Causally efficacious intentions and the sense of agency: In defense of real mental causation

Markus E. Schlosser, m.schlosser@hum.leidenuniv.nl

Forthcoming in the *Journal of Theoretical and Philosophical Psychology*

This is the author's copy, which might differ in minor details from the final print version

Abstract: Empirical evidence, it has often been argued, undermines our commonsense assumptions concerning the efficacy of conscious intentions. One of the most influential advocates of this challenge has been Daniel Wegner, who has presented an impressive amount of evidence in support of a model of "apparent mental causation". According to Wegner, this model provides the best explanation of numerous curious and pathological cases of behavior. Further, it seems that Benjamin Libet's classic experiment on the initiation of action and the empirical evidence concerning the confabulation of reason explanations provide further support for this view. In response, I will propose an alternative model of "real mental causation" that can accommodate the empirical evidence just as well as Wegner's. Further, we will see that there is plenty of evidence in support of the assumption that intentions are causally efficacious. This will provide us with ample reason to endorse the model of real mental causation.

Keywords: intentional action, mental causation, illusion of conscious will, sense of agency

1. Introduction: The conscious self as a spectator

We tend to believe that our conscious intentions and goals make a real difference to how we act. This assumption of mental causation lies at the very heart of the commonsense conception of human agency (Horgan & Woodward 1985, D'Andrade 1987, Greenwood 1991, Malle 1999 and 2004, for instance). It plays a central role in philosophical theories of human action (Davidson 1963, Goldman 1970, Bratman 1987, and Enç 2003, for instance), and it has also been at the core of many psychological theories of intentional action and motivation after Freud and after the fall of behaviorism (Fishbein & Ajzen 1975, Triandis 1977, Ajzen 1985, Locke & Latham 1990, Heckhausen 1991, Gollwitzer 1993, and Austin & Vancouver 1996, for instance).

More recently, however, the belief in mental causation and conscious control has come under heavy attack. Psychologists, social scientists, and cognitive neuroscientists have produced an impressive amount of evidence that challenges most parts of the commonsense conception of human agency, including the claim that conscious intentions are efficacious in the causation of behavior. One of the most influential and persistent advocates of this challenge has been Daniel M. Wegner, who has produced and collected an impressive amount of empirical evidence in favor of a model of "apparent mental causation" (Wegner & Wheatley 1999, Wegner 2002, 2004,

2005, and 2008, for instance). Wegner's view has been criticized for various conceptual ambiguities and argumentative flaws (Nahmias 2002, many of the peer commentaries to Wegner 2004, Bayne 2006, Malle 2006, Dennett 2008, and Mele 2009, for instance). Despite this, the view has remained very influential, and it is still widely acknowledged and discussed (Hassin et al. 2005, Pockett et al. 2006, Ross et al. 2007, and Baer et al. 2008, for instance). The main reason for this consists, I think, in the sheer amount of evidence that Wegner has presented. Even if there are ambiguities and argumentative shortcomings, it seems nevertheless clear that the bulk of the evidence supports a model of *apparent* mental causation. For this reason, many commentators accept the general picture of human agency that Wegner advocates, even though they disagree with some parts or aspects of Wegner's view. Roughly, this general picture is that the conscious self is a mere spectator in the performance of intentional behavior. Or, if one wants to avoid the reference to "the self", the view is that conscious choices and intentions are epiphenomena: they precede and accompany actions, but they do not cause them (Dennett 2003, peer commentaries by Ito, Kirsch & Lynn, Pylyshyn, Tweney & Wachholtz, Velmans, and Young to Wegner 2004, Frith 2007, Davies 2009, and Damasio 2010, for instance; influential predecessors are Nisbett & Wilson 1977 and Libet 1985; see also Wilson 2002).

In a recent article, Baumeister & Masicampo have argued that any theory which posits the efficacy of conscious intentions must either "incorporate or refute" Wegner's view (2010: 946). They have proposed a theory that incorporates the view. My aim is to refute Wegner's arguments by way of providing a constructive response. As mentioned, the criticisms of the view have failed to be decisive, mainly due to the large amount of evidence that seems to support it. It is unlikely that further criticisms of certain parts or aspects of Wegner's view would change this. What is needed is a response which addresses the overall argument, and which shows that the bulk of the evidence does not support a model of apparent mental causation. I will propose an alternative model of "real mental causation" that preserves the core of the commonsense conception of human agency, and I will argue that this model can accommodate the evidence just as well as Wegner's. The opponents of the commonsense conception have often suggested, implicitly at least, that the assumption of mental causation lacks empirical support. We will see, however, that this is not the case. There is plenty of evidence in support of the view that reasons and conscious intentions are causally efficacious in the initiation and guidance of behavior. In conjunction with the response to Wegner's challenge, this will provide us with ample reason to endorse the model of real mental causation.

2. Apparent mental causation and the illusion of conscious will

The most detailed defense of Wegner's view can be found in his book *The illusion of conscious will* (2002). In the first chapter, he explains the title and main thesis of the book as follows:

> (ICW) (*Illusion of conscious will*): Conscious will is "an illusion in the sense that *the experience of consciously willing an action is not a direct indication that the conscious thought has caused the action*" (ibid.: 2).

