refrain from, say, eating all of the chocolate at once appears to be an obscure phenomenon, much more obscure than the physical energy or force necessary to lift a weight. But then think of someone who wants to lift a weight which is a little heavier than the one she lifted before with utmost effort. She tries twice and fails both times. Now she thinks of what it would mean to her to succeed; she concentrates, tries a third time – and lifts the weight. Obviously, between the failed and the successful attempts she did not gain physical power. It was an effort of will that made the difference. The lifting of the weight was the result of a unified 'physico-psychological' effort. This example may even suggest that we can regard *every* action as the execution of both physical and psychological powers or abilities. Sometimes the physical and sometimes the

[8] I draw here on Davidson's account of weakness of the will in Davidson (1970).

psychological element is harder to carry out. Typical examples of will-power, then, would be those extreme cases where the physical element is simple, but the psychological element is exceedingly difficult to perform.[9]

Of course we must ask whether or not these notions of will-power and psychological effort can be further analyzed, and how the corresponding aspect of our 'practical' view of ourselves can be incorporated into our 'theoretical' beliefs about ourselves and the world. These are important and difficult questions. For what follows, though, it will not be necessary to answer them, as long as we suppose that their admissible answers do not imply that in fact there is no causal influence of the person on her behavior and motivation. If a fundamental feature of our life such as the attribution of responsibility depends on that idea, we should insist that no theory and no conceptual analysis is acceptable that denies the causal role of the person.

6. The influence a person has on her own action is not captured by the fact that it is *her* beliefs and desires that cause her behavior. This leaves open various ways in which a person's behavior may be caused by her own beliefs and desires, but nevertheless is not her action. Let us call the mental states of a person which, under appropriate conditions, can cause as well as justify a person's behavior that person's *motivation* (and particular items of it *motives*). Thus, motivation includes everything that might figure in the explanation of a person's actions and that contributes, at least in some weak sense, to their justification from the agent's point of view: momentary whims, urges, emotions like love, hatred, or fear, desires, elaborate plans, character-traits, general dispositions, accepted values and norms, as well as beliefs about instrumental connections, any of which the person may or may not be conscious of.[10] Motivation that in fact causes a certain piece of behavior will be called *effective* motivation. We now can rephrase the point of the preceding paragraph by saying that, for a piece of behavior to be the action of a person, the motivation causing the behavior has to be made effective by the person. (One way of making motivation effective is by not making use of the ability to prevent motivation from becoming effective.)

[9] There even are extremely difficult actions without any physical aspect, such as doing mental arithmetic or avoiding thinking about the upcoming visit to the dentist. I consider these mental or 'inner' actions in Willaschek (1992b).

[10] This notion of motivation is meant to be neutral with respect to the distinction between so-called 'internal' and 'external' reasons for acting. I will briefly address this issue below (sect. 27).

7. This is where normative standards originally enter the picture of agency. To be sure, actions have been defined from the start as behavior caused by *justifying* attitudes. But the justificatory potential of these attitudes depends on their being made effective by the agent, since it is this choice – which motivation to make effective – that has to be governed by some normative principle or other.[11] I now want to turn to the content of these principles by asking why a person makes only part of her motivation effective, and why some part of it rather than another.

The answer to the first question is quite obvious, but it still seems worth mentioning. First, of course, there may be contradictory motives. But presumably there would not be many of them were it not for the fact that as beings of finite powers we are limited to a finite number of actions at a given time; moreover, opportunities to act on a certain motive are also limited by our environment. Therefore we simply can't do everything we are motivated to do. Some of these things can be delayed, but many cannot. It is a consequence of our temporally finite existence that we do not even succeed in doing all of the things we delay.

That much is true of many animals, too. Animals with a certain physiological and psychological complexity do not simply move one way or another, but are motivated to do so by psychological states, some of which presumably have an experiential character. Apart from some possible borderline cases, though, it seems clearly wrong to ascribe to animals the ability to make some motives, and prevent other motives from being, effective. In animals the strongest motive prevails. Of course, a person, too, may always act on the motive that is currently strongest. But there seems to be a remarkable difference between such a person and an animal: the person, being in full possession of her powers, would have to *make* the strongest motive effective, or *allow* it to become so, whereas in an animal the fact that the strongest motive prevails simply happens to it. Most people do not always, or even mostly, act on their currently strongest motive. A person can do *A* rather than *B*, even if, at the time, she has stronger motives for doing *B*. It may be tempting to say that this only shows that the motives for doing *B* have not been strongest, after all. But if the strength of a motive in fact is that psychological quantity we often take into consideration *before* we decide what

[11] Korsgaard (1992) argues that it is the "reflective" structure of human consciousness, distancing us from our desires, that forces us to adopt one or another principle of choosing. In the previous pages I mainly tried to establish the fact that the person has to be regarded as an element in the causal history of her actions and that in this she has to be guided by some normative principle. Korsgaard's proposal can be seen as an explanation of why the specific causal role of a person is so different from that of other things in the world.

to do, the claim that people always act on their strongest motive leads to a glaring misdescription of what actually happens when we decide to act against a very strong motive. In these cases, we do not discover that, after all, we have an even stronger motive to do something else. We consider the relative strength of many different motives, and end up doing something we are only weakly motivated to do. This may be *hard*, but we can do it. What is required to act against one's strongest motives is *will-power*. (One is inclined to say that in these cases, the person must be stronger than her strongest motive: "I was tempted to pull back, but than I *overcame* my fear and jumped".) This suggests that the idea that people always act on the strongest motive is part and parcel with the picture of human agency that leaves the agent out of account.

8. But why would someone act contrary to her strongest motive? This takes us to the second question: Why does a person make some part of her motivation effective rather than another? Typically, people will not act on their strongest motive if they believe that, for some reason or another, it will be *better* to do something else instead. Human beings are *rational*. This means that they can do what they believe to be best, *all things considered*. What course of action a person believes to be best, all things considered, therefore depends on *all* of her motivation. Simple cases of animal behavior and completely irrational human actions are best regarded as motivated exclusively by the very motive, or set of motives, that is currently strongest. Competing weaker motives do not, at the given time, enter into the effective motivation at all. By contrast, even in cases of rational action where in fact we decide to act on the currently strongest motive, it is just a convenient simplification to cite only that particular motive as *reason* for that action. In fact, a great variety of considerations for and against that action enter our effective motivation even in the simplest cases. This is obscured in part by our tendency to characterize actions by action-type descriptions ("swimming", "reading") of rather high generality.

