Philos Stud (2011) 155:345–369
DOI 10.1007/s11098-010-9577-1

The metaphysics of rule-following

Markus E. Schlosser

Published online: 7 July 2010

Abstract This paper proposes a causal–dispositional account of rule-following as it occurs in reasoning and intentional agency. It defends this view against Kripke's (Wittgenstein on rules and private language, Harvard University Press, Cambridge, 1982) objection to dispositional accounts of rule-following, and it proposes a solution to the problem of deviant causal chains. In the first part, I will outline the causal–dispositional approach. In the second part, I will follow Martin and Heil's (Philos Perspect 12:283–312, 1998) realist response to Kripke's challenge. I will propose an account that distinguishes between two kinds of rule-conformity and two kinds of rule-following, and I will defend the realist approach against two challenges that have recently been raised by Handfield and Bird (Philos Stud 140:285–298, 2008). In the third part, I will turn to the problem of deviant causal chains, and I will propose a new solution that is partly based on the realist account of rule-following.

Keywords Rule-following · Reasoning · Deviant causal chains · Dispositions · Metaphysics of mind · Metaphysics of agency

1 The causal–dispositional approach

What happens, metaphysically speaking, when we make an inference or perform an action on the basis of reasoning? It is widely assumed that instances of reasoning and rational agency consist in transitions of certain kinds of states and events: they are chains of mental states, mental events, and overt behavior. Paradigm examples are the formation of beliefs on the basis of antecedent beliefs (theoretical reasoning), the formation of intentions on the basis of beliefs, desires, and prior

M. E. Schlosser (✉)
Department of Philosophy, University of Leiden, PO Box 9515, 2300 RA Leiden, The Netherlands
e-mail: m.schlosser@hum.leidenuniv.nl

intentions (practical reasoning), and the performance of actions on the basis of beliefs, desires, and intentions (action based on practical reasoning). What they have in common is that the output is *based on* the mental antecedents. It is controversial how exactly this basing relation should be construed.[1] But the following two necessary conditions provide a plausible starting point. Firstly, it is plausible to assume that the antecedents must *cause* the output, because it seems clear that an output is based on mental antecedents only if the agent forms the output *because of* having the antecedents (Davidson 1963; Harman 1973; Goldman 1979, for instance). Secondly, it is plausible to assume that the mental antecedents must *rationalize* the output. By this I mean, roughly, that the formation of the output must appear intelligible in the light of the antecedents. This is a standard condition in the theory of action (Davidson 1963; Mele 2003, for instance). But it is also a plausible condition on causal chains of reasoning in general (compare Wedgwood 2006). We will return to this further below.

A defining feature of reasoning, I take it, is that it is subject to normative assessment. It is correct or incorrect, rational or irrational, or simply good or bad. It is plausible to assume that reasoning is correct or good just in case it conforms to the relevant norms and rules of rationality, and that an agent who reasons correctly is in some sense *governed* by those norms and rules. What does this rule-governance consist in?

Let us consider a paradigm example, in which an agent's reasoning is governed by *modus ponens*. The agent forms the belief that q in response to considering the beliefs that p and that if p then q. Why and in what sense is the inference governed by *modus ponens*? In order to give an answer, one might assume that the agent has a background belief with the inference rule of *modus ponens* as its content. But the force and application of this additional antecedent would itself be in need of explanation. It would, in particular, itself require the application of *modus ponens*. Reference to yet another background attitude with another rule as its content would quite obviously lead to a vicious regress (Boghossian 2005, for instance).

An alternative approach explains the rule-governance in terms of dispositions. It says that the agent's reasoning is governed by *modus pones* because the agent is manifesting a disposition to make inferences in accordance with that rule. This explains the inference by reference to what can be called a *rational* disposition. It is a rational disposition because it disposes the agent to make inferences in accordance with a rational rule, and because the agent is a rational *agent* in virtue of possessing dispositions of this kind.[2]

Philosophers often focus on either the causal or the dispositional component of reasoning. It is not always clear whether this is due to a decision to focus on one of the two components, or whether they mean to give either a purely causal or a purely dispositional account. In any case, we have just seen that there is, prima facie, good reason to pursue an account that is *both* causal *and* dispositional. How should we

[1] Note that the fact that a belief is based on mental antecedents does not entail that it is *justified*. Arguably, the basing relation is necessary for justification, but not sufficient (compare Swain 1981).

[2] Compare Baker (2008), who argues that mental state transitions that are not based on reason-giving inferences can be rational, provided that they result from manifestations of the right kind of disposition.

combine the appeal to causality and dispositionality? Let us consider a simple example.

Suppose that a baseball, being thrown hard enough, shatters a window. What happens, according to the causal approach, is that one event (the baseball's hitting the window) causes another event (the window's shattering). According to a dispositional approach, what happens is that the window is manifesting its disposition to break in response to the stimulus. A combination of the two approaches seems straightforward, as the cause-event just is the occurrence of the disposition's stimulus condition, and the effect is its manifestation. Both events involve the window, and it seems that the cause-event causes the effect partly in virtue of the window's disposition to manifest the effect-type in response to the occurrence of the cause-type. This disposition is clearly causally relevant, and we might say that it *mediates* between the events, as the cause and the effect are the stimulus and the manifestation of this disposition. This admits of at least two interpretations.

We may construe the disposition not only as causally relevant, but as a separate cause. It seems clear that it cannot be construed as a causally efficacious event. But we can think of it as being a *structuring cause*.[3] Alternatively, we could construe the event-causal relation as consisting solely in the manifestation of the disposition in response to the stimulus. This interpretation does not deny the existence of the event-causal relation, but it amounts to a reductive account of the causal relation in terms of dispositionality. I will not attempt here to adjudicate between the two interpretations. Both provide a plausible integration of event-causation and dispositionality, and it does not matter, for our purposes, which one we choose.

An application to the case of reasoning is straightforward. Return to the example of following *modus ponens*. The agent forms the new belief in response to considering the antecedent beliefs and is thereby manifesting a disposition to form beliefs in accordance with *modus ponens*. Again, it seems clear that the cause (the agent's considering the antecedent beliefs) just is the disposition's stimulus condition and that the effect (the new belief) just is its manifestation. In other words, the agent forms the new belief partly because he or she is disposed to form beliefs in accord with *modus ponens*.

This illustrates the causal–dispositional approach to the metaphysics of reasoning and rule-following that I will defend in this paper. It construes reasoning-based inferences and actions in the way outlined: whenever an output of type E is based on mental antecedents of the types $A_1,\ldots, A_n$, the antecedents cause the output and the agent is thereby manifesting a disposition to form an output of type E in response to antecedents of the types $A_1,\ldots, A_n$. I shall assume that reasoning need not involve the conscious and explicit consideration of reasons, as it may consist in tacit, fast, and automatic transitions that are not at the forefront of the agent's mind, as it were.

My aim here is not to develop this account in detail, but to defend the causal–dispositional *approach* against two serious challenges: Kripke's objection to dispositional accounts of rule-following and the problem of deviant causal chains.

[3] Dretske (1988) distinguishes between *triggering* and *structuring* causes. The former are plausibly associated with event-causes and the latter with the causal influence of states.

The main reason why I engage with the latter problem in connection with the former is that the solution to the rule-following problem will pave the way for a new solution to the problem of deviant causal chains. But I should also note that these two problems are, in my opinion, the two most pressing metaphysical problems for the causal–dispositional approach.[4]

2 Kripke's challenge

In his discussion of Wittgenstein's rule-following paradox, Kripke (1982) has identified a difficult problem for the dispositional approach to rule-following. Kripke, and many of his commentators, did make substantial assumptions concerning the connections between language mastery, meaning, and the ability to follow rules. My concern here is with the phenomenon of rule-following as it occurs, more narrowly, in reasoning and agency. I will not make or discuss any assumptions concerning language mastery and meaning, and I will also set aside related epistemological issues. My focus is on the metaphysics of rule-following and the metaphysical problem that arises for the dispositional approach.