The theoretical core of the book is the mentioned model of apparent mental causation. What complicates the issue is that Wegner sometimes uses the term "apparent mental causation" in order to refer to a theory of behavior causation, while on other occasions he says that the model provides an account of how the experience of conscious will is generated. It is important to distinguish clearly between those two models, and I will therefore use the term "apparent mental causation" ("AMC", for short) only in order to refer to Wegner's account of behavior causation:

> (AMC) (*Apparent mental causation*): Conscious intentions and subsequent actions are generated by two distinct sub-personal mechanisms. There may or may not be any causal interaction between those two sub-personal mechanisms, but the connection between conscious intentions and actions is only *apparently* causal. See figure 1. (ibid.: 67-68)

> (SCW) (*Sources of the experience of conscious will*): The experience of consciously willing an action arises when we *interpret* a conscious intention as the cause of the subsequent action. In particular, we interpret our actions as consciously willed when (a) the conscious intention occurs before the action (*priority*), (b) the action is consistent with the intention (*consistency*), and when (c) the action is not accompanied by other potential causes (*exclusivity*). (ibid.: 69)

SCW offers a model of how the experience of conscious will is generated. This model is independent from ICW and AMC—SCW may be true, even if ICW and AMC are false, and *vice versa*. AMC proposes a model of behavior causation.

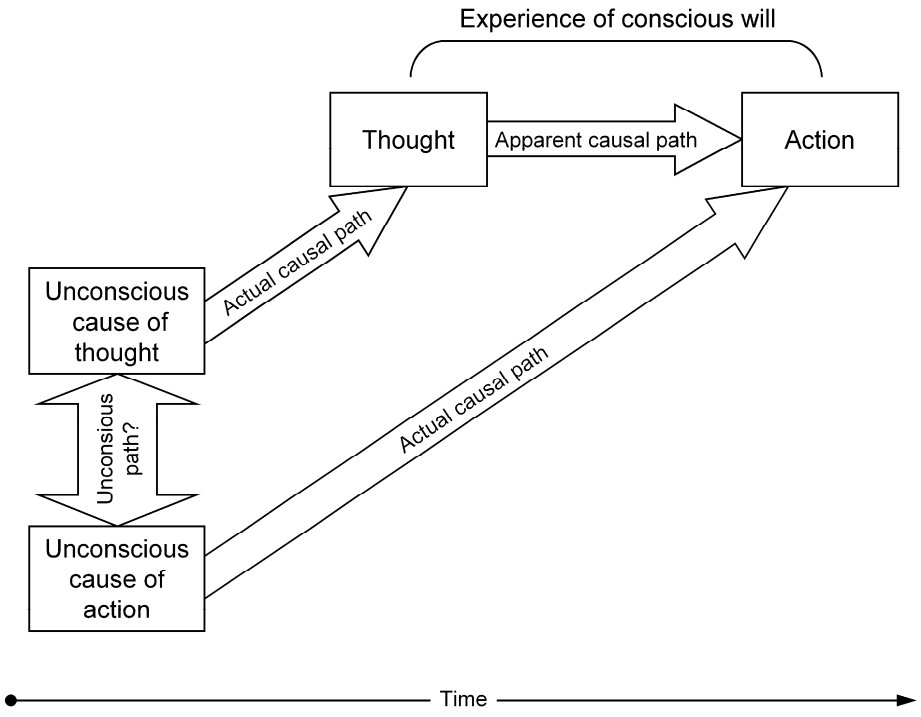

Figure 1. Model of apparent mental causation (AMC)
From Wegner 2002: 68. See also Wegner & Wheatley 1999: 483.

My main concern here will be with the model of AMC, primarily because it entails the claim that conscious intentions are not causally efficacious in the initiation and guidance of actions. I will, in particular, defend the commonsense assumption that conscious intentions are causally efficacious against Wegner's argument for the model of AMC. Before we turn to this argument, I shall say more about the commonsense conception of human agency, and I shall explain why a defense of this view does *not* require an engagement with Wegner's thesis that conscious will is an illusion (ICW). Throughout, I shall assume, by way of a terminological stipulation, that intentions are accessible to consciousness (Pacherie & Haggard 2010), and that choices and decisions just are formations of intentions. (More clarifications will be provided in due course.)

It is often pointed out that many of our actions are habitual or automatic in the sense that they are not preceded and accompanied by conscious intentions. A *considered* commonsense view of agency should acknowledge this, simply because phenomenological reflection alone can reveal that many of our actions are habitual or automatic. Moreover, this is unproblematic, from a commonsense perspective, because the vast majority of automatic actions are in the service of

conscious intentions and goals. Consider, for instance, all the actions that one executes while driving a car. The fact that most of them are automatic is unproblematic, provided that these are overlearned sub-routines that serve the pursuit of conscious goals. Further, it seems that we could exercise conscious control over such sub-routines, if we intended to. This is compatible with the empirical evidence on the automatic initiation and control of action, which shows only that many actions *can* be executed automatically, in the sense that they do *not require* conscious control (for reviews see Bargh 1994, Bargh & Chartrand 1999, and Hassin et al. 2005).[1]

Nevertheless, many of our actions are preceded by conscious intentions, and we tend to think that our conscious intentions make a real difference to our behavior. This assumption is at the very core of the commonsense conception of human agency. Some philosophers have argued that a commitment to the explanatory relevance of intentions does not entail a commitment to the claim that intentions *cause* actions (Anscombe 1957, Melden 1961, and Sehon 2005, for instance). But according to the widely accepted standard view, the commonsense conception is committed to the assumption of mental causation: conscious intentions and other mental states are genuinely explanatory of behavior only if they cause behavior.