Here's an example: I cut tomatoes. I do this, of course, because I want to have the tomatoes cut and I believe that I have to cut them (myself) in order to achieve this. But this explanation only gives a very limited picture of how my action is dependent on my motivation, a picture which is incomplete in two dimensions: First, it leaves out of account *why* I want the tomatoes be cut. Any action is motivated, in part, by its place in more encompassing courses of action, larger plans and general aims. This can even have an effect on how the action is executed: Whether I cut my tomatoes carefully may in part depend on whether the guests for whom I prepare dinner are welcome or not. This

is the second dimension in which the initial explanation is incomplete: There are myriads of ways of cutting tomatoes. Not all of the special features of the way I cut these tomatoes are intentional, but many are. If they are, I must have some reason for them which will relate what I do to my motivation for doing it. Thus, the rational person can take into account many different parts of her motivation in every choice of what to do, when to do it, and a particular way of doing it. These choices will be governed by principles of practical rationality.

Perhaps the most basic of these rational principles is that of individual instrumental rationality. According to this principle, one should engage in that course of action which has, relative to one's motivation, the highest probability of achieving the best possible outcome; if there is no such action (as will usually be the case), some trade-off between probability of success and desirability of result is necessary.[12] As has been noted before, the main reason that we cannot make all of our motivation effective is that acting to achieve some goal is diminishing one's chances of achieving some other goals. But in acting according to the principle of instrumental rationality, an agent diminishes her chances of achieving other goals only to the extent justified by the relative desirability of the goal she currently pursues. The remarkable effect of instrumental rationality is therefore to *integrate* the agent's divergent motivation. Since any motive that is relevant in a given context can influence the decision what to do, the rational person (ideally) transforms *all* of her motivation in *one* reason for one particular course of action. The result of this integration can be seen as unified *plans* or *strategies* according to which that person acts as long as she acts rationally.

From this all-inclusiveness derives the central characteristic of rational choice: Once someone has established what she believes to be the rational thing to do, there can be *no further question* for her as to why to do it or whether or not to do it at all. Because all one's motivation is supposed to be considered in determining what is rational to do, there can't be any further factor, relevant for the deciding person, that could reinforce or question the decision.

9. Acting according to the principle of instrumental rationality therefore contributes both to the synchronic and diachronic identity of the acting person. At any given moment, the aim of practical rationality is the resolution of conflict among the diverging motives of a human being. To a large extent this synchronic integration is achieved by *postponement*, as becomes obvious

[12] It may be misleading to call this conception of rational choice "instrumental", since sometimes this term is used for means-ends-considerations only and contrasted with "prudential" considerations. In this sense, then, what I am talking about is "prudence".

when we consider the effect on our decisions if we learned that we will die tomorrow. (Even if we would still plant a tree, we could no longer resolve a conflict between diverging motives by delaying the planting of the tree until next week.) The temporal perspective of rational decision, which is forced upon us by the impossibility of doing 'everything at once', makes a person the executor of projects she originated in the past and which she will continue to pursue in the future. Thus, rational decisions also give diachronic cohesion to a person's life. We may say that to act rationally is to regard oneself as *representative* of one's future self.[13]

Of course, yesterday's decision can only bind me today if I still have the relevant motivation. But as already noted, it is an important feature of human motivation that it is not necessary to consciously decide on every single action. Rather, motivation comes in chunks, defined largely by social and pragmatic roles: We see ourselves as grocers, philosophers, or insurance agents, mothers, brothers, friends, chess players, car owners and much more. Every role we 'play' brings with it the motivation to act in whatever way may be appropriate or necessary – as long as there is no conflict with a more important role one plays or a more urgent desire one feels. Also, we generally are motivated to comply with the practices and customs of the society, social group, or culture of which we regard ourselves members. When the situation comes for your acting as a fair sportsman, that's what you do. This is one of the reasons why so often we don't have to decide what to do in order to know what to do. On the other hand, these roles and customs immediately reveal their normative force when compliance with them faces resistance.

Rationality (ideally) transforms all of a person's motivation into one reason to act, at the given time, in a specific way. The required integration organizes the different and diverging motives in a hierarchical way, with more specific motives subsumed under general aims, long-term plans, and self-conceptions. We can call this product of the rational integration of a person's motives the *character* of that person. Someone's character is her general conception of what kind of person she wants to be. Typically, we are not consciously aware of our entire character, but only of some of its more specific parts or elements. Also, there may be some difference or conflict between a person's character and her actual behavior, which also can give rise to normative demands on

[13] Korsgaard (1989) has argued convincingly that within a Kantian, "practical" conception of the person the reason to care about one's future self is not some assumed psychological connectedness, but rather the necessity to delay the realization of so many of one's motives. Similar "practical" accounts of personal identity have been proposed in recent years by Gerhardt (1988) and Sommer (1988).

the person.[14] Nevertheless, it should be possible to infer a person's character with some accuracy from what she does, unless the person is irrational in an unusual degree. We may therefore say that in acting rationally a person is able to regard her actions as expressive of her own character.

10. Up to now I have tried to bring out two main features of agency: First, that the acting person plays a central role in the causal history of her actions by making part of her motivation effective; and secondly, that the normative principles according to which the person decides what part of her motivation to make effective allow the integration of the person's divergent motives into one continuously consistent and coherent character. These two features can be seen as closely related: In rational action, the psychological 'force' or causal power of a motive has no immediate impact on the choice or decision which eventually leads to action. Rational choice is guided by *normative* principles. Therefore, its outcome cannot simply be regarded, in a pseudo-mechanical fashion, as the resultant of diverging forces. Rational choice takes the relative strength of motives into account, but its outcome is in no way determined by these motivational forces. Rather, it is determined according to the weight *given* to them in the agent's practical deliberations. Thus, in rational choice and decision, the causal impact of the motives which enter practical deliberation has to be seen as *suspended*: It takes a further causal factor to make particular motives causally effective. This is the role of the person.

11. It may seem as if a person's character was the product of only two factors: the person's motivation and the application of certain principles of rationality. But that is not quite right. First, not all of a person's motives necessarily contribute to determine what her character is. Secondly, what motives she has is not independent of her character.

