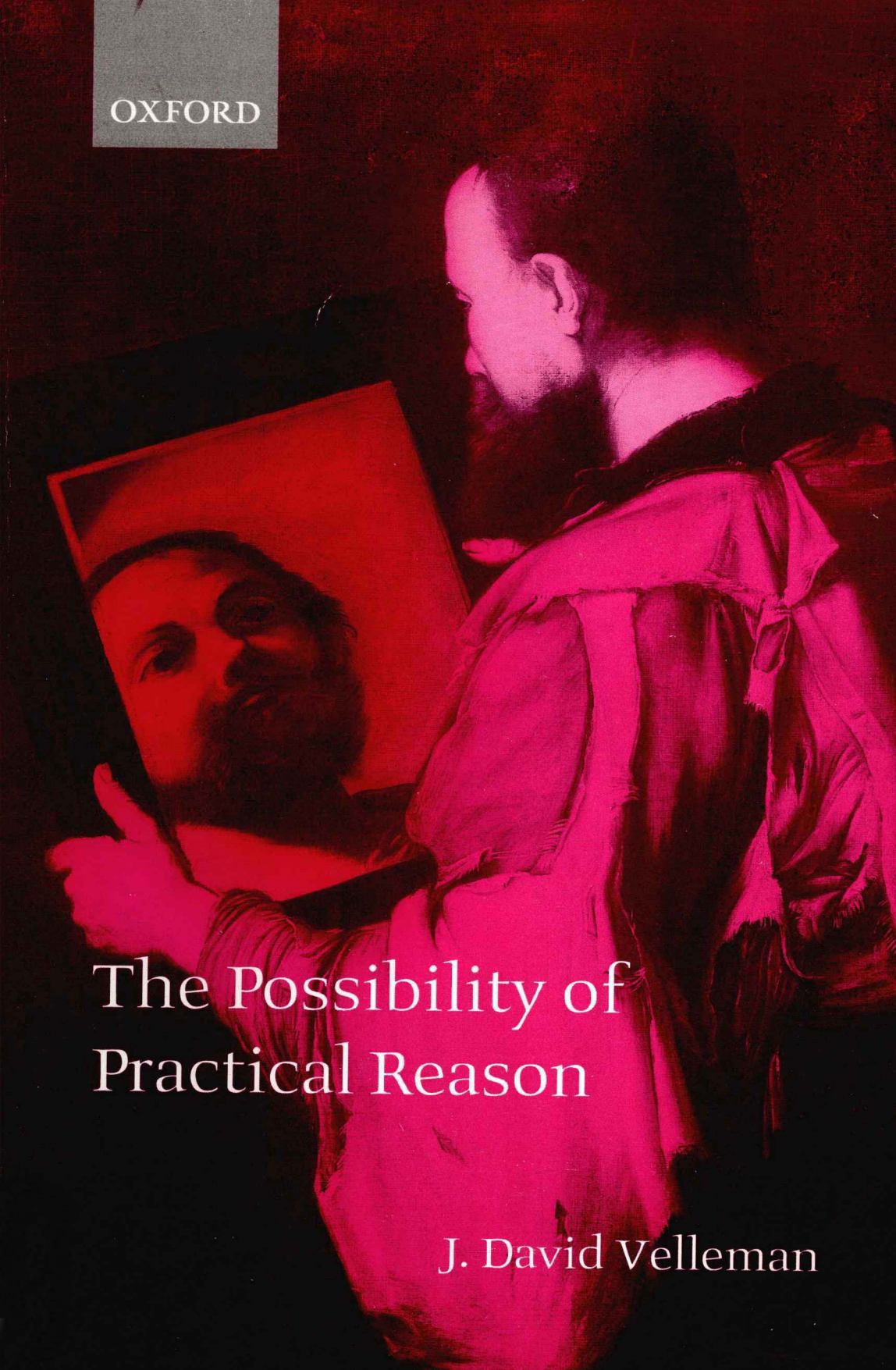
OXFORD
The Possibility of Practical Reason
J. David Velleman