However, no one should assume that conscious intentions *guarantee* the performance of matching actions. All kinds of accidents, interventions, or breakdowns may prevent the execution of an intended action, or the agent might simply revise his or her decision. This holds for basic and non-basic actions. Roughly, *non-basic* actions are things that we do *by doing something else*, and *basic* actions are things that we do without doing something else (see Enç 2003, for instance). Usually, non-basic actions are naturally described as goals or act consequences. Suppose, for instance, that you are moving the cursor on your computer screen by moving the mouse. You are moving the mouse by moving your hand, but you are not moving your hand by *doing* something else (although many other things, such as neuron firings and muscle contractions, *occur* when you move your hand). Moving the hand is the basic action. Moving the mouse and moving the cursor are non-basic actions (goals and consequences). Under normal conditions, control over basic actions is more reliable than control over non-basic actions. The former usually requires the successful execution of motor skills, whereas the latter requires also the satisfaction of various external conditions. In our example, the computer must be working, the mouse must be connected to the computer, and so on. We might say here that the external world must

[1] According to a narrow definition, an action is *automatic* only if it is unconscious, effortless, and uncontrollable. But it has become clear that few, if any, actions satisfy those conditions (Bargh 1994). It is therefore permissible, and more interesting, to assume a less demanding notion of automaticity that does not require uncontrollability.

"cooperate" for a successful performance of non-basic actions. But events in the external world may also result in lucky coincidences. Suppose, for instance, that the mouse is not connected to the computer. The cursor may nevertheless move in the intended way. In this case, the intention to move the cursor is followed by the intended consequence, but the intention clearly plays no causal role. This problem can be avoided if we restrict the claim concerning the causal efficacy of intentions to *basic* actions. But it seems that a similar problem can arise even for the execution of basic actions. Brain stimulation experiments have shown, for instance, that it is possible to bring about coordinated and controlled movements without bringing about any corresponding conscious state (Penfield 1975 and Desmurget et al. 2009). Given this, one can imagine circumstances in which it is a mere coincidence that an intention to perform a basic action, such as the intention to raise an arm, is followed by a matching movement. But such circumstances are highly unusual, and we can exclude them by means of a *ceteris paribus* clause. Furthermore, we can exclude cases in which the agent revises the intention by restricting the claim concerning the causal efficacy of intentions to cases in which the intention is in fact followed by the relevant action. Given all this, the thesis that I shall defend can be stated as follows:

> (RMC) (*Real mental causation*): *Ceteris paribus*, conscious intentions are causally efficacious in the initiation and guidance of the relevant basic actions.

The relevant basic actions are actions that *match* with the intention's content.[2] If one intends to perform a basic action, this would be the match between an intention to perform an action of type A and the performance of that type of action (an "A-ing", for short). And if one intends to perform a non-basic action, this could be the match between A-ing and an intention to A in order to bring about B, for instance. I will defend the thesis of RMC by way of defending a model of real mental causation ("model of RMC", for short) against Wegner's arguments for the model of AMC. This requires some further elaboration, because in some passages Wegner seems to affirm the reality of mental causation. For instance, he says that "it must be the case that *something* in our minds plays a causal role in making our actions occur" (2002: 96). According to the model of AMC, this is a "set of unconscious *mental* processes that cause the action" (ibid., my emphasis). So where, exactly, lies the disagreement?

Common experience suggests that many of our actions are preceded by conscious intentions (and Wegner agrees; ibid.: 97). It follows, according to RMC, that many of our actions

[2] This matching is basically what Wegner calls "consistency" (2002: 78-81).

are caused by conscious intentions. It follows, in particular, that some of the real causes of many of our actions are *accessible* to consciousness. The model of AMC denies this. It says that the real causes of our actions are inaccessible, despite the fact that they are "mental" causes. Given this, we can capture the disagreement in terms of the distinction between *personal* and *sub-personal* levels of explanation (see Elton 2000, for instance). At the personal level, actions are described and explained in terms of accessible mental states or events (desires, beliefs, intentions, judgments, and so on). At sub-personal levels, we find scientific explanations in terms of the underlying and inaccessible computational, neural, or physical mechanisms. This distinction allows for mental states that are inaccessible to consciousness, provided that information processing states can be called "mental" states. So, according to RMC, many actions are caused by personal level states and events (conscious intentions, in particular). The model of AMC denies this. It says that the "real causes of human action" are "never present in consciousness" (2002: 96 and Wegner & Wheatley 1999: 490).[3]

Does a defense of the commonsense view require a response to Wegner's claim that conscious will is an illusion (ICW)? My aim is to defend RMC against Wegner's argument for the model of AMC. ICW does not entail AMC, and the rejection of ICW is not in any other obvious way required for a defense of RMC. Nevertheless, Wegner clearly thinks that the argument for ICW constitutes a serious challenge to the commonsense view. In order to evaluate this, it will be helpful to make explicit how ICW construes the content of the experience of consciously willing an action:

> (CCW) (*Content of conscious will*): The content of the experience of consciously willing an action includes the representation of a causal connection between a conscious intention and a matching action.

To my knowledge, there is no account of the commonsense view according to which something along the lines of CCW is part of the commonsense conception of human agency (see Horgan & Woodward 1985, D'Andrade 1987, Greenwood 1991, Malle 1999 and 2004). Given this, the

[3] In response to critics, Wegner says that he never denied the possibility of causation by conscious mental states. He is surprised by the misunderstanding and he wonders "what book these folks were reading" (2004: 683-84). But admitting the *possibility* of causation by conscious states is not the same as affirming its frequent reality. Moreover, the book is quite clear on this: it says that the real causes of human action are never present in consciousness. More recent writings seem to confirm this: "When we look at ourselves, we perceive a simple and often astonishing apparent causal sequence […] when the real causal sequence underlying our behavior is complex, multithreaded, and unknown to us as it happens" (2008: 227-28). In any case, I shall engage here with the *perceived view*, which is that the model of AMC denies the efficacy of conscious intentions in the causation of behavior.

answer to our question seems straightforward: a defense of the commonsense view does not require a response to ICW, because ICW rejects a view concerning the content of conscious willing (CCW) that is not commonly attributed to the commonsense conception of human agency. But there are further and independent reasons to reject CCW.