A motive, as defined above, is every attitude that can contribute to the causing, as well as the justification, of a person's behavior. Under normal circumstances, every motive can enter into the person's practical deliberation, and to the extent the person is rational, every motive actually will contribute to determine her decisions and thus her character. But circumstances aren't always normal. There are a variety of ways in which a person may be alienated from part of her motivation. Not all of these ways exclude the motives in question from determining the person's character, but many do. Consider someone who always wanted to be a pediatrician. During his first internship,

[14] Outlines of the dependence of normative claims on the conception the person has of herself can be found both in Gerhardt (1988) and in Korsgaard (1992).

he notices that really he does not like to be with children at all. Nevertheless, in an abstract way he still would very much like to be a pediatrician. When he asks himself why, he comes to believe that the only reason for this is that his mother would have loved to be a pediatrician, and that she had conveyed to her son the feeling that, since she could not realize her dream, he would have to do it for her. This insight in the origins of his motive may or may not eventually make it disappear. At any rate, the person in question will no longer regard his wish to be a pediatrician in the same way as part of *his own* motivation as before. To be sure, as long as he has the motive, it will in some psychological sense be his. But when it comes to deciding on what professional education to pursue, he can't simply take it for granted that he has, among other things, a relevant motive to be a pediatrician. Rather, he has to *decide* whether to *reject* or to *acknowledge* the motive as something to be considered in his practical deliberation. If he comes to reject it, he will try to block it from having any influence on his decision (except, perhaps, in some instrumental considerations). If a thus rejected motive still influences his decision, he would regard that as a distortion of his motivation and the resulting decision. Actions that are motivated in part by motives which the person rejects do not express that person's character.

12. But there is also the possibility that the person's character influences his motivation. Imagine, for example, that our would-be pediatrician does acknowledge his motive, after all. Then he might try to alter his attitude towards children. This will be a difficult task, but it can be met. One possible way to achieve this is to emphasize, for himself, what is nice about children, and not to take too seriously what he does not like about them. But this *cognitive* shift of emphasis alone will not do. The would-be pediatrician will also have to *act* like someone who really takes delight in children. Thus he makes being someone who takes delight in children part of his character, even though it is not yet part of his motivation. Of course, this change in character itself is justified by his current motive to be a pediatrician. But nevertheless, if an immediate motive to be with children eventually evolves, it derives in part from a conception of what kind of character the person wants to have that precedes the evolving motive.

In the usual course of events, a person's motivation determines her decisions, her plans, her practical self-images, or, in short, her character, while these in turn determine her actions. But as we just have seen, the influence *can* go in either direction. A person can exclude motives from influencing her decision, and her decisions and her character can generate motives. And since rational actions are expressive of a person's character, and there is nothing more

to having a certain character, in the above sense, than can be expressed by acting accordingly, one even can achieve certain character-traits by acting consistently in some way or other.

According to this conception of agency and personhood, neither a person's motivation nor her character are determinately 'given' to her. What she wants and how she acts (and therefore also who she is) to a certain extent lies in her own hands. To be sure, there are limits as to how far we can model ourselves on our idea of who we want to be. Sometimes these limits are painfully tight. But at least in a negative way, in excluding motives we do not acknowledge as our own, we have a remarkable degree of freedom. This freedom from being forced to act on motives one cannot recognize as one's own is called individual *autonomy*.[15]

13. Autonomy, in the sense discussed here, is the ability to distance oneself from part of one's own motivation, thereby gaining the freedom to acknowledge or reject a motive as a basis for one's practical deliberation. To be sure, an autonomous person does not exercise this ability continuously. Rather, we need some special occasion, some uneasiness or doubt about a particular motive, to ask ourselves whether we really should regard it as part of our *own* motivation. The most typical occasions for such questions involve the discovery that a certain motive has been acquired in a way the person believes to be uncommon or at least different from what he had previously supposed. Brainwashing, motive-formation under unusual social or personal pressure, or motivational changes due to physiological dysfunctions are extreme examples for this type of cases. Less blatant examples may include a strong dependence on parents and peers during childhood and adolescence or changes in motivation driven by the wish to be loved by a loved one. All these are circumstances the discovery of which can lead to doubts as to whether the motive in question is really one's own or rather the Trojan Horse of some external force. Sometimes the motive will immediately vanish under the pressure of these doubts. But even if it does not, we cannot simply go on as before and base our decisions on that particular motive. We are forced to take a stand by either acknowledging or rejecting the motive as our own. What is at stake in that decision

[15] "Autonomy" originally is a political term meaning self-legislation. Since the days of Kant it is also applied to individuals in order to designate their ability to act in accordance with self-imposed laws. This relation to laws or principles will also figure largely in the concept of individual autonomy introduced here. As will be obvious, this is not my only dept to Kant. In fact, I take the view proposed here to be not substantially different from the view Kant held in the *Critique of Practical Reason* and other writings. For a reading of Kant along these lines, cf. Willaschek (1992a).

can be called the *authenticity* of the motive. Thus, autonomy can be also given a positive formulation: It is the ability to act only on one's authentic motivation.

14. Now, non-standard causal histories are not the only threats to a motive's authenticity. Hunger, for example, is a motive caused by certain biological functions of one's body. As the special kind of animal we are, it is difficult to see anything uncommon in motives generated by the natural functioning of our bodies. Nevertheless, there are situations, like a hunger strike or a ritual fasting, where a person may come to believe that her hunger really is some alien force that should be suppressed rather than taken into rational consideration as a motive. There is a fundamental difference between a situation where someone is hungry and acknowledges it, but still does not want to eat for some reason or other, and a situation where someone is hungry, but rejects this feeling as part of his own motivation and therefore does not eat. The relevant difference can be put this way: In the first case you need a *reason* not to eat, which you don't need in the second case, since there the hunger has not even entered the realm of motives that could serve as (prima facie) reasons for or against a course of action. Being autonomous is being in a position to decide which motives should enter this 'space of practical reasons'.[16]

15. If the hunger you feel can be a motive alien to your own motivation, what could not be? In principle, it seems, no motive is immune to doubt with respect to its authenticity. But how can this doubt be resolved, once it has arisen? What are the grounds for rejecting some motives and acknowledging others? Which *is* a person's authentic motivation, and how can she find out?