To begin with, consider the following variation on Kripke's main example. Suppose that John is adding 68 and 57 (either by means of mental arithmetic or with the help of pen and paper). The sum is 125, but John's answer to the task is "115". It is only natural to think that John made a mistake when he tried to add the numbers. We would assume, in other words, that John made a mistake in an attempt to follow the plus-rule.

Kripke pointed out that it is perfectly possible that John was in fact following some other rule. It is possible, in particular, that he followed some other rule *correctly*. I will express this by saying that John may have followed a *deviant* rule.[5] This gives rise to the following metaphysical question: in virtue of what fact is it true that John made a mistake in following the plus-rule and false that he followed a deviant rule?

According to a simple dispositional account of rule-following, the possession of the ability to follow a rule consists in the possession of a disposition to reason and act in accordance with that rule. Kripke argued that this view cannot answer the metaphysical question. The argument says, basically, that a dispositional account cannot identify a fact that would ground the claim that an agent made an error in following a rule (rather than following a deviant rule), because an appeal to dispositions can only ever account for what agents actually do. Answering "115", John did what he was disposed to do, and so he manifested a disposition to conform to some deviant rule.

This has three important ramifications. It seems to show, firstly, that the dispositional account cannot explain *mistakes* in rule-following, as it cannot

[4] Many philosophers would also name the problem of mental causation as an equally difficult problem. But I think that non-reductive physicalism *without downward causation* can provide a coherent and tenable account of mental causation (compare Owens 1989 and Gibbons 2006, for instance).

[5] We could follow Kripke and say that John may have followed the *quus*-rule, which is defined as follows: x *quus* y = x plus y, if x and y < 57, and 115 otherwise (1982, p. 9).

distinguish between incorrect rule-following and deviant rule-following. This, in turn, seems to show that the dispositional account cannot even explain *correct* rule-following: for any given case of apparently correct rule-following, it cannot rule out the possibility that the agent is following some deviant rule instead. Thirdly, that seems to show that the dispositional account cannot explain the kind of *normativity* that is characteristic of rule-following: it fails to identify a fact that determines what correct rule-following would amount to (or would have amounted to), and so it fails to identify a fact that determines what the agent should do (or should have done).

In defense of the dispositional approach, Martin and Heil (1998) have pointed out that Kripke's argument depends on the assumption that dispositions can be analyzed in terms of conditionals. The general schema for this conditional analysis is, basically, the following (where *S* is the bearer of the disposition, *A* is a response-type, and *C* is a stimulus-type):

S is disposed to *A* in response to *C* if and only if *S* would *A* if *C* obtained.

Given this reductive analysis, and given the dispositional account of rule-following, error is impossible and the normative dimension of rule-following remains unaccounted for. Whenever an agent *S* gives an incorrect answer, according to some rule *R*, it follows that *S* is not disposed to give the correct answer. From this it follows, in conjunction with the dispositional account, that *S* does not possess the ability to follow *R*. Moreover, if *S*'s answer is in accordance with a deviant rule, then it follows that *S* is following this deviant rule (compare ibid, p. 289, see also Handfield and Bird 2008).

But the conditional analysis of dispositions is subject to counterexamples. There are two main types. There are examples that involve so-called *finks*, and there are examples that feature *masks* (or *antidotes*).[6] A fink is a mechanism or agent that prevents the manifestation of a disposition by removing its categorical basis when its stimulus is about to occur. One standard example is Lewis' sorcerer, who protects a glass from breaking by changing its intrinsic properties when it is about to be struck (Lewis 1997). Masks and antidotes also prevent the manifestation when the disposition's stimulus occurs, but they do so without changing its categorical basis. Adjusting the example, we can think of the sorcerer as protecting the glass by shielding or masking it.

Examples that feature finks and masks show that the left-to-right conditional of the analysis is false. The glass is fragile, but it is false that it would break when struck. In general, the fact that *S* would not *A* if *C* obtained does not entail that *S* is not disposed to *A* in response to *C*. Appeal to finks and masks can also show that the right-to-left direction of the biconditional does not hold. One can imagine mechanisms or agents that bring about the characteristic manifestation of a disposition in an object that does not possess this disposition. This shows that an object can lack a disposition even though it would manifest the characteristic response.

Given all this, there is good and independent reason to reject the conditional analysis and to embrace a broadly non-reductive and realist approach to

[6] Johnston (1992), Martin (1994), and Lewis (1997). The outline provided here follows Handfield and Bird (2008).

dispositionality. Returning to our example, this allows us to make the following crucial assumptions. We can assume now that John did possess a disposition to follow the plus-rule, even though he did not answer in accordance with that rule. We can assume, moreover, that the manifestation of this disposition played a role in his giving the wrong answer. Assuming a non-reductive and realist view, we can think of dispositions as interacting and interfering with each other. We can assume, in other words, that the manifestation of some other disposition, such as the disposition to lose concentration, interfered with the manifestation of the disposition to conform to the plus-rule, and that John's answer resulted from an interaction of these two dispositions (compare Martin and Heil 1998, pp. 290–291).

Handfield and Bird, however, have identified two problems for this realist approach. The first stems from the fact that the interfering factors in reasoning and rule-following are often *intrinsic* dispositional properties of the agent. The second is what they call the *privileging problem*.

2.1 Intrinsic finks and masks

In the standard counterexamples to the conditional analysis, the finks and masks are *external* to the objects that possess the dispositions in question. These examples feature usually an object with a salient disposition, such as a fragile glass or a conductive piece of wire, and some external mechanism or agent that would change or protect the object. Handfield and Bird (2008, pp. 289–292) point out that there is good reason to construe finks and mask as external to the disposition's bearer. For if they were a part or an intrinsic property, then it would seem that the object does not possess the disposition in question. Consider, for instance, the fire in a household fireplace that has the disposition to set the house on fire. One can imagine finks or masks that reliably protect the house from catching fire, and that are parts or intrinsic properties of the house. If there is no such fink or mask, the house has the disposition to be set on fire. But if there is an *intrinsic* fink or mask of this kind, then it seems that the house does not possess this disposition at all. It seems, in other words, that intrinsic finks or masks would remove or obliterate the very disposition they are supposed to interfere with.

On the basis of this, Handfield and Bird argue that the dispositional approach to rule-following faces the following problem. If the interfering conditions in rule-following are intrinsic, they remove or obliterate the dispositions they are supposed to interfere with. The interference in rule-following often *is* due to an intrinsic property of the agent, such as the disposition to lose concentration, which means that the interfering conditions in rule-following often do remove or obliterate the dispositions *that are supposed to account for the agent's rule-following*.

Note, first of all, that intrinsic finks and masks give rise to an obvious problem only if we assume that they are *reliable*, in the sense that they would remove the disposition or protect the object *whenever* the stimulus was about to occur. But the interfering conditions in rule-following are usually not reliable in this sense. In our example, there is simply no reason to think that John would make a certain type of mistake whenever he was trying to add numbers. More importantly, finks *fully* remove or alter the categorical basis of their dispositions, and masks *fully* prevent

their manifestations. Both render the dispositions altogether inefficacious. But this is not how we think about the interfering condition in our example and in other cases of rule-following. John is trying to add numbers. According to the dispositional approach, this involves the manifestation of a disposition to conform to the plus-rule, and we assume that the incorrect answer results from the interaction or *joint manifestation* of this disposition to conform and some other disposition to make mistakes. But the latter neither removes the former disposition, nor does it altogether prevent its efficacy. In this respect, the interfering conditions in rule-following differ clearly from both finks and masks. Given this, however, there is no obvious reason to think that the interfering conditions in rule-following cannot be intrinsic properties of the agent.

One may wonder, of course, how all this works. It seems that there is only one way in which an agent can manifest a disposition to conform to the plus-rule—namely, by giving the correct answer. How can one manifest this disposition by giving the incorrect answer? Can dispositions be partly manifested? How can they combine with interfering conditions?