In recent years, the phenomenology of acting has been the subject of much investigation and theorizing—partly due to the influence of Wegner's work. The phenomenon that Wegner called the "experience of conscious will" is now commonly referred to as "the sense of agency". It is widely agreed that the sense of agency is a complex and graded phenomenon, and it is now common to distinguish between a basic *sense* of agency and post-act *judgments* about one's agency (Marcel 2003, Bayne & Pacherie 2007, Gallagher 2007, and Synofzik et al. 2008, for instance). The basic sense is an online experience that accompanies the performance of actions. It does not require the presence of a conscious intention and it can be a minimal sense of agency that is phenomenologically rather thin. In contrast, judgments of agency are usually offline and post-act, and they are usually assumed to be subject to various biases. Many researchers also maintain that even the basic sense of agency involves a sense of causal efficacy: the sense that, by acting, *I am bringing something about* (Aarts et at. 2005, Pacherie 2007, Gallagher 2007, and Synofzik et al. 2008, for instance). What does this mean? Does this entail something like CCW?

First of all, note that there is a clear difference between the claim that *I* am efficacious and the claim that *my intentions* are efficacious. This distinction generates difficult questions. Does one's sense of being causally efficacious consist in the sense that one's intentions are causally efficacious? This is rather implausible. I believe that my intentions are causally efficacious. But, for all I can tell, I usually do not have a conscious belief or awareness with that content *when I am acting* (see Wakefield & Dreyfus 1991, Horgan et al. 2003, and Bayne 2006). Instead, when I am acting, or when I am about to act, my conscious awareness is usually focused on the parts of the world that I have to change in order to attain my goals (Gallagher 2006).

Another possibility is that the sense that I am efficacious consists in the sense that *I*, an irreducible mental substance, cause my actions. This view of substance, self, or agent causation is not only philosophically problematic and empirically inadequate. But the suggestion is also phenomenologically implausible. One may have the belief that the self or agent is an irreducible substance that causes its actions. But it is rather implausible to suggest that this controversial metaphysical doctrine is represented in the content of one's sense of agency.

A much more plausible interpretation is that the sense of being causally efficacious consists in the sense that *my actions* are causally efficacious in bringing about my goals. In order to see what this entails, it is important to take the distinction between basic and non-basic actions into account. Return to our example. You are moving the cursor on your computer screen by moving the mouse, and you are moving the mouse by moving your hand in a certain way. Arguably, even such a simple action is accompanied by a sense of being causally efficacious. But this consists, most plausibly, in the sense of being efficacious by bringing about the intended consequences—the sense that one's basic action is efficacious in the pursuit of one's goals (Engbert et al. 2008 provide empirical evidence for this view). This interpretation does not entail the implausible claim that the sense of agency includes a representation of one's intentions as being causally efficacious, and it does not entail a contentious notion of substance, self, or agent causation. And it applies to the vast majority of everyday actions, because most of our basic actions are in the service of non-basic actions. Usually, we perform basic actions in order to bring about something else. In fact, it is hard to think of any ordinary action that is purely basic, in the sense that we do not intend to bring about something else by performing it. Given all this, we can explain the sense of causal efficacy for ordinary actions in terms of the sense that one's actions are causally efficacious in the pursuit of one's goals.

It is a further question whether or not the performance of basic actions is itself accompanied by a sense of agency that represents a causal relation. This would have to be either a causal relation between the self and a basic action or between an intention and a basic action. As explained, both options are problematic. But they are also phenomenologically implausible. When I raise my arm, for instance, without any further goal, then it does not seem to me that my intention is causing the movement. It does not even seem to me that I am causing the movement. Rather, it seems to me that raising my arm is something that I *do*, not something that I *cause*.[4] Again, I believe that my intention causes the action. But this is a theoretical belief about the workings of my agency. It is not something that I am aware of when I am performing the action.

To summarize, CCW is not commonly thought to be a part of the commonsense view of human agency; it fails to take into account the important distinction between basic and non-basic actions; and it is phenomenologically implausible. We have, I think, sufficient reason to reject

[4] According to non-causal theories in philosophy, reasons and intentions are not causally efficacious states or events (Anscombe 1957, Melden 1961, and Sehon 2005, for instance). This approach is widely rejected. But no one has argued, as far as I know, that non-causal views are phenomenologically inadequate. In fact, the very existence of non-causal views would be a rather curious state of affairs if even the performance of basic actions presented itself as efficient causality.

CCW. But once we reject CCW, Wegner's claim that conscious will is an illusion (ICW) loses all its challenging force. In fact, the very possibility of this illusion disappears, once we reject the suggestion that the sense of agency represents one's intentions as causally efficacious. However, this response to ICW does not affect Wegner's argument for the model of AMC.

3. The argument for the model of AMC

Wegner's argument is based on a large number of empirical studies, experiments, and observations about pathological, abnormal, or simply curious instances of human agency. Due to limitations of space, it is impossible to provide an account of all the cases, and it is also impossible to show for each individual case that it is compatible with RMC. Fortunately, Wegner has divided the bulk of the evidence into two groups: "automatisms" and "illusions of control" (2002, chapter 1). This makes the task of providing a response to the overall argument manageable. We can show that a model of RMC can accommodate the evidence by showing that it can accommodate paradigm examples of automatisms and illusions of control.

In general, automatisms are cases of "doing without the feeling of doing" (ibid.): the agent performs an action without having the relevant conscious intention and without a sense of agency. In illusions of control there is a "feeling of doing without doing": the agent has a conscious intention and a sense of agency despite the fact that no matching action is performed. Examples of automatisms are the anarchic hand syndrome, utilization behavior, table turning, automatic writing (Ouija boards), pendulum divining, hypnosis, and other spiritualist "experiments".[5] Illusions of control occur, for instance, in cases where the consequences of an action coincidentally match with the intended consequences, in cases where one perceives someone else's (or an artificial) limb that coincidentally moves in place of one's own, and in cases where phantom limb patients report a sense of agency with respect to their missing limb.