There seems to be something paradoxical about the picture we have arrived at. On the one hand, the holistic feature of rational choice and the interdependence of choice and motivation suggest an anti-essentialist conception of human motivation, according to which the question what someone *really* wants has no determinate answer, but depends on the standards of rationality we employ and the result of trade-offs on different levels. On the other hand, the same view leads quite naturally to the idea of authenticity, which in turn suggests a strict motivational essentialism. To make things worse, both views of motivation seem to be presupposed by some of the different ways we attribute responsibility. Thus, we sometimes hold people responsible for parts of their motivation, but also excuse them for acting on motives that turn out not 'really' to be their own.

[16] This term is modelled on Wilfried Sellars's expression "space of reasons", cf. Sellars (1956), p. 299.

This air of paradox, though, can be resolved by taking the essentialist notion of authentic motivation not as a 'substantial', but rather a 'procedural' notion. There is no need to suppose that a person simply must accept her 'real' motivation as something 'given' in order to settle questions of authenticity. Instead, doubts as to what motives are really one's own are themselves a pivotal element of our concept of authentic motivation. Obviously, we do not actually question all of our motivation. But if a motive is authentic, this means that it *would* be acknowledged, once its authenticity was doubted. There is nothing more to 'really' being part of someone's *own* motivation than what is captured by this counterfactual claim. This kind of essentialism is in perfect agreement with the view that a person's character and motivation to a large extent lies in her own hands.[17]

16. Of course, this procedural notion of authentic motivation shifts the problem of autonomy to the question of when to acknowledge a motive as one's own.[18] This question can have many kinds of answers, depending on the conditions under which it is asked. Sometimes it may be enough to notice that the motive in question has been acquired in a standard way, after all. More generally, I can measure my motive against some conception of myself, some idea of who I am and who I want to be. Whether I acknowledge a certain motive often will simply depend on whether I can 'live with it', whether it allows me to lead the kind of life I want to live, be the kind of person I want to be. On this level, the motive will be rejected only if, once admitted to the "space of reasons", it would change the person's character and action in a way she cannot accept. (Of course, the motive may be so strong that it will force its way to influence the person's behavior anyhow. But then the person can distance herself from her own behavior. This is reflected in

[17] The claim that autonomy is the ability to act on one's authentic motivation now needs a slight qualification. After all, it seems not to be enough to act on a motive that in fact is authentic according to the criterion given here, if that motive has been unreasonably barred from doubt – even if eventually it would be acknowledged. Autonomy is based in part on a self-critical attitude and includes a certain willingness to *actually* put part of one's motivation to the test.

[18] In recent years, individual autonomy has been the subject of renewed interest among English-speaking philosophers, partly in response to work by Dworkin (1970), Frankfurt (1971) and Watson (1975). Some contributions to this debate are collected in Christman (1989). Unfortunately, a discussion of these different approaches is not possible within the limited scope of this article. I would at least like to note, though, that I take the question of when to acknowledge a motive as one's own to be the same question Watson (1975) addresses to Frankfurt's approach. I first thought of the account given in the remainder of this paper as a Kantian solution to that problem. Frankfurt himself is dealing with this question in a more recent article (1987).

our attributions of responsibility.) Ultimately, that means that on this level the motive will be measured against the person's desires, values and norms.

17. But the result of this test will not always be a stable resting-point. This becomes most obvious when not just one motive, but a larger part of one's motivation is called into question. Why should I measure the motivation that justifies a central element of my practical identity, such as my being a philosopher, against motives whose authenticity can equally be doubted? Of course, I often *can* rely on the rest of my self-conception to evaluate some part of it. But I may also find myself doubtful with respect to the other aspects of my character, not willing, or maybe even not able, to rely on any part of it. This can be the case in situations of deep crisis, or simply in moments of sober and distanced reflection and self-analysis. There is an internal dynamic to questions of this kind that can make it necessary to adopt some external standard, independent of the person's own contingent motivation. Such standards we find in other people, their expectations, demands and interests.

This may seem to be quite the opposite of autonomy. But in fact, it need not be, if the 'external' standards, in part imposed by other people's interests, are *shared* by the person herself. This does not just mean that various people happen to have the same interests or values. Rather, sharing, say, a norm means holding it jointly, in such a way that one person's endorsing the norm is related to (if not dependent on) the other people's endorsement. Thus, a sports-team might share their goal to win, or their rule not to party before important games.[19] Of course, sharing norms and values with others just *is* incorporating a certain social role in one's practical identity, and thus we are led back to the person's own character. But then my own standards gain in reliability and lose part of their subjectivity and contingency when they are shared by others. This doesn't make them immune to criticism. But they can serve as a more than purely subjective guideline for my autonomous self-assessment.[20]

[19] This meaning of "sharing" thus is different from each of the four senses of "sharing ends" Korsgaard distinguishes in her (1993), p. 41, fn. 31.

[20] It has been objected that in passing from the individual to the social, and eventually to the moral level, one is leaving the question of individual autonomy and entering a completely different issue. After all, how can external standards contribute to determine who *I* am? But this objection rests on a substantial notion of authenticity which assumes, wrongly, that it somehow is a 'given' fact which of my motives are authentic and which are not. If the adequate notion of authenticity is 'procedural', then there is a strong link between individual autonomy and external standards through the need to find a reliable principle for the acknowledging and rejecting of parts of one's own motivation.

18. The first natural step in transcending one's purely private motivation will be to turn to one's immediate social environment: one's family, friends, or colleagues, for example. Sharing certain values, interests and norms with a group of other people has two important consequences. First, it provides some external standards for self-assessment and autonomous decision. Second, it has an effect on the group analogous to the effect the principle of instrumental rationality has on the individual: it resolves potential conflicts among members of the group by *integrating* their motivation into shared interests. This makes the group of people a single "unit of action"[21] and allows them to engage in joint courses of action.

This means, among other things, that an individual can act as a *representative* of that group, in much the same way as a rational person can act as a representative of her later self. It also means that, in the relevant practical domain (say, at home, or at work, or in sports), my decisions will have to take the interests of the other members of my group into account as much as my own. If *we* want to win the match, I have done something wrong if I try to win the audience's favor at the expense of my team's performance. It would be a mistake by my own standards, since it obstructs a goal I share. Of course, such a conflict can put my loyalty to the group to the test and I may prefer to leave the group rather than give up some individual aims or values. But if I remain, say, a member of the team, then our decision, what *we* want to do, can have more authority over me than considerations which are based solely on my private motivation.