The Possibility of Practical Reason

J. DAVID VELLEMAN

CLARENDON PRESS · OXFORD

This book has been printed digitally and produced in a standard specification in order to ensure its continuing availability

OXFORD
UNIVERSITY PRESS
Great Clarendon Street, Oxford OX2 6DP
Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in
Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto
With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan South Korea Poland Portugal
Singapore Switzerland Thailand Turkey Ukraine Vietnam

Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries

Published in the United States
by Inc., New York

Reprinted 2004

ISBN 0-19-823826-6

Cover illustration:

Man Holding a Mirror, oil on canvas, by Francesco Fracanzano (1612-1656),

Derek Johns Ltd., London.

Antony Rowe Ltd., Eastbourne

PREFACE

This volume contains the work that I have done in the philosophy of action since completing the book *Practical Reflection*. One of the papers, 'Epistemic Freedom,' is an expanded version of a chapter in the book; it was written after the book went to press, though it was published first. I have included it here because I think that it significantly improves on the corresponding portions of the book. One or two of the other papers are attempts to build upon or renovate the theory set down in the book, but most of them are attempts to dig beneath it. Without even referring to that theory, I have tried to unearth more fundamental reasons for wanting a theory of its general form—reasons for thinking that there ought to be a theory of its kind.

In most cases, I have done my digging in areas familiar to philosophers of action: problems about agent causation, internal and external reasons, direction of fit, the normative force of formal decision theory, and the rationality of resolute choice. The papers therefore do more to situate my view on the philosophical map than I previously could, though they do less by way of filling in the details.

The Introduction is an attempt to fashion a single narrative out of the main themes that appear in the rest of the collection. In concentrating on the flow of this narrative, I have tended to gloss over argumentative details, relying on the other papers to provide them. I have tried to indicate in the footnotes where detailed versions of the arguments can be found in the other chapters. The Introduction also records recent changes of mind about various issues.

I have not revised or updated the previously published material in substantive respects. (I have made some minor adjustments in Chapters 9 and 10.) I have also retained the acknowledgements that originally appeared with the papers, thus ensuring that each paper contains at least one true statement—namely, that of my indebtedness to friends and colleagues. I have several debts, however, that are not adequately represented in those acknowledgements, and I would like to mention them here.

Although each paper thanks many of my colleagues individually, none records my debt to the collective that they make up: the Department of Philosophy at the University of Michigan, Ann Arbor. Whatever virtues my

work displays are largely a reflection of the intellectual community in which it was carried out. I also want to thank those colleagues who have chaired the Department during my tenure here—Jaegwon Kim, Allan Gibbard, Stephen Darwall, and Louis Loeb—each of whom has provided significant support to my research.

Michael Bratman is the person who first suggested that I publish this collection. Michael's contributions to the philosophy of action include not only many important publications but also many years of good-natured advice and encouragement to others working in the field. I am fortunate to have been a beneficiary of his intellectual generosity since the very beginning of my career.

Finally, I want to thank Ted Hinchman for help with the Bibliography and Index; James Bell for help with proofreading; Nancy Higginbotham for copy-editing; and, at Oxford University Press, Peter Momtchiloff, Enid Barker, and Charlotte Jenkins.

Ann Arbor
October 1999

CONTENTS

I

Introduction

Behavior, Activity, Action

Philosophers of action have traditionally defined their topic by quoting a bit of Wittgensteinian arithmetic: "What is left over if I subtract the fact that my arm goes up from the fact that I raise my arm?"[1] The difference between my arm's rising and my raising it is supposed to illustrate the difference between a mere occurrence involving my body and an action of mine. And the difference between mere occurrences and actions is what the philosophy of action seeks to explain.

Yet there is reason to doubt whether Wittgenstein's computation has a unique result. As Harry Frankfurt has pointed out, my raising an arm may be something less than an action:[2]

> Actions are instances of activity, though not the only ones even in human life. To drum one's fingers on the table, altogether idly and inattentively, is surely not a case of passivity: the movements in question do not occur without one's making them. Neither is it an instance of action, however, but only of being active. . . . One result of overlooking events of this kind is an exaggeration of the peculiarity of what humans do. Another result, related to the first, is the mistaken belief that a twofold division of human events into action and mere happenings provides a classification that suits the interests of the theory of action.