Note, first of all, that partial manifestations of dispositions are not unusual. Consider a sugar cube that dissolves partly, a fragile glass that cracks without breaking, and so on. More importantly, many higher-level cognitive abilities, such as the ability to follow the plus-rule, consist in various sub-abilities to follow certain sub-rules. Whether we are adding in our heads, by means of pen and paper, or by means of using some other tool, we are always applying a number of sub-rules that constitute the steps of some procedure or algorithm (unless the numbers are very small, of course). Given this, we can say that our ability to follow the plus-rule is constituted by sub-abilities to follow certain sub-rules. Accordingly, the dispositional approach should construe the ability to follow such rules in terms of a number of dispositions that correspond to the abilities to follow the relevant sub-rules. Given this, we can then assume, plausibly, that John made the mistake because he failed to follow one of the sub-rules. On the face of it, it seems that John answered "115" (in response to 68 + 57) because he forgot to carry the number 1 after adding 7 and 8. In any case, it is plausible to assume that John made a mistake in following the plus-rule, because he made a mistake in following a sub-rule (or because he failed to follow a sub-rule). Given this, we can assume that the manifestation of a disposition to make mistakes interfered with the manifestation of the disposition to conform to the plus-rule by interfering with the manifestation of a disposition to conform to a sub-rule.

This shows how an interfering disposition can combine with the disposition to conform to the plus-rule so that it interferes only *partly*, resulting in a partial manifestation of the disposition to conform to the plus-rule. We might say here that it is both false and true that the agent manifests this disposition. It is false in the sense that the agent does not give the correct answer, and it is true in the sense that the disposition is partly manifested. The important point is that the relevant disposition is *efficacious* or *operative*, even though it is not fully manifested.

Not all mistakes in rule-following are of this kind and of this complexity. If they were, we would face regresses of ever more basic sub-rules. Suppose, for instance, that John made the mistake because he forgot to carry, and that he forgot to carry

because he got distracted. Given this, John made the mistake in following the plus-rule by virtue of making this particular mistake, which is explained by the fact that he got distracted. There is no further mistake. Given that he forgot to carry, it is not even correct to say that he made a *mistake* in following the sub-rule in question. He simply forgot, and so he did not even try to follow the corresponding sub-rule.

Other mistakes may be solely due to lapses in memory. Suppose, for instance, that John is prone to make mistakes in multiplying small numbers. This may be entirely due to weak and unrehearsed memory functions. John may be disposed to make such mistakes not because he possesses dispositions to make certain mistakes, but simply because his disposition to retrieve the answers from memory is weak and unreliable.

A further objection stems from Handfield and Bird's observation that we all have *permanent* intrinsic dispositional properties that appear to undermine certain rule-following dispositions. In the case of the plus-rule, these are properties such as the inability to hold very long numbers in our heads or the inability to write them down (pp. 292–293). The fact, however, that one is unable to add very large numbers raises a problem only if we think of the ability to follow the plus-rule as constituted by one single disposition. The dispositional account is not committed to this, and it is plausible, as I have just argued, to construe such abilities in terms of sub-dispositions that correspond to the relevant sub-rules. Given this, the inability to hold long numbers in our heads, or to write them down, does not raise any obvious problems, because there is, then, no good reason to think that we must be able to add very long numbers in order to be able to follow the plus-rule. In order to make this clearer, I shall briefly consider the related objection that the dispositional account cannot accommodate the fact that many rules range over an infinite range of cases, as human agents can possess only a finite number of dispositions that can be manifested only on a finite number of occasions.

Following Martin and Heil, Handfield and Bird point out that this objection misconstrues of the nature of dispositions. Arguably, dispositions are entities that are necessarily directed at an indefinite range of characteristic manifestations (pp. 286–287). We can add to this the following. In order to possess the disposition to conform to the plus-rule, an agent need not possess infinitely many dispositions. Nor is it required that the agent possesses a single disposition that can be manifested, fully and correctly, for any two numbers. Rather, an agent can possess a disposition to conform to a rule that covers an infinite range of instances by virtue of possessing a finite number of dispositions that constitute the ability to follow certain procedures or algorithms. Following the right sub-rules, the agent can at least *begin* to follow a rule such as the plus-rule, for any two numbers, by beginning to apply the sub-rules. And the agent can be said to begin to follow the plus-rule *correctly* or *incorrectly*, for any two numbers, because the agent can begin to apply the sub-rules correctly or incorrectly.

2.2 The privileging problem

We assume that the manifestation of a disposition to make a certain type of mistake can interfere with the disposition to conform to a rule. According to Handfield and

Bird, interfering conditions are often intrinsic and permanent properties that *dispose* the agent to give the *incorrect* answer (p. 294). In our example, this would be a disposition to answer "115" in response to 68 + 57. Let us assume that John possesses a disposition to make a specific type of error. Suppose that he is disposed to forget to carry. Adding 68 to 57, John has to carry only once, and so it seems that John is disposed to answer "115" by virtue of being disposed to forget to carry.[7] Given, moreover, that this is the correct answer according to some deviant rule, John is manifesting a disposition to conform to a deviant rule.

This shows, according to Handfield and Bird, that some cases give rise to a privileging problem: the problem to find a principled reason to give privilege to one particular disposition as being causally relevant to the agent's behavior (pp. 294–296). In our example, that is the problem to find a principled reason to give privilege to John's disposition to conform to the plus-rule (over the disposition to give an answer in accordance with a deviant rule).

First, we need to be clear about what the challenge consists in. Consider the following claims about John's case:

(1) John is disposed to answer "125" in response to 68 + 57 because he has the disposition D_R to conform to the plus-rule.
(2) John has the disposition D_F to forget to carry.
(3) John makes a mistake in following the plus-rule and gives the wrong answer "115" in virtue of manifesting both D_R and D_F.
(4) John manifests the disposition D_D to answer "115" in response to 68 + 57.
(5) D_D is a disposition to conform to a deviant rule.
(6) John follows a deviant rule in virtue of manifesting D_D.

It is important to note that the conjunction of the claims (1)–(4) alone does not give rise to a privileging problem. First of all, the fact that more than one disposition is causally relevant does not by itself give rise to a privileging problem, because we explain why John makes a mistake in following the plus-rule by appeal to the manifestation of *both* D_R and D_F. This shows, also, that (4) does not give rise to a privileging problem, because it seems clear that the possession (and manifestation) of D_D is nothing over and above the joint possession (and manifestation) of D_R and D_F.

Claim (4) does not give rise to a privileging problem as long as we can account for the claim that John makes an error in following *one* particular rule, the plus-rule, and as long as the manifestation of D_D does not constitute an instance of deviant rule-following. We would, however, have to find a principled reason to give privilege to D_R if there was a disposition that constitutes an instance of deviant rule-following. This is the claim made by (5) and (6). Neither (5) nor (6) follow from (1)–(4), which shows that the problem arises only if we add (5) and (6) to (1)–(4).

[7] The fact that John answers "115" in response to 68 + 57 does not by itself show that he is disposed to give that answer. Given that we reject the conditional analysis of dispositions, it shows only that John is able to give this answer, in a broad sense of *ability* or *capacity*. Note also that not all cases are as straightforward. If John has to carry several times in adding longer numbers, for instance, it is not clear that he is disposed to give a particular answer by virtue of being disposed to forget to carry.

John is disposed to answer "115" because he is disposed to forget to carry. Given this, it seems undeniable that John is disposed to answer in conformity with some deviant rule. This is claim (5). Does this show that John *follows* some deviant rule? According to Handfield and Bird, the simplest and most natural dispositional account of rule-following says just this: an agent possesses a rule just in case the agent has a disposition to give responses in conformity with that rule (p. 294). Given this simple account, John follows a deviant rule in virtue of manifesting a disposition to conform to it. And given this—claim (6)—we face a privileging problem. We want to say that John made a mistake in following the plus-rule, and we want to deny that he followed a deviant rule. On the basis of what further fact can we give privilege to John's disposition to conform to the plus-rule?

2.3 Dispositions and rule-conformity

The privileging problem arises for the simple dispositional account of rule-following. But the simple account is neither particularly plausible, nor is the causal–dispositional approach committed to it. In this section, I will introduce an important distinction between two kinds of rule-conformity, and in the following section, I will propose a further distinction between two kinds of rule-following. Once these distinctions are in place, we will be able to see that the simple account is inadequate, and we will be in a position to solve the privileging problem.