19. The smaller the group of people, the more substantial and concrete the shared norms, values and interests will tend to be. Only from being a member of a relatively small group with strong interpersonal ties, like a loving couple, a pair or circle of friends, a family, a sports team, or a group of people working together, can the individual gain much additional *content* for his practical conception of himself. On the other hand, though, the smaller the group and the more specific the shared interests, the more parochial the shared practical attitudes will appear when employed as external standards of critical self-assessment. When a career scientist comes to doubt the authenticity of her striving for knowledge, her colleagues may not be the best group to turn to for critical assessment. Maybe her family is, but it may be the case that all people she interacts with on a personal basis share the very value which, she suspects, rests only on prejudice. But of course her social identity stretches further than the area of her personal relations. She is a citizen of her local community and her state. She may be member of a political party or of a religious community. With respect to her problem, she might consider the fact that she is part of a large and long-standing cultural tradition, a tradition, we may assume, which values knowledge highly and in which a scientist is generally an esteemed personality. In seeing herself as part of that culture and, moreover, in seeing that culture as part of herself, she has a reason to acknowledge her quest for knowledge and the motives that drive her scientific work. Again, the fact that a culture is a system of *shared* values, norms and practices on the one hand makes them relevant for the individual person, while on the other it lends to them a degree of objectivity and reliability that purely personal standards cannot achieve. And again, acting in accordance with the norms of one's cultural community makes the person a *representative* – in the sense of "embodiment" of that culture, but also in the sense of acting on behalf of the other members of one's community. (If scientific inquiry is an integral part of the culture I share, then I can speak for 'us' when I present established scientific results or defend scientific method against 'external', say religious, criticism.)

[21] I borrow this term from Korsgaard (1989).

20. What we call scientific knowledge, though, has not been valued by all people of all times. From a historical or a global perspective, even the most impressive cultures can seem parochial. Therefore, any standard that is shared only by a restricted group of people, however many and however competent they may be, can be doubted. This finally takes us to the question whether there is some standard which is *universally shared* and which can serve as an ultimate measure of self-assessment and the settling of doubts about the authenticity of one's motives.[22]

It may seem obvious that the answer must be in the negative. People are too diverse, their motivation and self-conceptions too much dependent on their circumstances, for them all to share only the most general values and norms. To be sure, everyone can agree that self-preservation and well-being are important values. But people do disagree over just how important they really are. Moreover, they may all have these values, but they don't share them in the relevant sense. Therefore the quest for an ultimate normative standard can appear hopeless or even misguided.

[22] It may be objected that there always is the possibility of simply *deciding* one way or another, without any reason or standard. This may be true, but then there always remains the possibility of further doubt. Only a general principle for our decisions and actions can really settle these doubts, once they arise. To be sure, in the end we have to *rely* on something without further guarantees, as Wittgenstein (1969) pointed out (cf. § 509). But that just means that we do not flip a coin, but have *something* we rely on. And, obviously, there are better and worse things to rely on.

I believe, though, that this line of thought starts from the wrong point. We have to keep in mind that we are looking for a standard that would allow a person to decide about the authenticity of parts of her motivation. What we need is an aspect of her practical identity, of her character as a person, which on the one hand she cannot regard as contingent or accidental, and which on the other hand involves a normative standard. Now, there certainly are aspects of a person's identity that could not have been different, at least from the person's own point of view, without making her another person altogether. Most of them, though, like the person's sex or other features of her biology, do not give rise to normative standards. On the other hand, those features that do so, like social roles or long-standing aims, are not tied closely enough to the person's identity to be regarded as non-accidental features of her character. But there is at least one aspect of a person's identity that seems to fulfill both conditions: *to be a person*, that is, to be intellectually capable of rational choice and biologically forced to make use of this capability. Being a person is definitely an essential aspect of anyone's practical identity. Moreover, it is an aspect that clearly implies normative standards, namely those of rationality. Nevertheless, this may seem to be of no avail, since these norms, too, are universally accepted, but not universally shared. Therefore they don't carry with them the kind of intersubjective authority which is central for the sought-for standard of authenticity.

21. What would such a standard have to look like? The principles of decision on individual and social levels showed four basic features: they have domains of different sizes, they resolve conflict within that domain by integrating the relevant motives, they allow for representation, and relative to that domain they bring an end to questions of what to do and why to do it. We would have to expect analogous features from a universal standard. First, it would have to be *shared by all persons.* Second, as a shared normative standard it would, at least to a certain degree, resolve potential conflict by *integrating* people's diverse motivation. Third, acting in accordance with this standard would make a person a *representative* of all other people: In so far as the person acts in accordance with that standard, she is acting on behalf of humanity. And finally, compliance with this standard would be the *ultimate normative requirement* on actions. Once one believes that a particular action, and that action alone, meets the standard, no further questions can arise as to why to do it or whether or not to do it at all.[23]

[23] There may be doubts as to whether a *universal* standard would have to be regarded as *ultimate*, and what can motivate my acting on that standard if I believe that doing so goes against my own best interest. I will return to these questions below (sects. 26 f.).

In being integrative, in making possible representation, and in ending the quest for practical reasons, this universal standard would have important features in common with our concept of instrumental rationality. We therefore can regard both principles as different instances of the same general conception of rationality. On the other hand, it seems fair to say that this standard would embody the central idea of universalist ethics, naming a (minimal) requirement on morally permissible action. Those and only those actions would meet that standard which can be regarded as done on behalf of humanity. This standard, however, is a merely *formal* one in that it leaves open what the relevant interests of 'humanity' are. Therefore, it may not appear to be helpful even as a minimal criterion of morality. But in fact it can be understood in a way that will yield quite substantial restrictions on action: An action complies with this standard just in case all people, if they genuinely tried to resolve their conflicts of interest, would eventually agree that this action may be done. (This would imply a *general* permission to perform actions of the relevant type.) Of course, finding out what all people eventually would allow to be done is not an easy task either, but at least in some cases it is quite clear what the results would be. Our democratic ways of legislation can be seen as locally restricted and rather imperfect implementations, on a political level, of that basic idea. On the level of individual decision, no distortions by factors like unequal political or economic power would occur, since the standard can only be applied by way of a thought experiment anyway. (Nevertheless, asking the people who are immediately concerned what their interests are may help.) Kant's Categorical Imperative can be read in just this way: Act in such a way that (if you were a legislator representing the will of all people) you could make your maxim into a general law.[24]

22. Thus, if I find myself in the situation where I need to rely on a universal standard of decision, something like a universalist principle of morality is the only candidate. Let us call this principle the Moral Law. Then the obvious question is whether the Moral Law is a standard I share with all other people. From my own perspective, this question falls in three parts: Do I endorse the Moral Law? Do the other people endorse it? And: Does one person's endorsing the Moral Law involve its endorsement by other people in a way constitutive of the sharing of a norm? (Remember that it is the property

[24] This interpretation of the Categorical Imperative, which finds much support in Kant's texts, unfortunately cannot be reconciled with his claim that the Categorical Imperative can be applied without taking into account real people's actual desires and interests. I do not see, though, how that claim could be justified. The conflict, therefore, seems to express a conflict *within* the Kantian text. For a more detailed discussion, cf. Willaschek (1992a), § 11.

of being shared which gives a standard both relevance for the individual and more than subjective authority.)