Automatisms and illusions of control are, in effect, evidence for the possibility and reality of dissociations between the sense of control and actual control. As I understand it, Wegner's main argument is that the model of AMC provides the best explanation of such dissociations. In particular, Wegner argues that the model of AMC provides a better explanation than any

[5] It has been argued that automatisms do not raise a problem for the commonsense conception of agency, simply because automatisms are not *intentional* actions (Malle 2006, for instance). But I think that this is beside the point. Wegner's challenge is based on the observation that *actions* can come apart from the sense of acting, where "action" seems to be defined as goal-directed movement that is not necessarily intentional. Of course, one could define overt actions as intentional movements. But this would beg the question against Wegner.

alternative model of RMC. We can distinguish here between the following three points on which this inference to the better explanation is based.

Firstly, according to the model of AMC, there is no causal connection between conscious intentions and matching actions, and there may be no causal connection between the sub-personal mechanisms that produce them (figure 1): the sub-personal mechanisms may generate intentions without generating matching actions, and they may generate actions without generating intentions. According to SCW, the sense of agency is produced by an independent mechanism of self-interpretation (that is governed by the principles of priority, consistency, and exclusivity). Given this (the conjunction of AMC and SCW), dissociations between actions, conscious intentions, and the sense of agency are not only possible, but they are to be expected.

Secondly, Wegner suggests that any model of RMC would have a theoretical disadvantage, because it would have to explain all the pathological, abnormal, and curious cases in a piecemeal fashion. Presumably, this would be quite difficult, and it is likely to give rise to various *ad hoc* explanations. It would, in any case, be inferior to the parsimonious and unified explanation that is provided by the model of AMC in conjunction with SCW (see ibid.: 143-44).

Thirdly, according to Wegner, an ideomotor theory of behavior causation could explain various automatisms. However, "most people who have thought seriously about ideomotor effects have been led to propose that such effects are caused by a system that is distinct from the intentional system of behavior causation" (ibid.: 130). In order words, a model of RMC could be supplemented with a theory that can explain some of the cases (such as an ideomotor theory of automatisms). But it would still be inferior to the model of AMC in at least two respects. It would not explain all of the cases, and it would be less parsimonious in the sense that it would have to stipulate a distinct mechanism of behavior causation.

One might think that this last point is rather weak, because dual process theories are empirically well supported (for a review see Evans 2008). However, it is not obvious that support for dual process theories amounts to support for the thesis that there are distinct mechanisms of behavior causation. It may be, for instance, that two distinct systems of cognition feed into one mechanism of behavior causation or that distinct types of processes are implemented by one and the same mechanism. Moreover, it is questionable that the empirical evidence really supports dual process theories (for critical reviews see Osman 2004 and Keren & Schul 2009). For this reason, I shall assume, with Wegner, that a model which does not resort to the assumption of

distinct behavior causation mechanisms is preferable to a model that does (other things being equal).[6]

In addition, it seems that Benjamin Libet's classic experiment on the initiation of action and the empirical evidence concerning the confabulation of reason explanations provide further support for the view (Wegner 2002: 49-55 and 171-186). We will turn to this further below, and I will argue that neither the Libet experiment nor the evidence on confabulation provide any direct support for the model of AMC. In the following two sections, I will propose a rival model of RMC, and then we will turn to examples of automatisms and illusions of control.

4. Towards a model of RMC: Assumptions

The model of RMC will be based on a number of assumptions that are introduced in this section. Wegner explicitly endorses only one of them. But it will become clear that all of them are consistent with the broadly naturalistic approach that Wegner subscribes to (2002: 21-26), and it will become clear that none of them begs the question against the model of AMC.

(A1) The first assumption is that consciously accessible mental states and events are *realized* by sub-personal states and events (Chalmers 1996, Kim 1998, and Shoemaker 2007, for instance). This is compatible with reductive physicalism, non-reductive physicalism, and functionalism (which I take to be a form of non-reductive physicalism).[7] It might be helpful to point out here that the states which realize accessible mental states are in the cognitive neurosciences commonly referred to as the "neural correlates" of the mental states in question. In contrast, Wegner assumes that conscious mental states are *caused* by sub-personal states and mechanisms. But nothing of substance hangs on this. In particular, neither the argument for the model of AMC nor his rejection of RMC depends on this. (Consider figure 1: we could assume that the upper-left upward arrow represents the relation of realization, rather than causation, without changing the substantial and controversial claims of the model.)

[6] RMC provides a straightforward explanation of the apparent fact that conscious intentions are often followed by matching actions. A theory that denies RMC must provide an alternative explanation. According to Wegner, the correlation between conscious intentions and actions can be explained in terms of its function for social interaction: conscious intentions provide us with "previews" of our actions, which allow us to communicate our goals and plans, and which prompt us to take responsibility (2002: 325-28 and Wegner 2008). This explanation is less parsimonious than the one provided by RMC. But I shall assume, with Wegner and for the sake of argument, that the theoretical virtues of the model of AMC outweigh this disadvantage.

[7] There is a hierarchical multitude of sub-personal levels of explanation, including levels of computational, neural, and physical explanation. I shall ignore this complication, and we may assume that personal level states are *ultimately* realized by physical states (Chalmers 1996, Kim 1998, and Shoemaker 2007).

(A2) Conscious intentions are not necessary for the initiation and guidance of controlled movements. As mentioned, one can bring about coordinated and controlled movements by means of brain stimulation without bringing about any corresponding conscious state (Penfield 1975 and Desmurget et al. 2009). This suggests that the sub-personal processes that realize conscious intentions are upstream of the motor control system, so that it is possible to bypass the formation of conscious intentions in the causation of coordinated and controlled movements (see Frith et al. 2000 and Desmurget & Sirigu 2009). This assumption is clearly compatible with the model of AMC.