According to the causal–dispositional approach, rule-following involves the manifestation of dispositions to conform to rules. It should be noted that not every manifestation of a disposition conforms to a rule or norm. Manifesting its disposition to dissolve in water, the sugar cube is in accord with the laws of nature (or the laws of science). But it does not conform to a rule. Rules and norms set normative standards for evaluation, and not every manifestation of a disposition is subject to a normative standard.[8]

More importantly, we can conform to rules without following them. This can come about in two rather different ways. Firstly, we may conform to rules without following them consciously. Consider an example. Suppose that when John left his house in the morning he was thinking about an upcoming business meeting. On the first pedestrian crossing he noticed that the lights were red, and so he stopped. John, it seems, was conforming to a rule (a conventional rule of traffic). We can assume, plausibly, that he did not consciously intend to conform to the rule, and that he did not have any occurrent thoughts about it. John, we can assume, was enacting a habit that he had acquired some time in the past. Nevertheless, it seems that John's behavior was governed by the rule.

Secondly, we may conform to rules without following them at all. Suppose that John has the habit of responding to business related e-mails within two days. This is in conformity with a rule of the company where he is employed. But, as it happens, John is not aware of this (assume that John is new in this job and that his superior forgot to mention the rule). Again, John is in conformity with a rule due to a habit.

[8] This holds, of course, also for many types of behavior exhibited by rational agents. When John scratches the back of his head, for instance, he neither conforms to nor violates any rule or norm.

But this case differs significantly from the previous one. It is a *coincidence* that John conforms to the company rule, whereas it is no coincidence that he stops when the lights are red in conformity with the rules of traffic.

Typically, a rule or norm prescribes some pattern of behavior or thought. We assume that the possession of the ability to follow a rule presupposes the possession of a disposition to conform to it. The fact that an agent is manifesting a rule-conforming disposition is part of the reason why the agent is governed by the rule. But we can see now that this is not sufficient, because the rule-conformity might be accidental. Given this, it seems clear that an agent follows a rule only if the agent's rule-conformity is not accidental.

What does the difference between accidental and non-accidental rule-conformity consist in? In general, an instance of rule-conformity is non-accidental if there is an explanation of why the agent is conforming to the rule. But not any explanation will do. It seems that an adequate explanation of why an instance of rule-conformity is non-accidental must somehow establish a link between the disposition and the rule. It must explain why the agent is governed *by the rule*. I shall express this by saying that an adequate explanation must explain *rule-governance*.

How can this be done? Rules and norms are propositions (abstract entities, perhaps). How can they explain why an agent possesses and manifests a certain disposition? Our example gives an indication of how this is possible. Stopping at the lights, John is enacting a habit. We can assume, plausibly, that he has acquired the relevant disposition in a process of learning *to conform to the rule*. An explanation of why John has acquired this disposition would have to refer to the learning process, and so it would have to refer to the rule. This, it seems, would establish the required link between the disposition and the rule, and it would provide an explanation of why John's behavior is governed *by the rule*.

Given this, it seems plausible to assume that, in general, an agent is governed by a rule if the agent manifests a rule-conforming disposition that has been acquired in a process of learning to conform to that rule. Further, it seems plausible to assume that learning of rule-conformity can come about by way of various kinds of implicit and explicit learning mechanisms, such as association, conditioning, habituation, imitation, trial-and-error, and so forth.

Alternatively, it might be possible to explain rule-governance by appeal to natural selection and innate cognitive abilities. The idea would be, roughly, that the manifestation of a rule-conforming disposition constitutes an instance of rule-governance either if the disposition's possession is innate, or if its acquisition is constrained by innate mechanisms, such that its possession or acquisition can be explained in terms of the adaptive function of the corresponding rule-conforming pattern. Whether or not this appeal to adaptation and innateness can in fact provide alternative explanations of rule-governance is an open and largely empirical question.[9] It seems only plausible to think that we learn many abilities, including abilities to conform to various rules, by means of more basic abilities that are not themselves learnt. But even if this is correct, it is still an open question whether the

[9] For speculative accounts of how norm-guided behavior can be explained in terms of biological adaptation see Gibbard (1990) and Millikan (1990).

possession of these basic abilities consists in the possession of dispositions that conform to *normative standards*. In other words, it is not clear whether the appeal to adaptation and innateness would give us alternative explanations of rule-governance, or whether it would merely supplement explanations in terms of learning.[10] In any case, it should be noted that an appeal to innateness would not entail a crude genetic determinism, because the development of the relevant dispositions may depend, partly and crucially, on interactions within the organism and on interactions of the organism with the right kind of environment (compare, for instance, Elman et al. 1996).

Given all this, I propose that all instances of rule-governance can and must be explained as follows: either by appeal to learning *or* adaptation (and innateness), or by appeal to learning *and* adaptation (and innateness).[11] This proposal raises a host of questions, and I shall now address some possible objections that will help me to clarify the view.

The proposal says that rule-governance can be explained, partly at least, in terms of learning. But, as Wittgenstein and Kripke have argued, any student may misunderstand the teacher. In particular, the student and the teacher may have different rules in mind, which are both compatible with the examples given in the learning process. What rule does the student learn? Would appeal to either the student's or the teacher's understanding of the rule not beg the question or lead to a vicious regress?

This objection is partly due to a misunderstanding. It confuses the proposed view with a mental state account, according to which facts about meaning and rule-following are to be explained by facts about the agent's occurrent mental states. It has been shown, conclusively I think, that this view is hopeless as an account of meaning (Kripke 1982; McDowell 1984; Boghossian 1989, for instance). But we have already rejected the related mental state account of rule-following, when we rejected the mental state account of what it is to follow *modus ponens* (see above).

There is, however, a related objection to the proposed dispositional account. This view assumes that rule-following is grounded in rule-conforming dispositions. An agent must acquire the relevant dispositions at some point in time, and the view suggests that this must occur in processes of learning (and, perhaps, in processes in which innate abilities are developed). But, the objection says, it is difficult to see

[10] Compare Evans and Over (1996, pp. 147–148). For a discussion of the relation between innate knowledge and learning in neural networks see, for instance, Bechtel and Abrahamsen (1990) and Elman et al. (1996).

[11] Millikan (1990) has also proposed a solution in terms of learning and biological adaptation. On her view, following a rule consists, basically, in conforming to a biological norm. This view explains the difference between rule-conformity and rule-following, and it can account for the normative aspect of rule-following. But there are significant differences to the solution proposed here. Firstly, we assumed a realist conception of dispositions in order to explain why one can be disposed to conform to a rule even when one makes a mistake. Millikan's view makes no claims concerning the nature of dispositions, and it does not explain why dispositions to conform can be ascribed when one fails to conform. Secondly, Millikan's view is committed to the rather contentious claim that all the standards for rule-following are either identical with or derived from biological norms. In particular, learning of new rules is restricted to rules that derive from biological norms. My view is compatible with this strong claim, but it is not committed to it. Thirdly, Millikan identifies rule-following with what I have called rule-governance. But I will distinguish, further below, between two types of rule-following on top of rule-governance.

how learning processes can possibly establish the required links between dispositions and general rules that cover infinite domains of possible applications, as learning processes cannot possibly exemplify such rules.

I argued that there must be an explanation of why a given instance of rule-conformity is not accidental. In order to explain this, we do not have to establish a link between the disposition and the rule *as such*. In fact, if rules are propositions or abstract entities, and if the required links are supposed to be causally explanatory of the formation of dispositions, then it is difficult to see how anything could establish such links. Provided, however, that the relevant learning processes exemplify finite patterns that fall under the corresponding rules, and provided that they result in the formation of rule-conforming dispositions, there is an explanation of why subsequent instances of rule-conformity are not accidental. They are not accidental, because their agents manifest dispositions that have been acquired in processes in which the corresponding rules have been exemplified. The assumption that it is not only possible for us to form general rule-conforming dispositions in response to finite learning processes, but that we also tend to form such dispositions, is part and parcel of the realist view. It is bedrock—the view *assumes* that we, and other animals, just are like that.[12]

But this raises the question of why it is legitimate to assume that agents acquire the *right* rules. Appeal to finite patterns that are exemplified in learning processes does not help, as any number of deviant rules and rule-conforming dispositions can be made out to fit any finite pattern.