It may seem obvious that at least the last two questions have to be answered in the negative. After all, there are people who do not seem to recognize moral considerations of a universalist kind. On the other hand, though, the fact that people are not able to state certain rational principles, for example in arithmetic, and make mistakes in their application, does not imply that they do not endorse the relevant principles. The same can be said about the Moral Law. Immoral action and immoralist rhetoric are both consistent with its endorsement. There are no clear limits as to what kind of behavior can be reconciled with its 'in principle'-acceptance. To say that someone endorses the Moral Law as a standard for his decisions is not simply to describe his behavior or to state some psychological fact. It is rather a certain way of understanding the life that person leads.

23. This is true even from the first-person-perspective: I cannot simply determine by introspection whether I do or do not endorse a general principle of morality. As noted earlier, a person's character may not always and in all parts be immediately accessible to her. The more general a feature of her character is, the more she will have to rely on interpreting her own actions and attitudes. You cannot be sure to have adopted some general principle of action or even just a long-term strategy simply on grounds of your good resolutions. Rather, the answer will depend in part on how you can make sense out of your own life. Does the assumption that you endorse the Moral Law help you to understand yourself and to find satisfactory reasons for the things you do?

For someone who questions the authenticity of his motivation, the answer is obvious. The quest for authenticity *is* a search for satisfactory reasons for one's actions and one's way of living. We need these reasons as much for understanding ourselves as to decide what to do. It is the very problem about the authenticity of my motivation which shows that I can arrive at satisfactory reasons and a real understanding of the life I lead *only* if I assume that I have accepted the Moral Law as a standard for my decisions. I therefore have a good reason to answer the first question – whether I myself are committed to the Moral Law – in the affirmative.

24. Do I also have a good reason to assume that all other people endorse the Moral Law? It may seem outrageous to say that a remorseless murderer 'really', but unwittingly has accepted the Moral Law as his guiding principle or that Eichmann in fact (as he claimed) acted on the Categorical Imperative, but only misapplied it. But this reaction rests on a misunderstanding. Endorsing the

Moral Law does not make someone a morally good person, it only makes her a person submitted to moral standards. Whether we should ascribe to someone else the acceptance of a moral principle will depend largely on whether we consider morality an appropriate standard for the evaluation and understanding of that person's actions. But of course this just is the attitude we take towards other people: We hold them morally responsible, thereby ascribing to them a fundamental endorsement of the Moral Law.[25] This attribution can be supported by reflecting on the way other people can settle *their* doubts with respect to the authenticity of their motives. The line of thought that leads me to the point where *I* need to rely on a universally shared standard is no quirk of mine, but a reasonable response to naturally arising doubts which cannot be ignored. It even seems safe to say that all people, *qua* rational, *should* practice the kind of critical self-assessment which will result in questions of authenticity. If the best way to make sense of my life involves my commitment to a moral principle, simply in virtue of my being a person, the same must be true of all other persons. This means that if we want to understand other people as leading their own lives, on autonomous decisions, acting largely on their authentic motivation, we have to attribute to them a fundamental commitment to the Moral Law. It does not tell against this argument that other people may not, for lack of occasion to thoroughly question their own motivation, have realized this commitment. Of course, this does not exclude that we eventually encounter a rational human being to whom even the most charitable interpreter could not ascribe a commitment to morality. But we would be justified in regarding this as a case of deficiency, a rare shortcoming in the qualities that generally characterize a person. Such a condition may not preclude the status of a person, but neither does it question the general link between being a person and a commitment to the Moral Law.[26]

---

[25] I assume that moral responsibility rests on the endorsement of moral principles, which of course is not self-evident. But within a conception of practical reasons and personhood that stresses autonomy in the way advocated here, this assumption really is unproblematic: Normative demands on an autonomous person can be justified *only* by appeal to a principle the person herself endorses. Though I cannot argue for the claim to exclusiveness here, I hope that it is plausible enough.

[26] Of course, there is also the problem of the possibility that foreign cultures might have completely different notions of practical rightness. (More particular differences can be regarded as internal incoherences in much the same way as we regard slavery in our own western history.) I cannot even begin to discuss this issue adequately. There are two obvious ways of dealing with the problem of societies that do not seem to endorse our universalist notions of morality: either we admit that their members are not persons in our sense, or we regard their moral views as insufficiently developed or even completely mistaken. I am not quite happy with either of these options and prefer to hope that we generally succeed in understanding different cultures as *other* ways of living *moral* lives.

25. Finally, we must ask whether all people endorse the Moral Law in a way constituting a *shared* commitment. As we have seen, my commitment to the Moral Law derives at least partly from my need to rely on an external standard for my self-assessment. The Moral Law can play this role only if it is universally shared. Thus, my commitment to it partly rests on the fact that other people are equally committed to it. Since we must assume that the same is true about other people in order to understand them as autonomous persons, we have to regard all people as sharing a common commitment to a fundamental moral principle. This is, of course, Kant's basic idea: To be autonomous, I have to consider myself a member of the community of rational beings, of the 'Kingdom of Ends'. The result of the above reflections supports this Kantian thought. Autonomy consists in the ability to have doubts about the authenticity of one's own motivation and to settle them by appeal to some principle of decision. Such doubts can arise, and be answered, on different levels. Ultimately, though, what is called for is a standard of decision and critical self-assessment which is shared by everyone, simply in virtue of being a person. I argued that the fundamental principle of universalist ethics in fact is such a universally shared standard. Of course, the importance of universal scope does not rest on the assumption that more people are less liable to make mistakes. It rests on the belief that once everyone's interests are taken into account, nothing remains that could be of any immediate practical relevance.[27] Therefore, morality is the ground on which we build our autonomous lives.