Frankfurt's distinction between action and mere activity reveals a potential ambiguity in the above quotation from Wittgenstein. "The fact that I raise my arm" can denote an instance of action, such as my signaling a request to be recognized by the chair of a meeting; but the same phrase can also denote an instance of mere activity, such as my idly and inattentively—perhaps even

For comments on earlier drafts of this Introduction, I am grateful to Joel Anderson, Pamela Hieronymi, Sigurdur Kristinsson, R. Jay Wallace; and to Philip Clark and other members of the Philosophy Department at Kansas State University.

[1] *Philosophical Investigations*, trans. G. E. M. Anscombe (Oxford: Blackwell, 1972), §621.

[2] 'Identification and Externality,' in *The Importance of What We Care About* (Cambridge: Cambridge University Press, 1988), 58–68, at 58. The second half of this quotation appears in the original as a footnote to the first.

unwittingly—scratching my head while engrossed in a book. When the fact that my arm goes up is subtracted from something called "the fact that I raise my arm," what is left will depend on whether the minuend is a case of action or mere activity.

Unfortunately, most philosophy of action is premised on the mistaken belief pointed out by Frankfurt, that human events can be divided without remainder into actions and mere happenings. The result is that the prevailing theory of action neglects the difference between action and activity.[3]

This difference is also illustrated by behaviors that call for psychoanalytic explanation.[4] Consider an example from Freud's *Psychopathology of Everyday Life*:[5]

> My inkstand is made out of a flat piece of Untersberg marble which is hollowed out to receive the glass inkpot; and the inkpot has a cover with a knob made of the same stone. Behind this inkstand there is a ring of bronze statuettes and terra cotta figures. I sat down at the desk to write, and then moved the hand that was holding the pen-holder forward in a remarkably clumsy way, sweeping on to the floor the inkpot cover which was lying on the desk at the time.
>
> The explanation was not hard to find. Some hours before, my sister had been in the room to inspect some new acquisitions. She admired them very much, and then remarked: 'Your writing table looks really attractive now; only the inkstand doesn't match. You must get a nicer one.' I went out with my sister and did not return for some hours. But when I did I carried out, so it seems, the execution of the condemned inkstand. Did I perhaps conclude from my sister's remark that she intended to make me a present of a nicer inkstand on the next festive occasion, and did I smash the unlovely old one so as to force her to carry out the intention she had hinted at? If that is so, my sweeping movement was only apparently clumsy; in reality it was exceedingly adroit and well-directed, and understood how to avoid damaging any of the more precious objects that stood around.

This explanation is simpler than many of Freud's, in that it portrays his action as a realistically chosen means to a desired end, rather than a symbolic wish-

[3] In 'What Happens When Someone Acts?' (Chap. 6, below), I tried to distinguish between action that is full-blooded, or fully human, and action that is something less. I now regard the terms "action" and "activity" as preferable for drawing the distinction that I had in mind.

[4] See Richard Wollheim, *The Thread of Life* (Cambridge, Mass.: Harvard University Press, 1984), 59–61; and Sebastian Gardner, *Irrationality and the Philosophy of Psychoanalysis* (Cambridge: Cambridge University Press, 1993), 188–9. As Gardner notes, this distinction is probably co-extensive with the distinction drawn by Brian O'Shaughnessy between sub-intentional and intentional action (*The Will*, vol. ii (Cambridge: Cambridge University Press, 1980), ch. 10). The distinction is also discussed by Jonathan Lear in his critical notice of Gardner, 'The Heterogeneity of the Mental,' *Mind* 104 (1995) 863–79. Note, however, that Wollheim and Gardner do not draw the distinction between action and activity as I shall draw it. They accept the desire-belief model as adequate to characterize action, whereas I shall argue that it at most characterizes a kind of activity.