(A3) Actions are distinct from bodily movements. Actions belong to the personal level of explanation, because they are typically explained in terms of accessible mental states, whereas bodily movements are typically explained in terms of neuro-physiological processes. Arguably, overt actions are token-identical with bodily movements, but there is good reason to think that they are not type-identical. For instance, whenever I raise my arm, my arm rises. But it is not the case that whenever my arm rises, I raise my arm. Given this, it is plausible to assume that overt actions are also realized by sub-personal events (bodily movements, in particular). Wegner does not distinguish between actions and movements, but neither the argument for the model of AMC nor the rejection of RMC depend on this issue.[8]

(A4) The philosophical problem of causal exclusion has a solution. The problem is, very roughly, that the causal sufficiency of physical states and events in the causation of behavior appears to exclude any substantial causal role for mental states and events (Crane 1995 and Kim 1998, for instance). This problem has been the subject of much debate within philosophy, and there is no uncontroversial solution. Nevertheless, we may *assume* that there is a solution, simply because Wegner's challenge to mental causation is not based on considerations concerning causal exclusion. Furthermore, we may assume that genuine mental causation does not require the "downward causation" of sub-personal events, provided that actions are located at the personal level (Gibbons 2006, for instance). This is also a controversial issue, but nothing of substance hangs on it here.

[8] In the philosophy of action, it is common to hold that actions *are actions* in virtue of being caused by personal level states and events. The assumption of this view would obviously beg the question against Wegner, who argues that actions do not have personal level causes at all. Assumption A3, however, assumes only that actions belong to the personal level because we explain them in terms of personal level states. This is compatible with the possibility that actions are not caused by personal level states, because it might be, for instance, that explanations in terms of desires, beliefs, and intentions are mere rationalizations.

(A5) The sense of agency is a complex and graded phenomenon. Wegner agrees that the sense of agency can be more or less vivid. In the "I-Spy" experiment, for instance, subjects are asked to report the perceived degree of intentionality on a percentage scale (Wegner 2002: 75). In more recent work, Wegner and colleagues have suggested that the sense of agency is modulated by both internal commands and external cues in various ways and to various degrees (Wegner et al. 2004, Aarts et al. 2005, and Moore et al. 2009). Wegner also shares the assumption that the sense of agency is complex. According to the original account (SCW), the sense of agency has three distinct sources (priority, consistency, and exclusivity), and Wegner suggested that the absence of any one of them tends to undermine the sense of agency (2002: 69-70). What was missing was a conceptual framework for the categorization of different kinds of the sense of agency. As mentioned (section 2), it is now common to distinguish between the sense of agency and post-act judgments of agency. In more recent work, Wegner and colleagues have acknowledged this distinction (Moore et al. 2009).

(A6) The sense of agency is generated, in part, by a sub-personal comparator mechanism of motor control. In broad outline, this comparator model says the following. Whenever a motor command for the performance of a bodily movement is generated, a copy of the command is used to produce a prediction of the movement. This prediction (also called "forward model") is then used for a comparison between the predicted end state of the movement and the intended end state, and for a comparison between states of the predicted movement and sensory feedback concerning the actual movement. This computational model is empirically well supported, and it is now widely assumed that the initiation and control of movements is achieved by such a sub-personal comparator mechanism of motor control (Wolpert & Kawato 1998, Scott 2004, Haggard 2005, Christensen et al. 2007, Andersen & Cui 2009, and Desmurget & Sirigu 2009, for instance). Further, it is now widely assumed that this system contributes to the sense of agency. It is assumed, in particular, that positive matches in the comparators contribute to the generation of the sense of agency and that mismatches result in error signals that undermine the sense of agency (Frith et al. 2000, Gallagher 2007, Bayne & Pacherie 2007, and Synofzik et al. 2008, for instance).

Initially, the comparator model of the sense of agency was the main alternative to the self-interpretation model proposed by Wegner and others. There is, however, now a growing consensus that these two approaches are better construed as complementing each other, rather than as rivals (Bayne & Pacherie 2007, Gallagher 2007 and Synofzik et al. 2008). The

comparator model seems well suited to explain the online and basic sense of agency, whereas the self-interpretation model can explain why judgments about our own agency are subject to various biases that may lead to post-act confabulations. Wegner and colleagues have acknowledged that internal signals and comparisons with feedback contribute to the sense of agency (Wegner et al. 2004 and Moore et al. 2009). However, they have also provided evidence which suggests that the comparator model cannot fully explain the sense of agency. Evidence shows, in particular, that the sense of agency is partly modulated by matches at the personal level between conscious representations and types of actions (Wegner et al. 2004; see also Synofzik et al. 2008). Intuitively, this makes sense, as it seems clear that positive matches between conscious intentions and subsequent actions should contribute to the sense of agency. However, the evidence suggests that the match with a mere representation or thought of the action is sufficient to enhance the sense of agency. Given that there is a clear difference between *thinking about* A and *intending to* A, this means that a match with a conscious intention is not necessary, at the personal level, to enhance the sense of agency. (We will return to this below.)