It is crucial to remember here that the proposed realist view aims to establish a *possibility*: it aims to show that facts about dispositions *can* ground facts about rule-following. In order to do this, it assumes, for the sake of argument, that agents possess rule-conforming dispositions. This, I think, suffices in response.[13] But we can say more, and it will be helpful to compare the objection to a related point that Noam Chomsky made in his celebrated poverty of the stimulus argument against behaviorist theories of language acquisition.

Chomsky (1959, 1965, for instance) argued that even if we grant that children form general dispositions that conform to the right language rules on the basis of finite examples and experiences, we still need an explanation of why they acquire those dispositions rather than others, because there is an indefinite number of alternative rules that are consisted with their experiences. According to Chomsky, only an appeal to innate knowledge of language rules or grammar structures can provide the missing constraints on rule acquisition. Now, I have already suggested

[12] This does not mean, of course, that we have no idea how rule-conforming dispositions can be acquired. We may, for instance, appeal to neural network models of learning. It is controversial whether or not these models can explain how abstract cognitive rules are represented, but it is generally agreed that they can explain how rule-conforming patterns are acquired and implemented (compare Bechtel and Abrahamsen 1990, for instance).

[13] One may wonder, also, how we can *know* that the right dispositions are formed. Here, I am defending a metaphysical solution to a metaphysical problem, and questions concerning the epistemology of rule-following are beyond the scope of this paper. Note, however, that there is good reason to think that the problem of rule-following is, at its core, a metaphysical problem and that Kripke's exposition in the form of a dialogue with a skeptic is misleading, as it may suggest, falsely, that the challenge is epistemological skepticism (compare Boghossian 1989, pp. 515–516).

that explanations of rule-governance in terms of learning might have to be supplemented by an appeal to innate abilities. So, if it is true that one can explain why we acquire the right dispositions in terms of innate knowledge or innate abilities, and if this holds for rule-conforming dispositions in general, then we have already made reference to a theoretical framework that can explain why agents tend to acquire the right dispositions.

It is widely thought that the question of whether or not language acquisition is based on innate knowledge or abilities is largely an empirical question. Likewise, it is only plausible to think that the question of whether or not rule acquisition is in general based on innate knowledge or abilities is also largely an empirical question. The important point for us is that we can see, in broad outline, how it is possible to explain rule-governance in terms of learning and innate abilities. The question of how this works, in fact and in detail, is not only beyond the scope of this paper, but beyond the scope of philosophy. But let me add a few more remarks on this. Firstly, a nativist view about rule acquisition is perfectly compatible with the claim that most rules are learnt. It may claim only that the knowledge or possession of some basic rules must be innate, and that the acquisition of further non-basic rules is in some sense based on or enabled by the basic rules. Secondly, I see no obvious reason why such a nativist view should have to refer to the innate *knowledge* or *representation* of rules (compare Elman et al. 1996). On the view that I have proposed, rule-governance is not grounded in mental states, but in dispositions that can be said to *embody* the rules. Thirdly, an appeal to innateness need not be an appeal to domain-specific abilities (to follow domain-specific rules). It may, instead, consist in an appeal to general cognitive abilities, such as the ability to recognize patterns or to extrapolate in certain ways, that are innate and that help to explain why we acquire the more specific rule-conforming dispositions (compare Tomasello 2003).

Finally, one might object that the proposed view is simply implausible. Consider the following example. Suppose that John's friend Sue brings along a new board game. The rules are very simple. John can memorize them at once, and he is able to follow them immediately. It seems implausible to suggest that John must have acquired this ability in a process of learning in which he has acquired dispositions to conform. And it seems that there is no need for this assumption, because the fact that John *intends* to follow the rules explains why the subsequent conformity is not accidental.

However, an explanation in terms of the agent's intention is just another incarnation of the already rejected mental state view. Suppose that John intends to follow some rule *R*. In order to apply *R* in particular cases, John must detach the consequent of *R* in response to recognizing that *R*'s antecedent obtains. But this is just to say that John must apply another rule (*modus ponens*, presumably), which gives rise to yet another instance of the familiar regress.

Further, John does not learn to follow a new *type* of rule. We can assume, plausibly, that the rules of the game have a common and familiar logical form, such as *modus pones* or *disjunctive syllogism*. Presumably, John already possesses the ability to follow rules with that form. There is, of course, a sense in which John learns new rules. The particular contents of the rules are new for John, and he must

memorize them in order to be able to apply them. This presupposes that John is able to apply a given rule to new cases. But this assumption is by no means extravagant. Even low-level abilities, such as face recognition, presuppose the ability to apply given skills to new cases. With this assumption in place, we can construe examples of this kind as cases of rule *application*, which are cases of rule *acquisition* only insofar as the agent learns new instances of familiar types of rules.

The same can be said about moral or social norms. For instance, when John learns and accepts a new moral rule, then it is likely that the content but not the form is new for him. Moral and social rules usually take the form of imperatives, and we can assume that John already possesses the ability to conform to imperatives. In this case, following the new rule may not be as easy as in the previous example. But, if this is the case, the difficulty stems most probably from opposing inclinations and habits, rather than from a difficulty to grasp the rule.

The claims of this section raise further questions that I cannot pursue here. Recall, however, that my aim is to defend the causal–dispositional approach, not to develop it in detail. The main point of this section was to show that there is an important distinction to be made between accidental rule-conformity and rule-governance. Given this, it is clear that the ability to follow a rule cannot simply be identified with the possession of a disposition to conform to it.

2.4 Rule-following

According to the causal–disposition approach, rule-following involves the manifestation of rule-conforming dispositions, and I have argued that rule-conformity amounts to rule-governance if there is an adequate explanation of why the conformity is non-accidental. But when is an instance of rule-governance an instance of rule-following?

I propose to construe rule-following as a kind of *intentional activity*, and I shall distinguish between two kinds of rule-following on the basis of the following distinction between intentional activity and acting with an intention.[14] Roughly, an agent *S* forms an output (or performs a mental or overt action) *intentionally* if the output (or action) is caused by some of *S*'s mental states that provide a reason explanation of that output (or action). And *S* performs a mental or overt action of type *A* with the *intention to A* if *S*'s *A*-ing is caused by a conscious intention to *A*. In both types of cases, the causation must be *non-deviant*—we will return to this below.[15]

Given this, I propose the following necessary conditions on two kinds of rule-following, which I shall call *explicit* and *implicit* rule-following. In general, *S* follows a rule only if *S*'s thought process or behavior is governed by the rule (as explained in the previous section). *S* follows a rule *explicitly* only if *S* acts with the

[14] Pettit (1990) construes rule-following as intentional rule-conformity. But he does not distinguish between intentional activity and acting with an intention.

[15] Note that I have deployed here a rather broad notion of *activity*, according to which the making of an inference can be intentional activity without being an exercise of mental *agency*. This is only plausible, as many judgments and intentions that are formed on the basis of reasons are not the result of an exercise of mental agency.

intention to conform to the rule. And *S* follows a rule *implicitly* only if *S*'s rule-conforming activity is intentional, and only if *S* has the ability to follow the rule explicitly.

Explicit rule-following entails implicit rule-following, but not vice versa. Consider again John at the traffic lights. John's behavior can be explained in terms of intentional psychology. Presumably, John has the background belief that one should stop when the lights are red and the background policy to conform to this rule. But it is possible, and likely, that John does not stop with the conscious intention to conform to the rule. Nevertheless, John would be able follow the rule explicitly if he intended to.

As already noted, an appeal to intentionality cannot ground rule-governance, because the execution of an intention to follow a rule requires always the application of a further rule. But causation by mental states can give an agent intentional control. The behavior of human and many non-human agents can be governed by rules. If rule-governed behavior is also caused by reason-giving mental states, such as beliefs, desires, and intentions, then it is, in this sense, under the agent's intentional control. Instances of both explicit and implicit rule-following entail this kind of control. But the control is obviously greater, or more robust, in cases of explicit rule-following.