[5] *The Standard Edition of the Complete Psychological Works of Sigmund Freud*, trans. James Strachey et al. (London: Hogarth Press, 1960), VI: 167–8.

fulfillment or the enactment of a phantasy. The agent wanted to destroy the inkstand so as to make way for his sister to give him a new one; and his desire to destroy the inkstand moved him to brush its cover onto the floor, thereby destroying it.

Freud's point about bungled actions is that they are no accidents: they serve an intention or purpose. Because there was a purpose for which the agent brushed the inkstand's cover onto the floor, his doing so cannot be classified as something that merely happened to him. He didn't just suffer or undergo this movement of his hand; he actively performed it.

Nevertheless, Freud acknowledges that a bungled action somehow differs from a normal attempt to accomplish the same purpose with the same bodily movement. This admission is clearest in Freud's explanation for a famous slip of the tongue:[6]

> You probably still recall [writes Freud's source] the way in which the President of the Lower House of the Austrian Parliament *opened* the sitting a short while ago: "Gentlemen: I take notice that a full quorum of members is present and herewith declare the sitting *closed!*" His attention was only drawn by the general merriment and he corrected his mistake.

In commenting on this case (which he does several times during his career), Freud sometimes emphasizes the similarity between the President's initial slip and his subsequent correction.[7]

> The sense and intention of his slip was that he wanted to close the sitting. 'Er sagt es ja selbst' we are tempted to quote: we need only take him at his word. . . . It is clear that he wanted to open the sitting, but it is equally clear that he also wanted to close it. That is so obvious that it leaves us nothing to interpret.

Here Freud implies that the President's utterance of the word "closed" was motivated by a desire to close the sitting, just as his subsequent utterance of the word "open" was motivated by a desire to open it. In other passages, however, Freud draws a contrast between the two utterances. In the first case, he points out, the President "said the contrary of what he intended," whereas "[a]fter his slip of the tongue he at once produces the wording which he originally intended"—and which he now presumably intends again.[8] The correction is therefore intentional in a sense that the slip was not. Indeed, Freud ultimately implies that the slip was committed not only unintentionally but

[6] *Ibid.*, 59. Freud is quoting R. Meringer, 'Wie man sich versprechen kann,' *Neue Freie Presse*, 23 Aug. 1900.

[7] *Introductory Lectures on Psychoanalysis*, SE XV: 40, 47.

[8] *Introductory Lectures*, 47; see also 'Some Elementary Lessons in Psycho-Analysis,' SE XXIII: 284.

unwillingly, since he says that the desire to close the session "succeeded in making itself effective, against the speaker's will."[9]

Thus, Freud's explanation of the slip as purposeful leaves unchallenged the speaker's own sense that his power of speech ran away with him, or that his words "slipped out." The explanation would contradict the speaker only if he went to the extent of denying that it was indeed his power of speech and his words that were involved. The Freudian explanation would then force the speaker to admit that "I declare the sitting closed" was something that he said—not, for example, a noise forced from his throat by a spasm. But the Freudian explanation still allows him to claim that he said it despite himself, and that it was therefore a slip, however motivated.

Such cases require us to define a category of ungoverned activities, distinct from mere happenings, on the one hand, and from autonomous actions, on the other. This category contains the things that one does rather than merely undergoes, but that one somehow fails to regulate in the manner that separates autonomous human action from merely motivated activity. The philosophy of action must therefore account for three categories of phenomena: mere happenings, mere activities, and actions.

Making Things Happen

The boundaries separating these categories mark increments in the subject's involvement as the cause of his own behavior. A slip of the tongue differs from a spasm of the larynx, we observed, in that it doesn't just issue from the subject: he produces it. But then, of course, there is also a sense in which his utterance is produced despite him, by a desire that he didn't intend to express. Similarly, a person can knock something off a desk in a manner that is adroitly clumsy—perfectly aimed, on the one hand, and yet also out of his conscious control, on the other.