26. This general outline of a moral account of autonomy must leave many relevant questions unanswered. In closing I briefly want to address some of them.

According to the view taken here, the source of our moral obligations lies in our endorsement of the Moral Law which in turn we must assume in order to understand ourselves as autonomous persons. Thus, the question arises whether we really have to conceive ourselves as autonomous. Why care for autonomy if life is so much easier without it? Can't we avoid being autonomous? Or is there, even if we could, a reason why we shouldn't?

Above, I have pointed to the analogies between the principle of instrumental rationality and the fundamental principle of universalist ethics. In fact, both can be seen as instances of the same general idea of rationality, especially because both serve as principles for decision which form the end-point for certain basic practical questions ("What should I do?", "Why should I do it?").

[27] This way of tying practical relevance to the interests of persons raises a problem about the moral rights of animals. I cannot deal with this problem here. Cf. Korsgaard (1992) for a sketch of a solution within a Kantian framework (sect. 3.6).

Since the application of the two principles can lead to conflicting or even contradictory decisions, one of them has to be appropriately confined.[28] Depending on which of the principles we regard as more fundamental, we will be led to different readings of the argument given above. If instrumental rationality is more fundamental, we will have to regard a person's interest in autonomy as the basis for her moral obligations. That would mean that there is no *further* reason why she should strive for autonomy. Moral obligations would derive their normative force from instrumental considerations and would bind her only to the degree fixed by the importance she attaches to her own autonomy. If, on the other hand, the moral principle is more fundamental, this will be related to a special status of autonomy as a non-accidental feature within the character of a person. If we cared for anything at all, we would have to care for autonomy.

I am not sure whether there is any 'fact of the matter' as to which of two alternatives is the right one. It may again be a question of how we can best understand ourselves and the life we lead. If one of the alternatives in fact is 'the' right one, the reason must lie in a fundamental normative fact about the kind of beings we are. There is some plausibility to the 'existentialist' thought that we cannot avoid being autonomous: We are forced to lead our own lives. Being autonomous and standing under moral obligation would then be the two sides of our true identity. Since this would mean that we necessarily have to ask questions of the kind that only can be answered with recourse to the Moral Law, moral obligation would be the ultimate, all-inclusive level of normative standards.

On the other hand, how can we decide what the right conception of our true identity is if not by probing it as a key to understanding our lives? But even then we are driven to regard morality as more fundamental than instrumental considerations. If the justification of moral demands was limited by considerations of instrumental rationality, we would have to reject as unjustified moral demands that require acting against one's instrumentally defined interest. But in morally wrong action, we often have to admit that a moral demand on ourselves in fact is justified, even if we have decided for instrumental reasons not to fulfill it. This, I believe, can only be understood if we regard moral claims as more fundamental than instrumental ones. Whether autonomy is a normative fact or not, we had better regard it as one.

[28] There is, of course, the possibility that both principles have to limited. But it seems that this could not be done in a 'principled' way. I will not consider this possibility here.

27. That takes us to the next question I would like to discuss. Many philosophers, most famously Hume, seem to find unacceptable the idea that there can be reasons for someone to do something which are in no way rooted in that person's desires. This denial of so-called 'external reasons' is known as 'internalism'. Now, the internalist will object that granting normative precedence to universalist moral claims just is allowing for external reasons. After all, instrumental considerations (in the sense defined above) are supposed to take into account *all* of that individual's motivation, including her desires and other 'pro-attitudes'. Therefore, moral claims which conflict with the instrumentally defined interest of the individual cannot, for that person, be reasons to do something.

I believe, though, that this internalist objection rests on a distorted picture of rational action. The internalist denies the possibility of external reasons because he rightly expects from a reason for acting that, if the person eventually acts for this reason, the reason must contribute to the causal explanation of that action. If the reason was in no way related to the person's desires, it seems that the action will have to be causally explained without any mention of that reason.[29] But this argument seems to assume that a motive which we take into account in practical deliberation by itself can be the cause of the resulting action. As I have argued above (sects. 7, 10), this is a mistaken, pseudo-mechanistic view of rational choice. No rational action can be brought about by the very factors which are taken into consideration in practical decision. (Otherwise, as in mechanics, no decision would be necessary.) Rather, it is the person who *makes* certain motives effective, according to rational principles. But if the explanation of rational action requires recourse to another causal factor besides the various motives, and this causal factor, the person, exercises her influence guided by rational principles, there seems to be no fundamental difference between the explanatory power of internal and external reasons. If in rational choice a person considers her motives according to the weight she attaches to them, then there is no reason why she should not be able to attach weight, for example, to other people's motives and even to more abstract ideas, as long as this is in accordance with the principle that guides her choice. This may well result in 'external' reasons, and perhaps it is more difficult to bring

[29] This worry, which fuels the rejection of external reasons, is most obvious in William's (1979) seminal article on internal and external reasons (cf. pp. 106 f.). He goes on to argue that external reasons could only enter the explanation of an action if the person's reflection on this external reason could rationally lead him to acquire a corresponding desire (much in the same way as, given an end, reflections on necessary means can bring him to acquire a motive to employ them). But the very fact that the external reason is supposed to have no basis at all in that person's desire-type motivation excludes the possibility that this can happen in a rational way.

oneself to act on them than it is on 'internal' ones. But since being rational presupposes independence from one's motivation and subjection of one's choice to rational principles anyhow, explaining actions done for external reasons is no more difficult than explaining rational action in general.

28. The next problem will take us back to the relation between autonomy and moral obligation. Does the moral account of autonomy force us to say that only a morally good life can be autonomous? Autonomy is the ability to act on one's authentic motivation, which presupposes, in order to determine which motivation is authentic, a moral standard. It does not presuppose exceptionless compliance with that standard. Therefore, an autonomous person can acknowledge motives as her own which will lead her to act against the Moral Law and even accept a whole way of life which is immoral in some important respect. But how, on this account, can there be autonomous decisions against morality? If the Moral Law is regarded as a principle of practical rationality, this case can be regarded as analogous to other shortcomings in rationality. In the same way in which we may fail to give proper weight to some of the motives concerning our more distant future, we can fail to appropriately consider moral reasons. Or we can even know that what we do is wrong, but fail in our effort to do what is right. In both cases, there will be some justification for what we do, some motives with respect to which it appears to be rational. All things considered, however, it is not. Thus, although immoral actions are irrational, they still can be autonomous as long as their motives are such that the person would, or does, acknowledge them as her own. Of course, there are limits to the extent of immoral action that can be reconciled with the ascription of a fundamental moral commitment. But as mentioned before (sect. 24), extreme cases of this kind are too rare to exclude a constitutive relation between standard cases of being a person and autonomy.