I have proposed necessary conditions on explicit and implicit rule-following. But these conditions are, I think, also plausible candidates for sufficient conditions. What complicates the issue here is a difficult question concerning *plasticity*. To what extent must an agent's rule-conformity be responsive to reflective assessment and belief revision for it to be genuine rule-following? I am inclined to think that the proposed conditions are also sufficient for rule-following, and that responsiveness to reflection and belief revision would constitute only more refined *kinds* of rule-following. But I shall not pursue this any further, as the proposed necessary conditions provide us with a plausible distinction between rule-governance and two types of rule-following.

All this shows that the phenomenon of rule-following is more complex than a simple dispositional view would suggest. Some agents can be governed by rules, others are also able to follow them intentionally. A plausible theory of rule-following must account for this difference, and the simple dispositional account is inadequate because it does not have the resources to make this distinction.

Now we can finally return to the privileging problem and to John's mistake in following the plus-rule. An implicit assumption in all examples of this kind is that the agent is an ordinary subject. We assume that John is a normal person or thinker—on the whole and individual differences aside, he is like you and me. Given this, there is simply no reason to think that John follows a deviant rule explicitly. Nor is there any reason to think that John follows such a rule implicitly. There is, in fact, no good reason to think that his conformity to a deviant rule is not accidental. We granted that John is manifesting a disposition to conform to a deviant rule. But there is no reason to think that John has learnt to conform to it, and there is no reason to think that he is enacting an innate ability. There is, in other words, no explanation that links John's response to a deviant rule, and so there is no reason to think that John's behavior is *governed* by a deviant rule.

Given this, the privileging problem disappears. John partly manifests a disposition to conform to the plus-rule. There is good reason to think that this conformity is not accidental, because it is only plausible to think that he has learnt to conform to that rule. According to claim (5), stated above, John has also a disposition to conform to a deviant rule. We can now see that this is not problematic, because there is no reason to think that this conformity is not accidental. According to claim (6), John *follows* a deviant rule by virtue of manifesting a disposition to give an incorrect answer. We can now see that this is false. Hence, there is no privileging problem.

One might object that this seems to be a mere appeal to common sense, which would be altogether beside the point. For if we could simply appeal to common sense, then we would not have acknowledged the rule-following problem in the first place. But we must be careful here to distinguish between the reply to Kripke's challenge and the response to the privileging problem. Kripke argued, by elimination, that there are no facts that can ground the relevant claims about rule-following. A reply to this challenge from common sense would indeed be hopeless. But the proposed realist response is based on considerations concerning the nature of dispositions, and it offers a metaphysical solution to the metaphysical problem. The privileging problem raises a different kind of challenge. The problem here is to find a *principled reason* to give privilege to the right dispositions. In the first step of my response, it is pointed out that there must be an explanation of why given instances of rule-conformity are non-accidental. In a second step, it is suggested that this explanation can be given in terms of learning (and, perhaps, innate and adaptive abilities). We identified, thereby, a principled reason to give privilege to certain dispositions—namely, the ones that have been acquired in the right way. The final step returns to the example, and it is argued that there is no good reason to think that John's conformity to the deviant rule is non-accidental. This final step is partly based on commonsense considerations. But this is unproblematic for the following two reasons. Firstly, it is plausible to think that common sense can provide reasons for the attribution of dispositions. It seems clear that one can have good reasons to ascribe fragility or solubility, for instance, without having any knowledge about the underlying physical mechanisms. And it is plausible to think that those reasons are provided by common sense (or what is sometimes called "folk physics"). The same, I think, holds for psychological dispositions and for some of the reasons we can have concerning the attribution of learning histories. More importantly, however, the last step of my response is by no means a mere appeal to common sense. Its main thrust stems from the distinction between accidental and non-accidental rule-conformity and from the suggested link to learning history (and innateness). These considerations stem from metaphysical theorizing and from background assumptions on learning theory and nativism, not from common sense.

3 Deviant causal chains

As explained in the introduction, it is plausible to assume that an output is *based on* mental antecedents only if the antecedents cause its formation. The proposed

account of rule-following provides us with a further reason for this appeal to mental causation. The view distinguishes between rule-governance and rule-following. This distinction is made by appeal to the notion of intentional action, and the standard theory of intentional action is a causal theory, which presupposes mental causation by mental states and events.[16] This appeal to mental causation brings with it the difficult problem of deviant causal chains. I will now argue that an appeal to the kind of dispositions that enabled the solution to the rule-following problem allows us also to exclude deviant causal chains.

3.1 Rationalization

The relation between mental antecedents and their outputs in reasoning involves clearly more than causation. Suppose that John comes to believe that he missed an important deadline. Annoyed by this, John kicks a table, which knocks over a bottle, and so red wine is pouring onto the carpet. Observing this, John then forms the belief that red wine is pouring onto the carpet.

This is a straightforward example of a deviant causal chain. The new belief is formed because of the mental antecedent in the sense of being caused by it (provided that the relation of causation is transitive). But the new belief is clearly not based on the antecedent. Examples of this kind violate the already stated condition that the antecedent mental states must rationalize the output. Clearly, John's belief that he missed the deadline does not rationalize the belief that wine is pouring onto the carpet.

But what is rationalization? I have in mind a broad notion, according to which rationalizing something is basically the same as rendering it *intelligible*. But this notion is itself in need of elucidation. A good starting point, and one source of rationalization, is provided by the standards of rationality, such as the rules of logic, the norms of decision theory, and so on. The formation of a belief in accordance with *modus ponens* is a paradigm example. In such cases, the mental antecedents rationalize the output simply because it is *rational* to make the inference in response to the antecedents. This involves usually some kind of *match* between the contents of the antecedents and the output. In the case of following *modus ponens* the match is obvious, because the content of the conclusion (that q) is contained in the content of one of the antecedents (if p, then q). In standard cases of actions that are based on means-ends reasoning there is a match between the *type* of action and the antecedent belief that specifies that type of action as the appropriate means to the intended or desired end.[17]

But we need an account of rationalization that goes beyond rendering rational, because inferences and actions can be based on fallacious or irrational reasoning. In reasoning in which the mental antecedents *merely rationalize*, the antecedents render the output intelligible, but not rational. We can distinguish between two possible sources of mere rationalization.

[16] Davidson (1963). For more recent and more detailed accounts see, for instance, Mele (2003) and Enç (2003).

[17] This is, basically, Davidson's (1963) narrow conception of rationalization: means-ends rationalization by desires and beliefs (compare Mele 2003, pp. 70–71).

An inference or action may appear as intelligible in the light of one's own judgments, experiences and intuitions. The theoretical framework for this is provided by simulation theory. Let us consider an example of irrational reasoning. Suppose that Sue is tossing a coin a few times in a row. As it happens, it lands tails up each time. John observes this and says: "How strange. Try again, I bet it will be heads up next time." Suppose that Sue knows that this judgment is false. Despite this, she may find John's remark intelligible on the basis of her own intuitions. Perhaps Sue finds herself inclined to agree with John, even though she knows better. Or perhaps she finds John's view intelligible in the light of her own past intuitions. In the past, she thought the same, but now she knows better. In general, we find each other intelligible if we find that we would have judged or acted in the same way, and we can find irrational responses intelligible by bracketing our own judgments and by taking up that agent's perspective.

Alternatively, psychological knowledge and theorizing may allow us to make sense of irrational inferences and actions. For instance, theorizing may help to explain irrational responses in terms of the misapplication of valid rules or in terms of invalid principles that resemble valid ones. In this way, inferences or actions can make sense, even if we judge them irrational.

The points just made highlight the fact that rationalizing relations are *general* relations between types of mental states, mental contents and actions, and that there are *patterns* of rational and irrational reasoning can be categorized and described by means of law-like generalizations.[18] This lends support to the assumption that rationalization involves the manifestation of dispositions: whenever mental antecedents rationalize an output, the agent is manifesting a disposition to form a certain type of output in response to certain inputs.