Mere activity is therefore a partial and imperfect exercise of the subject's capacity to make things happen: in one sense, the subject makes the activity happen; in another, it is made to happen despite him, or at least without his concurrence. Full-blooded human action occurs only when the subject's capacity to make things happen is exercised to its fullest extent. To study the nature of activity and action is thus to study two degrees in the exercise of a single capacity.

This capacity merits philosophical study because it seems incompatible with our conception of how the world works more generally. We tend to think that

[9] 'Some Elementary Lessons,' *loc. cit.*

whatever happens either is caused to happen by other happenings or *just* happens, by chance: events owe their occurrence to other events or to nothing at all. But if we make things happen, those events owe their occurrence to us, to persons. How can people give rise to events?

On the answer to this question hangs the viability of innumerable concepts indispensable for everyday life—concepts of human agency, creativity, and responsibility. Nothing that happens can genuinely be our idea, our doing, or our fault unless we somehow make it happen. Without a capacity to make things happen, we would never be in a position to choose or reject anything, to owe or earn anything, to succeed or fail at anything. We would simply be caught up in the flow of events, and our lives would be just so much water under the bridge.

We don't seem to be adrift in the flow events: we seem to intervene in it, by producing some events and preventing others. Yet our intervention invariably consists in thoughts and bodily movements, which either happen by chance or are caused to happen by other thoughts and movements, which are themselves events taking place in our minds and bodies. Our intervening in the flow of events is just another part of that flow. So how can it count, after all, as an intervention—or, for that matter, as ours?

The Standard Model

The standard answer to this question goes like this. We want something to happen, and we believe that some behavior of ours would constitute or produce or at least promote its happening. These two attitudes jointly cause the relevant behavior, and in doing so they manifest the causal powers that are partly constitutive of their being, respectively, a desire and a belief. Because these attitudes also justify the behavior that they cause, that behavior eventuates not only *from causes* but *for reasons*. And whatever we do for reasons is consequently of our making.

Thus, for example: I want to know the time; I believe that looking at my watch will result in my knowing the time; and these two attitudes cause a glance at my watch, thus manifesting their characteristic causal powers as a desire and a belief. The desire and belief that cause my glance at the watch are my reasons for glancing at it; and because I engage in this behavior for reasons, I make it happen.[10]

[10] The example is borrowed from Donald Davidson, 'How is Weakness of the Will Possible,' in *Essays on Actions and Events* (Oxford: Clarendon Press, 1980), 21–42, at 31. Davidson is, of course, the foremost exponent of the standard model.

This model seems right in several respects. To begin with, it treats my making something happen as a complex process composed of simpler processes in which events are caused by other events. I can make something happen even though it is caused by other events, according to this model, because their role in its production can add up or amount to mine. If the model identifies events whose causal role really does amount to mine, then it will have succeeded in reconciling my capacity to make things happen with the causal structure of the world.

The model is at least partly successful on this score. The events that it picks out in the causal history of my behavior are closely associated with my identity, and the causal operations of these events consequently implicate me, at least to some extent. What I want and what I believe are central features of my psychology, which is central to my nature as a person. My wantings and believings are therefore central features of me, and whatever they cause can be regarded as caused by me, in some sense.

The question remains, however, whether the causal role of my desires and beliefs adds up to the role that I play in producing an action or whether alternatively, it amounts to the role that I play in producing a mere activity. The claim made for the standard model is that it is a model of action, in which my capacity to make things happen is exercised to its fullest extent. Is this claim correct?