Intuitively, it seems also clear that the awareness of movement initiation is an important and perhaps necessary part of the sense of agency (Pacherie 2007). Empirical evidence shows that this "motor awareness" is generated by internal signals, rather than by the execution of the movement itself (Engbert et al. 2008 and Desmurget et al. 2009). It has been suggested that the relevant internal signals are the formation of the movement prediction and the release of the motor command (Frith et al. 2000 and Desmurget & Sirigu 2009). This awareness of movement initiation can be distinguished from prior intentions. The former has been described as an "urge to move" (Fried 1991 and Desmurget & Sirigu 2009), and, perhaps more plausibly, as the conscious awareness of "being about to act" (Pacherie & Haggard 2010). Prior intentions, on the other hand, have usually conceptually complex contents that can be characterized as "act-plans" (Goldman 1970, Bratman 1987, Andersen & Buneo 2002, and Mele 2009, for instance). Typically, act-plans specify goals and means (as in the intention to move the cursor by moving the mouse, for instance). In the limiting basic case, the plan specifies only a certain type of action (as in the intention to raise an arm). But even in this limiting case, it seems that we can first form the prior intention and then have the sense of initiating the movement. Brain stimulation experiments support this distinction. They suggest, in particular, that "parietal cortex stimulation generates conscious intentions to move", whereas "stimulation of the SMA [supplementary motor

area] triggers feelings of an urge to move that reflect the imminence of a motor response" (Desmurget & Sirigu 2009: 413).

We have now distinguished between a number of sources that contribute to the sense of agency: (1) the formation of a movement prediction, (2) the release of the motor command, (3) matches in the sub-personal comparator mechanism, (4) the match between a conscious representation and the type of action, and, in particular, (5) the match between a conscious intention and the type of action. I shall assume that all five factors contribute to the sense of agency. In the light of the empirical evidence on the sense of agency, it is plausible to assume that the combined occurrence of *some* of the components can result in a minimal or basic sense of agency, whereas only a coordinated occurrence of all the components generates a full sense of agency. But the evidence suggests also that the contribution or weight of the individual components can vary from case to case (see Frith et al. 2000, Synofzik, et al. 2008, and Moore et al. 2009). For instance, in the absence of efferent motor commands or proprioceptive feedback, more weight might be given to visual feedback. The extent to which judgments of agency are based on veridical memories of an experienced sense of agency may also vary from case to case. It seems plausible to assume that judgments of agency are usually based on experiences of a sense of agency (Bayne & Pacherie 2007). But post-act judgments are also subject to various biases. For instance, an unrealistic self-conception of one's agency may occasionally override weak or vague memories concerning the presence (or absence) of a sense of agency.

A full and satisfactory specification of the weighting mechanisms that modulate the sense of agency and the formation of judgments of agency has not been developed yet, and further research is required in order to determine the contribution and interaction of the mentioned components. But the outlined model of the sense of agency is sufficiently precise for my purposes, and we can *assume* this model, for the following reasons. Firstly, my aim here is not to disprove Wegner's original model of the sense of agency (SCW), but to defend RMC against the argument for the model of AMC. Secondly, as pointed out, Wegner and colleagues have now acknowledged that the sense of agency is partly based on internal signals and feedback comparisons (as postulated by the comparator model). Thirdly, and most importantly, the outlined model of the sense of agency is in principle compatible with the model of AMC and it does not presuppose RMC. It says that matches between conscious intentions and actions contribute to the sense of agency, but it does not presuppose that conscious intentions *cause* matching actions.

5. A model of RMC

On the basis of these assumptions, I can now propose a model of RMC, according to which the mental causation of actions at the personal level is realized by causal processes and mechanisms at the sub-personal level (A1). The sub-personal processes that realize intentions are upstream of the motor control system (A2). The movement is initiated by the release of a motor command and controlled by means of internal predictions (feed-forward models) and comparisons with sensory feedback (A6). This process realizes the mental causation of an intentional action (A3), provided that the bodily movement realizes an act-type that matches with the intention's content. Figure 2 illustrates this model of RMC.

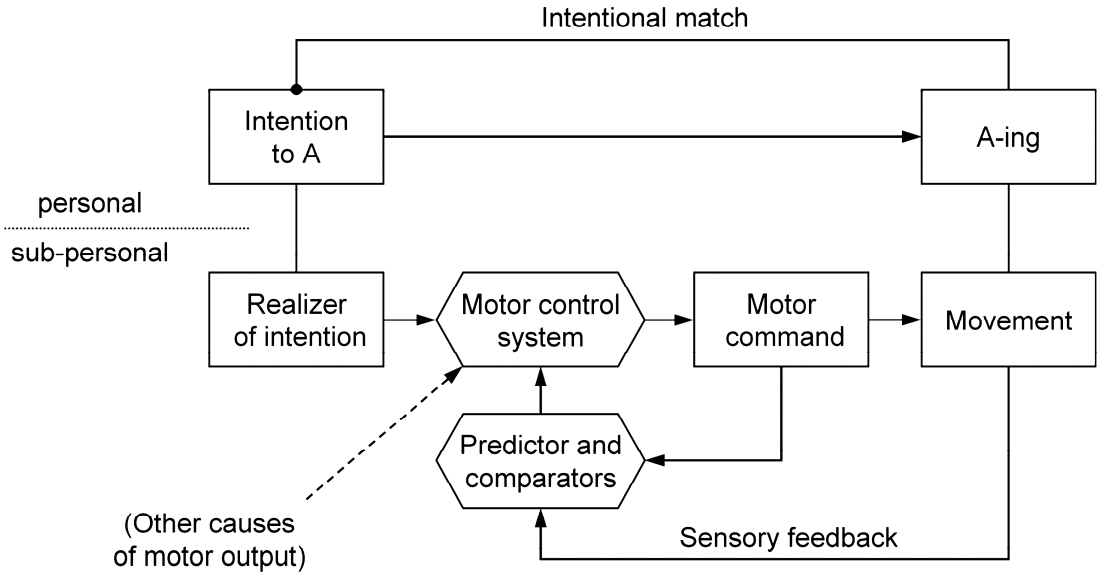

Figure 2. Model of real mental causation (RMC)
(Including sources of the sense of agency)

The model assumes that personal level states and events are realized by sub-personal level states and events, indicated by the vertical lines that connect the intention and the action to the sub-personal level. At the neural level, the selection and execution of actions involves a complex prefrontal-parietal network (Scott 2004, Haggard 2008, Andersen & Cui 2009, and Desmurget & Sirigu 2009).[9] The levels should be understood as levels of description and explanation, and the

[9] Roughly, dorsolateral prefrontal cortex is associated with the selection of goals (Rowe et al. 2000); regions of both medial and lateral prefrontal cortex are involved in the representation and storage of prior intentions (Miller & Cohen

model is compatible with the thesis that personal level entities are (type- or token-) identical with lower level entities, but it is not committed to it. It also stays neutral on the question of whether or not personal level theories are reducible to lower level theories.