Further, it is plausible to assume that the dispositions that underlie the formation of outputs on the basis of mental antecedents must be an integral part of an assembly of abilities that constitute the agent's ability to engage in reasoning. In general, and very roughly, the ability to reason presupposes the ability to access the contents of the mental antecedents and the ability to register that the outputs are formed in response to the antecedents. Further, I shall assume that an agent who forms an output on the basis of reasoning is disposed to use the output as a premise in further reasoning, and that the agent is disposed to refer to the contents of the antecedents in order to explain and justify the output (in case an explanation or justification is called for). An agent may fail, of course, to exercise these abilities or to manifest the mentioned dispositions, and an agent may lose them for various kinds of reasons. I submit, however, that if these abilities and dispositions are not present or acquired when the output is formed, then the output is not properly based on reasoning. (An agent who altogether lacks these abilities and dispositions is at best capable of, say, computation, but not genuine reasoning.)

More could be said here, of course. But this rough characterization of reasoning and rationalization suffices to substantiate the claim that, in the example, John's

[18] There is a vast amount of literature in psychology which shows that ordinary subjects frequently violate the standards of logic and decision theory. But the same research also suggests that mistakes are systematic and that there are general patterns of irrational thought and behavior (compare Evans and Over 1996, for instance).

antecedent belief about the missed deadline does not rationalize the new belief, and it provides the basis for solutions to more difficult cases of causal deviance, as we will see below.

3.2 Causal deviance in action

The most difficult cases of causal deviance are the ones where the mental antecedents cause *and* rationalize the outcome. In the theory of action, we can distinguish here between *consequential* and *basic* deviance. In cases of consequential deviance, the deviant causal chain connects an agent's basic action to some non-basic outcome or consequence (that is brought about by the performance of the basic action). In one standard example, a sniper fires a shot with the intention to kill an enemy. He misses, but the noise of the shot stampedes a herd of wild pigs, which trample his target to death (see Bishop 1989, for instance). In cases of basic deviance, the agent fails even to perform a basic action, as the deviant causal chain runs between the mental antecedents and an event that *would have been* a basic action (had the chain not been deviant). A standard example is Davidson's climber who intends to rid himself of the weight and danger of holding another man on a rope by loosening his grip. This intention unnerves him so much that it causes him to loosen his grip on the rope (Davidson 1973).

In both cases, the outcome is caused and rationalized by the mental antecedents, but it is not properly based on them. In the first case, the outcome is not brought about in the right way. In the second, no action is performed at all. The problem of consequential deviance has a standard solution. What goes wrong is that the outcome is not produced in accordance with the agent's action-plan. In our example, the sniper intends to kill the enemy by shooting him, not by having him trampled to death. This action-plan, which specifies how the outcome is to be brought about, is part of the intention's content. Cases of consequential deviance can therefore be excluded by the condition that the intended outcome must be produced in accordance with the agent's action-plan (Bishop 1989, for instance).

Elsewhere I argued that the problem of basic deviance can be solved under the assumption that mental states and events are causally efficacious and explanatory in virtue of their contents. The climber's intention causes and rationalizes the movement of loosening the grip, but it does not cause and explain this movement in virtue of its content. The state of nervousness, which connects the intention to the movement, preserves the relation of causation, but it does not preserve the relation of being causally relevant and explanatory in virtue of content. Cases of basic deviance can therefore be excluded by the condition that the rationalizing mental antecedents must be causally efficacious and explanatory in virtue of their intentional contents (Schlosser 2007).

3.3 Causal deviance in reasoning

Suppose that John suspects his wife, Joanne, of having an affair. John, in fact, has a number of beliefs about various bits of evidence that give him reason to infer that Joanne is having an affair. But John fails to make this inference. He fails to consider

the reasons together and in the right order, and so he fails to see the crucial connections between them. Suppose, further, that Joanne is actually having an affair, and that she becomes aware of the fact that John knows about the relevant bits of evidence. Unnerved by this, Joanne delivers a full confession, and so John forms the belief that his wife is having an affair. This belief is rationalized and caused by the antecedent beliefs, but it seems clear that it is not properly based on them. As in cases of consequential deviance, this involves a causal chain that is in part *external* to the agent.

Consider now the following variation. Again, John's antecedent beliefs would justify the conclusion, but John fails to make the inference. Suppose now that some of the antecedent beliefs remind John of a similar past incident, where a former partner did have an affair. Upset by this memory and by the thought that he finds himself in same situation again, John jumps to the conclusion that Joanne is also having an affair. Again, the formation of the belief is rationalized and caused by antecedent beliefs. But the inference is not properly based on the antecedent states. As in cases of basic deviance, the deviant causal chain is *internal*, involving only John's mental states and events.

Despite the similarities, cases of external and internal deviance in reasoning cannot be solved in analogy with the problems in the theory of action. The antecedents in reasoning do not involve action-plans, and the causal chains appear to preserve the relation of causation in virtue of content. Intuitively, the problem in both cases is that the formation of the conclusion is a *response* to the antecedents *only insofar* as it is connected to them by way of a causal chain. The problem, in other words, seems to be that the inference is not a proper and direct response to the rationalizing antecedents. The causal–dispositional approach can capture this intuition, because it requires that the outputs in reasoning must result from manifestations of the right kind of dispositions. On the basis of the characterization of rationalization presented above, I propose the following conditions on reasoning that make this idea more precise.

Firstly, the dispositions manifested in reasoning must be the *agent's* dispositions: a transition from mental antecedents to an output is an instance of reasoning only if it consists solely in manifestations of the agent's dispositional properties. This condition is intuitively very plausible, and it should be noted that it does not raise any obvious problems with respect to the use of external computing tools. In such cases, the conclusion is only *partly* based on the agent's own reasoning. The outputs of the agent's reasoning provide inputs for external computation, and the outputs of the computation deliver new inputs for the agent (inputs for belief formation or for further reasoning). The same holds for cases where an agent's conclusion is partly based on the reasoning of another rational agent.

Secondly, the dispositions manifested in reasoning must be of the right kind. As suggest above (Sect. 3.1), they must conform to some rationalizing pattern of inference. We could say that they must be *rule-conforming* dispositions, if we assumed that all rationalizing patterns instantiate rules. But this would be to assume that there are invalid and irrational rules (as there are many invalid and irrational rationalizing patterns). I shall say, rather, that the dispositions manifested in reasoning must be *rationalizing* dispositions. In general, a transition from mental

antecedents to an output is an instance of reasoning only if the output is connected to each of the antecedents by way of a chain that consists solely in the manifestations of rationalizing dispositions.[19] In accordance with the proposed account of rule-governance (Sect. 2.3), the conformity to rationalizing patterns must be non-accidental, and rationalizing dispositions must be accompanied by the ability to access the antecedents and by a disposition to refer to them in justification (as suggested in Sect. 3.1). For instance, the manifestation of a disposition to conform to *modus ponens* is an instance of reasoning only if the conformity is non-accidental, and only if the agent is able to access the antecedents and disposed to refer to them in justification.

Third, the dispositions manifested in reasoning must be *basic* in the following sense. Roughly, a disposition is a basic disposition at a given level of explanation if its manifestation in response to a given stimulus does not involve the manifestation of any other disposition at that level of explanation. Suppose that John is disposed to fall asleep on long train journeys, and that he is disposed to snore when asleep. John, it seems, is thereby disposed to snore on long train journeys. This is not a basic disposition, as the response to the stimulus involves the manifestation of two dispositions. In contrast, the disposition to reason in accord with *modus ponens* is basic at the level of intentional psychology. Its stimulus is to consider two beliefs with certain contents, its characteristic manifestation is to form another belief, and the formation of the new belief does not involve the manifestation of any other disposition at the level of intentional psychology. This disposition is basic and the formation of the output is therefore a *direct response* to the rationalizing antecedents.[20]

These conditions provide plausible constraints on any causal–dispositional account of reasoning and agency.[21] With them in place, we can now return to the two remaining problems of causal deviance in reasoning. The first condition provides a solution to the first problem, where the chain is in part external. This is straightforward. The mental antecedents rationalize John's conclusion (that his wife is having an affair), but the causal chain that connects them does not consist solely in manifestations of the agent's dispositional properties (because it involves the manifestation of some of Joanne's dispositions).