The standard model is at least correct, I think, about what this claim will require for its vindication. The model assumes that the processes constituting a person's role in producing an action must be the ones that connect his behavior to reasons in such a way that it is based on, or performed for, those reasons. If a person's constitution includes a causal mechanism that has the function of basing his behavior on reasons, then that mechanism is, functionally speaking, the locus of his agency, and its control over his behavior amounts to his self-control, or autonomy.[11]

Why would behavior produced by such a mechanism be any more attributable to the person than that produced by other causes? The answer is that a person is somehow identified with his own rationality. As Aristotle put it, "Each person seems to be his understanding."[12] Hence causation via a person's rational faculties qualifies as causation by the person himself. Of course, this statement raises more questions than it answers; but I hope to answer those questions, too, by the end of this Introduction. For now,

[11] The terms 'autonomy' and 'autonomous' are ambiguous. On the one hand, they express a property that distinguishes action from mere behavior and (I claim) from mere activity as well. On the other hand, they express a property that differentiates among actions, or styles of action. To be subservient or conformist is to lack autonomy in the latter sense. But subservience and conformism can be displayed in actions that are still autonomous in the sense that distinguishes them from mere behavior or activity.

[12] *Nicomachean Ethics* 1178a.

I simply want to endorse this inchoate intuition underlying the standard model of agency.

One might object, at this point, that responding to reasons is the function of an entire person, not of a causal mechanism within him. The phrase "responding to reasons," one might insist, already describes something done by the agent and hence cannot describe a mere chain of events.

To be sure, the concept of basing behavior on reasons belongs to the same conceptual vocabulary as that of performing an action or making things happen, and so it cannot provide the desired reduction of those concepts into the vocabulary of event-causation. But it isn't meant to provide that reduction. "Basing behavior on reasons" is not proposed as an event-causal replacement for agential concepts; rather, it is proposed as that agential concept whose reduction will be the key to reducing the others. In order for a chain of events to constitute a person's making things happen, in the fullest sense of the phrase, it will have to constitute, more specifically, his doing something for a reason.

So says the standard model of action—rightly, in my view. But my agreement with the standard model ends here. The model goes on to say that the chain of events constituting a person's doing something for a reason is that in which his behavior is caused by a desire and belief in the manner that's characteristic of those attitudes. I think that this aspect of the model runs afoul of obvious counterexamples.

Failings of the Standard Model

The standard model already contains a clause designed to rule out some counterexamples, in which behavior is caused by a desire and belief but fails to constitute an action performed for reasons. This clause appears in my formulation as the requirement that the desire and belief causing behavior must exercise the causal powers that are characteristic of those attitudes.

Here is an example in which desire and belief operate uncharacteristically. A speaker's desire to win the sympathy of his audience, and his belief that nothing short of tears would suffice, might frustrate him to the point of tears. In causing behavior through the medium of frustration, his desire and belief would not manifest their characteristic causal powers. Characteristically, these attitudes cause whatever behavior is specified in the content of the belief as conducive to the outcome desired.[13] But the speaker in this case could have

[13] We can imagine a version of this case in which the speaker, upon sensing the purely involuntary flow of tears, is moved to exploit it by actively crying, thus transforming a mere bodily event into an activity. The point is that the difference between the initial event and the

been frustrated to the point of tears by many different beliefs about the difficulty of attaining his goal. It's just an accident that the belief frustrating his desire, and thereby producing tears, is a belief about the necessity of tears. The mechanism thus actuated—that is, the mechanism of frustration—would not in general conform the subject's behavior to the instrumental content of his belief. Hence his motives do not exercise their characteristic powers in causing his behavior.

Adherents of the standard model believe that by ruling out such cases, in which behavior is caused but not motivationally guided by a desire and belief, they have succeeded in narrowing their analysis to behavior that is caused in the right way to qualify as an action performed for reasons. I think that they have made considerable progress in narrowing their analysis to behavior that qualifies as motivated activity. But I do not think that motivated activity necessarily constitutes an action performed for a reason.

Recall the first Freudian slip examined above.[14] The agent wants to destroy his inkstand and he is thereby moved to do what he knows will destroy it. His behavior thus satisfies the standard model, but it doesn't qualify as an action: it's a defective instance of the agent's making something happen.

Note that this example is not ruled out by the requirement that desire and belief exert their characteristic powers in causing behavior. In Freud's explanation of his mishap these characteristic powers are indeed at work. It's no accident that the agent is caused to do what's specified in the content of his belief as conducive to the desired outcome of obtaining a new inkstand: the instrumental content of his belief is what's governing his behavior. So the agent really does brush the inkstand's cover off the desk for the purpose of

subsequent activity would be—not *that* the latter was caused by the speaker's desire and belief—but rather *how* it was caused by them.