29. Finally I would like to consider what the endorsement of the Moral Law really contributes to the autonomy of our lives. How can moral considerations determine the authenticity of my motivation? If I wonder whether my motive to be a pediatrician really is my own, how can the reflection on a universal Moral Law help me to find an answer? After all, both options, we may assume, are morally permissible. In order to appreciate the contribution of morality to an autonomous life, we have to consider that particular motives first will be measured against the person's private and purely subjective conceptions of herself and of the kind of life she wants to lead. If these conceptions do not settle the matter, it will be because they, too, did not seem reliable under pressure. So our original doubt will carry over to the broader aspect of one's

practical identity to which the more particular motive belongs and on whose authenticity its own authenticity depends. Thus, recourse to moral principles is adequate, and necessary, only on a very general level. Should I really live my life the way I do, really share the practice of my fellow countrymen or of my cultural tradition? To be sure, moral considerations of the 'thin', universalist kind will not always determine decisions even on that most general level. But often they will. And on that general level it even helps to find that the life one leads at least is not morally reprehensible, even if there are other equally permissible ways of living. Morality does not eliminate the contingency of our lives. It just allows us to organize them around a stable and reliable core.

References

Aune (1967). Bruce Aune: "Hypotheticals and 'Can': Another Look", *Analysis* 27 (1967).

Brand (1984). Myles Brand: *Acting and Intending*, Cambridge, Mass., 1984.

Brand/Walton (1976). Myles Brand and D. Walton (eds.): *Action Theory*, Dordrecht 1976.

Chisholm (1966). Roderick M. Chisholm: "Freedom and Action", in Lehrer (1966).

Chisholm (1976). Roderick M. Chisholm: "The Agent as Cause", in Brand/Walton (1976).

Chisholm (1978). Roderick M. Chisholm: "Comments and Replies", *Philosophia* 7 (1978).

Christman (1989). John Christman (ed.): *The Inner Citadel*, Oxford 1989.

Davidson (1963). Donald Davidson: "Actions, Reasons, and Causes", in Davidson (1980); article first publ. in 1963.

Davidson (1970). Donald Davidson: "How Is Weakness of the Will Possible?", in Davidson (1980); article first publ. in 1970.

Davidson (1980). Donald Davidson: *Actions and Events*, Oxford 1980.

Dworkin (1970). Gerald Dworkin: "Acting Freely", *Noûs* 4 (1970).

Frankfurt (1971). Harry G. Frankfurt: "Freedom of the Will and the Concept of a Person", in Frankfurt (1988); article first publ. in 1971.

Frankfurt (1987). Harry G. Frankfurt: "Identification and Wholeheartedness", in Frankfurt (1988); article first publ. in 1987.

Frankfurt (1988). Harry G. Frankfurt: *The Importance of What We Care About*, Cambridge 1988.

Gerhardt (1988). Volker Gerhardt: "Selbstbestimmung: Über Ursprung und Ziel moralischen Handelns", in Henrich/Horstmann (1988).

Gerhardt/Herold (1992). Volker Gerhardt and Norbert Herold (eds.): *Perspektiven des Perspektivismus*, Würzburg 1992.

Henrich/Horstmann (1988). Dieter Henrich and Rolf-Peter Horstmann (eds.): *Metaphysik nach Kant?*, Stuttgart 1988.

Kaulbach (1974). Friedrich Kaulbach: "Sprache, Stand der praktischen Vernunft und Handeln", in Riedel (1974).

Korsgaard (1989). Christine Korsgaard: "Personal Identity and the Unity of Agency: A Kantian Response to Parfit", *Philosophy and Public Affairs* 18 (1989).

Korsgaard (1992). Christine Korsgaard: *The Sources of Normativity*, Tanner Lectures, delivered at Cambridge 1992.

Korsgaard (1993). Christine Korsgaard: "The Reasons We Can Share: An Attack on the Distinction between Agent-Relative and Agent-Neutral Values", *Social Philosophy and Policy* 10 (1993).

Lehrer (1966). Keith Lehrer (ed.): *Freedom and Determinism*, New York 1966.

Moore (1911). G. E. Moore: *Ethics*, London 1911.

Prauss (1986). Gerold Prauss (ed.): *Handlungstheorie und Transzendentalphilosophie*, Frankfurt/Main 1986.

Riedel (1974). Manfred Riedel (ed.): *Rehabilitation der praktischen Philosophie*, vol. 2, Freiburg 1974.

Rohs (1986). Peter Rohs: "Gedanken zu einer Handlungstheorie auf transzendentalphilosophischer Grundlage", in Prauss (1986).

Sellars (1956). Wilfried Sellars: "Empiricism and the Philosophy of Mind", in Sellars (1963); article first publ. in 1956.

Sellars (1963). Wilfried Sellars: *Science, Perception and Reality*, London 1963.

Sommer (1988). Manfred Sommer: *Identität im Übergang: Kant*, Frankfurt/Main 1988.

Velleman (1992). J. David Velleman: "What Happens When Someone Acts?", *Mind* 101 (1992).

Watson (1975). Garry Watson: "Free Agency", *Journal of Philosophy* 72 (1975).

Willaschek (1992a). Marcus Willaschek: *Praktische Vernunft: Handlungstheorie und Moralbegründung bei Kant*, Stuttgart 1992.

Willaschek (1992b). Marcus Willaschek: " 'Inneres' Handeln: Handlungstheoretische Überlegungen zu einem Grundbegriff des Perspektivismus", in Gerhardt/Herold (1992).

Williams (1979). Bernard Williams: "Internal and External Reasons", in Williams (1981); article first publ. in 1979.

Williams (1981). Bernard Williams: *Moral Luck*, Cambridge 1981.

Wittgenstein (1969). Ludwig Wittgenstein: *Über Gewißheit/On Certainty*, Frankfurt/Main 1994 (vol. 8 of the Wittgenstein *Werkausgabe*); *Über Gewißheit/On Certainty* first publ. in 1969.