The second problem of internal deviance violates the second and the third condition. John's conclusion is not connected to the antecedent mental states via a chain that consists only in the manifestation of basic rationalizing dispositions. The most problematic step is the first one, in which John remembers a past episode. Let us assume that John's remembering this episode results in the formation of beliefs

[19] The performance of actions in response to intentions is based on rationalizing dispositions due to the rationalizing role of intentions (compare Enç 2003, pp. 183–187).

[20] In the theory of action, some philosophers secure direct responsiveness by requiring that the mental antecedents must be the *proximate* causes of actions (Mele 2003, for instance). But this appears to be *ad hoc* as a condition on causation (given that the relation of causation is transitive). The same cannot be said about dispositions, because it is part of their nature to be manifested *in response to* certain stimuli.

[21] Wedgwood (2006) has suggested similar conditions. But his account differs radically from the one proposed here, because he suggests that the dispositions that govern reasoning must be responsive to *normative facts*.

that are genuinely based on memory. We can assume, plausibly, that these beliefs are formed in one of the following two ways. Either they are formed in response to a memory image, or they are directly retrieved from memory. If a belief is based on a memory image, then it is formed in response to this image. This violates the second and the third condition. The output is not a direct response to the rationalizing antecedents, because it is not based solely on the manifestation of basic rationalizing dispositions. If, on the other hand, a belief is retrieved from memory, then it is not based on the right kind of disposition. Firstly, memory beliefs differ from beliefs that are formed in response to rationalizing antecedents due to their functional role. A genuine memory belief must be retrieved from memory and it must, presumably, be causally connected to some of the agent's past mental states. In forming such a belief, the agent retrieves a stored memory and thereby manifests a type of disposition that differs from rationalizing dispositions (roughly, a disposition to retrieve beliefs from memory in response to certain mental stimuli). Secondly, the possession of a memory belief does not entail the possession of a disposition to justify that belief by reference to the contents of the antecedents (and the formation of a memory belief does usually not result in the formation of such a disposition). The agent may have or form a disposition to explain the formation of a memory belief with reference to the contents of mental antecedents. But this is not the same as a disposition to *justify* it in the light of the antecedents' contents.

It may well be that the example violates the proposed conditions in further respects. But I will not pursue this any further. I shall close, instead, with the following two remarks. Firstly, I think that similar responses can be given for cases where the formation of an output is mediated by mental states that are the result of some kind of free association, wishful thinking, emotional attachment, or the like. They all involve the manifestation of dispositions that differ in important respects from rationalizing dispositions. Secondly, the proposed conditions on reasoning provide also a solution to the problem of basic deviance. In cases of basic deviance, such as Davidson's climber, the movement occurs in response to an intermediate state, typically a state of agitation, that links the movement to the rationalizing antecedents. So, the movement is not a direct response to the antecedents, and it is not based on the manifestation of a basic rationalizing disposition. This violates two of the proposed conditions. Given some further plausible assumptions, this solution entails the solution mentioned above, which requires that mental states are causally efficacious in virtue of their contents. The proposed causal–dispositional account says that reason-based outputs and actions must be based on the right kind of basic dispositions. Arguably, the manifestation of these dispositions must be sensitive to both the mental state *types* and *contents* of the antecedents, which is just to say that the antecedents must cause the output in virtue of being of a certain mental state type and in virtue of having a certain type of content.

4 Conclusion

Rational agents can be governed by norms. To say that an agent is governed by norms appears to be very different from saying that something is in accordance with

the laws of nature (or the laws of science). And so one might wonder how reasoning and rule-following can possibly be accounted for in causal, dispositional, or broadly naturalistic terms. The proposed causal–dispositional account shows how this can be done. An agent's thoughts and actions are governed by norms if they result from manifestations of dispositions to conform, and if this conformity has an adequate explanation. Manifestations of rule-conforming dispositions obey the laws of nature, and they can be described in terms of more specific psychological generalizations. Given this, and given the arguments presented here, the normative aspects of reasoning, including rule-following and the possibility of error, are perfectly compatible with a naturalistic account of mentality and agency.

The view that I have defended is decidedly causal *and* dispositional. The appeal to intentional action and mental causation plays an important role in the proposed account of rule-following, and the appeal to rule-conforming and rationalizing dispositions provides a solution to the problem of deviant causal chains. There is, then, good reason to endorse a causal–dispositional approach, as the combined appeal to causation and dispositionality allows us to solve two of the most notorious problems in the philosophy of mind and action.

Acknowledgements This paper was written during a post-doctoral research project which was funded by the FWF (Austrian Science Fund) and hosted by the University of Bristol. I would like to thank Alexander Bird, Jennifer Hornsby, Jimmy Doyle, and an anonymous referee for very helpful comments on an earlier draft.

References

Baker, J. (2008). Rationality without reasons. *Mind, 117*, 763–782.

Bechtel, W., & Abrahamsen, A. (1990). *Connectionism and the mind: An introduction to parallel processing in networks*. Cambridge, MA: Blackwell.

Bishop, J. (1989). *Natural agency: An essay on the causal theory of action*. Cambridge: Cambridge University Press.

Boghossian, P. (1989). The rule-following considerations. *Mind, 98*, 507–549.

Boghossian, P. (2005). Rules, meaning and intention—discussion. *Philosophical Studies, 124*, 185–197.

Chomsky, N. (1959). Review of B. A. Skinner's *Verbal Behavior*. *Language, 35*, 26–58.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Davidson, D. (1963). Actions, reasons, and causes. Reprinted in his *Essays on actions and events* (pp. 3–20). Oxford: Clarendon Press.

Davidson, D. (1973). Freedom to act. Reprinted in his *Essays on actions and events* (pp. 63–82). Oxford: Clarendon Press.

Dretske, F. (1988). *Explaining behavior: Reasons in a world of causes*. Cambridge, MA: MIT Press.

Elman, J. L., et al. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.

Enç, B. (2003). *How we act: Causes, reasons, and intentions*. Oxford: Oxford University Press.

Evans, J., & Over, D. (1996). *Rationality and reasoning*. Hove: Psychology Press.

Gibbard, A. (1990). *Wise choices, apt feelings: A theory of normative judgment*. Oxford: Clarendon Press.

Gibbons, T. (2006). Mental causation without downward causation. *Philosophical Review, 115*, 79–103.

Goldman, A. (1979). What is justified belief? In G. Pappas (Ed.), *Justification and knowledge*. Dortrecht: D. Reidel.
Handfield, T., & Bird, A. (2008). Dispositions, rules, and finks. *Philosophical Studies, 140*, 285–298.
Harman, G. (1973). *Thought*. Princeton: Princeton University Press.
Johnston, M. (1992). How to speak of the colors. *Philosophical Studies, 68*, 221–263.
Kripke, S. (1982). *Wittgenstein on rules and private language*. Cambridge: Harvard University Press.
Lewis, D. (1997). Finkish dispositions. *Philosophical Quarterly, 47*, 143–158.
Martin, C. B. (1994). Dispositions and conditionals. *Philosophical Quarterly, 44*, 1–8.
Martin, C. B., & Heil, J. (1998). Rules and powers. *Philosophical Perspectives, 12*, 283–312.
McDowell, J. (1984). Wittgenstein on following a rule. *Synthese, 58*, 325–364.
Mele, A. R. (2003). *Motivation and agency*. Oxford: Oxford University Press.
Millikan, R. (1990). Truth rules, hoverflies and the Kripke-Wittgenstein paradox. *Philosophical Review, 99*, 323–353.
Owens, D. (1989). Levels of explanation. *Mind, 98*, 59–79.
Pettit, P. (1990). The reality of rule-following. *Mind, 99*, 1–21.
Schlosser, M. (2007). Basic deviance reconsidered. *Analysis, 67*, 186–194.
Swain, M. (1981). *Reasons and knowledge*. London: Cornell University Press.
Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.
Wedgwood, R. (2006). The normative force of reasoning. *Noûs, 40*, 660–686.