[14] Note, by the way, that the second slip does not fit the standard model, because the President knows that he cannot close the session simply by uttering the words "I declare the session closed." Hence his utterance is not motivated by the belief that it will accomplish the desired result.

Other slips of the tongue do fit the standard model, however. Consider, for example, a case reported to Freud by Viktor Tausk. Tausk committed his slip of the tongue when the hostess entertaining him and his young sons began to rail against the Jews, unaware that Tausk himself was Jewish. On the one hand, Tausk wanted to set his sons an example of moral courage in the face of anti-Semitism; on the other hand, he wanted to avoid a scene, which could potentially have ruined the family's vacation. Deciding to hold his tongue, he tried to dismiss the boys from the room, lest they precipitate the confrontation that he had reluctantly decided to avoid: 'I said: "Go into the garden *Juden* [Jews]", quickly correcting it to '*Jungen* [youngsters]". The others did not in fact draw any conclusions from my slip of the tongue, since they attached no significance to it; but I was obliged to learn the lesson that the "faith of our fathers" cannot be disavowed with impunity if one is a son and has sons of one's own' (*Psychopathology of Everyday Life*, 92–3). Tausk wanted to show his sons how they should declare their Jewish ancestry when under social pressure to conceal it; and he was moved to say something that amounted to just such a declaration.

destroying it, unlike the frustrated speaker imagined above, whose tears are shed out of frustration and hence not for any purpose.

In sum, the agent's movement is caused in the way that's designated as right by the standard model; and yet it is only an activity. The standard model thus appears to specify the wrong "right way" for behavior to be caused. It specifies the way in which behavior must be caused in order to qualify as a purposeful activity, but not the way it must be caused in order to qualify as an autonomous action.

If we want to know why the standard model has failed to specify the right way for autonomous action to be caused, we need look no farther than the requirements that the model set for itself. The idea behind the model, remember, is that the causal processes constitutive of action will be the ones in virtue of which behavior is based on or performed for reasons. Those processes are the ones that the model aspires to specify as the right way for action to be caused. But the model has not lived up to its own aspirations: it hasn't specified the processes in virtue of which behavior is based on reasons.

A reason for acting is something that warrants or justifies behavior. In order to serve as the basis for the subject's behavior, it must justify that behavior to the subject—that is, in his eyes—and it must thereby engage some rational disposition of his to do what's justified, to behave in accordance with justifications. When someone just knocks over something that he unconsciously wants to destroy, or blurts out something that he unconsciously wants to say, he has not necessarily seen any justification for his behavior, nor has his rationality been engaged, although he has indeed been motivated by a desire.So although his behavior has been caused by something that may in fact be a reason, it has not been caused in the right way to have been done for that reason.

This flaw in the standard model is papered over, in some versions, by a characterization of desire itself as entailing the grasp of a justification for acting. Engagement of the agent's rationality is thus claimed to be inherent in the very nature of desire.

Desiring something entails regarding it as *to be brought about*, just as believing something entails regarding it as *having come about*, or *true*. And regarding something as to be brought about sounds as if it entails seeing a justification for acting. Proponents of the standard model therefore claim that if a subject desires something, and believes some behavior conducive to it, then he already sees a justification for that behavior, and his responsiveness to reasons is thereby engaged.

Unfortunately, this argument trades on confusions in the language of "seeing" and "regarding as." To say that desiring something entails regarding it as to be brought about is simply to describe the so-called direction of fit that's characteristic of desire. Desire has what is called a mind-to-world direction of

fit, in that its propositional object functions as a model for what it represents rather than as modeled after it. When the President wants the session of the Senate to be closed, for example, he has a mental representation of the session's being closed, and that representation serves as an archtype for the state of affairs that it represents rather than as an ectype of it. But to regard the session's closure as to be brought about in this sense is not to think of it as appropriate or fitting or correct to bring about: it is not to judge that the session's closure is desirable or good. It's simply to hold the thought "session closed" in a conative rather than cognitive mode. Thus, wanting the session to be closed does not entail seeing any justification or warrant for behavior conducive to closing it.

The objection therefore stands that the standard model fails to specify the way in which action involves causation by reasons, although it succeeds in specifying the way in which purposeful activity involves causation by desires and beliefs. The standard model is a model of activity but not of action.[15]

Let me pause for a brief summary. I began by drawing a distinction between mere activity and action, which differ with respect to the degree of the subject's agency—the degree to which he makes things happen. I then posed the question how a person can make things happen, in a world where events are caused by other events. An answer to this question, I suggested, would have to show how causation by events could add up to or amount to causation by a person.

I next examined a standard model of agency, which rests on the premise that a person causes his behavior when it is caused by reasons in such a way that it is based on or performed for those reasons. The model claims that behavior is performed for reasons whenever it is caused by desire and belief in the characteristic way. But some instances of characteristic desire-belief causation yield no more than mere activity, because the resulting behavior is not based on the desire and belief as reasons. Hence the standard model is sufficient for motivated activity but not for autonomous action.

I shall now consider a proposal for improving the standard model by adding to it. This proposal will bring us closer to an account of agency, but still not close enough. My critique of this proposal will suggest a third—and, in my view, correct—account of agency.

Adding to the Standard Model

I have argued that when desire and belief cause behavior in such a way as to operate as its motives, they do not necessarily operate as its reasons—that is,

[15] The argument of this section is developed more fully in 'The Guise of the Good,' (Chap. 5, below).

as reasons for which the behavior is performed. But I do not claim that their being motives for acting somehow excludes their also being reasons; nor do I claim that their operating as motives somehow excludes their also operating as reasons. All I claim is that their operating in the one capacity doesn't amount to their operating in the other. Autonomous action requires something more than motivation by desire and belief, as is demonstrated by motivated slips that are not autonomous; but the "something more" that's required can be provided by the same motivating desire and belief, operating in an additional capacity.

Consider an alternative version of Freud's story, in which the agent not only is motivated by his desire to destroy the inkpot but also acts on the grounds of that desire, in its capacity as a reason. In acting on that desire as a reason, let us suppose, he is aware of the desire and regards it as justifying a movement of his hand; and he makes the movement partly because of seeing it as justified. His desire thus causes his behavior via his disposition to behave in accordance with perceived justifications—a mechanism that wasn't operative in the original story, where the agent was unaware of the justifying desire.

Yet even in the alternative version of the story, where the desire influences the agent via his perception of it as justifying action, it can continue to operate as a motive, as it did when it was hidden from view. The new influence that it now exerts in its capacity as a reason can be to enlist some reinforcement for, or remove some inhibition of, its own motivational force. Even when the subject is persuaded by rational reflection on his desire—a process different from simply being moved by its inherent force—his response to being persuaded can be to acquiesce in being so moved. On the grounds of his desire conceived as warrant, he may accede to its impetus as a motive.

The interaction of these causal mechanisms is not as mysterious as it may sound. Suppose that you were charged with the task of designing an autonomous agent, given the design for a mere subject of motivation.[16] If you like, you can imaginatively assign yourself to divine middle-management as project leader for the sixth afternoon of creation; or you may prefer to take the role of natural selection over the corresponding millennia. In either case, you face a world already populated with lower animals, which are capable of motivated activity, and your task is to introduce autonomous agents.

In neither case would you start from scratch. Rather, you would add practical reason to the existing design for motivated creatures, and you would add

[16] Michael Bratman has pointed out to me that the methodology of "creature design" was first proposed by Paul Grice, in his Presidential Address to the APA, 'Method in Philosophical Psychology (From the Banal to the Bizarre),' *Proceedings and Addresses of the APA* 48 (1975) 23–53. Bratman uses the same methodology, to reach different conclusions, in a paper entitled 'Valuing and the Will,' (MS).