This model of RMC is fully compatible with the outlined model of the sense of agency (A6), and figure 2 indicates some of its components (the sub-personal comparator system and the "intentional match" at the personal level). It is partly based on the models proposed by Frith et al. 2000, Haggard 2005, Synofzik, et al. 2008, and Desmurget & Sirigu 2009. But there are some significant differences. Unlike the models proposed by Frith et al., Haggard, and Desmurget & Sirigu, it embeds the comparator system within a model of levels of explanation, and it preserves thereby the distinction between actions and bodily movements (A3). Synofzik et al. have proposed a model of the sense of agency, which is silent on questions concerning the causal efficacy of intentions. Their model suggests top-down and bottom-up interactions between two levels, whereas the model of RMC holds that personal level states are *realized* by sub-personal level states (A1).

Note that figure 2 illustrates only the basic case: the mental causation of a *basic* action by a prior intention. As mentioned, most actions are complex, in the sense that we usually pursue some further goal when we perform a basic action. But all actions require the performance of some basic action. Recall also that prior intentions should be distinguished from the awareness of movement initiation, and note that the term "prior" intention is potentially misleading, as prior intentions may precede and *accompany* the performance of actions. We can further distinguish between "distal" and "proximal" intentions (Mele 2009). As the terms suggest, the former are further removed in time from the execution of the action than the latter. But I take it that the distinguishing feature consists in the fact that distal intentions are first stored in memory and later retrieved or activated, whereas proximal intentions are transformed into motor commands without delay.

Finally, I should point out that the model is clearly empirically testable. According to the model of AMC, conscious intentions are merely *followed* by matching actions. The model of RMC gives an account of how conscious intentions can be the real causes of action, and it holds, in accord with the thesis of RMC, that conscious intentions *cause* the relevant basic actions

---

2001 and Haynes et al. 2007); premotor areas and the primary motor cortex are involved in motor planning and self-monitoring of movement execution (Berti et al. 2005 and Christensen et al. 2007); activity in regions of the parietal cortex are correlated with conscious intentions to move (Andersen & Buneo 2002 and Desmurget et al. 2009) and with the processing of feedback signals (Farrer et al. 2003); the SMA is associated with the release of motor commands (Ball 1999 and Sumner 2007) and with the awareness of movement initiation (Desmurget et al. 2009).

(*ceteris paribus*). This difference between mere correlation and causation—a crucial difference between the two models—, can be empirically tested by means of controlled interventions on intentions. We will return to this further below. Further, the model holds that conscious intentions are not necessary for the causation of controlled movements, and it predicts that other factors, such as unconscious mental states or environmental stimuli, can lead to the execution of coordinated movements. As mentioned (see A2), there is empirical evidence for this.

In the following two sections, I will show that this model of RMC can accommodate paradigm examples of automatisms and illusions of control. The main task will be to show that each type of case is compatible with RMC. But I will also indicate how the presence (or absence) of a sense of agency can be explained for each type of case.

6. Illusions of control

6.1. Lucky coincidences

The most straightforward and common instances of illusions of control occur when there is a coincidental match between the agent's intention and some event or movement. This is familiar, as Wegner (2004) points out, from interactions with machines. Return to the example of moving the cursor on the computer screen. Suppose that you move the mouse to the left in order to move the cursor to the left. Unbeknownst to you, the mouse is not connected to the computer. But due to some fluke, the cursor moves in accordance with the movement of the mouse (at the right time, the right distance, and so forth). As a result, you feel in control over something that you have no control over.

All cases of this kind are unproblematic for the model of RMC, because they concern only act consequences (or non-basic actions). The agent feels in control over bringing about a consequence that matches with the content of the intention by coincidence. But there is no reason to think that everything that the agent does matches with the intention by coincidence. In particular, there is no reason to think that the successful performance of the basic action is due to a coincidence. In the example, it seems to you that you are moving the cursor. You are wrong about that. But it also seems to you that you are moving the mouse by moving your hand. For all we know, you are right about that—for all we know, this is not a lucky coincidence. So, the illusion is only *partial*. Given this, it may well be that the intention is the real cause of a matching action. It may well be that your intention to move the cursor (and mouse) by moving your hand causes the movement of your hand (and mouse). Moreover, it is no surprise that the agent feels in

control, because all the contributors to the sense of agency are in place: the action and its consequences match with the content of the intention; we have reason to assume that there are positive matches in the sub-personal comparator mechanism, because the relevant basic action is executed; and we have reason to assume that a sense of movement initiation is generated by the release of a motor command.

6.2. Phantom limb movement

Consider next the pathological phenomenon of "phantom limb movement". The term "phantom limb" is commonly used to refer to the sensation that a missing limb is still attached. The phenomenon of phantom limb "movement" occurs when phantom limb patients experience also a sense of control over their missing limb. More specifically, when they are asked to move their phantom limb, they report a feeling of control—they have the sense that the phantom limb is moving in the intended way. This phenomenon is difficult to explain, partly because there are subtle differences between different cases and because the phenomenon changes over time. In particular, the illusion is more common and more vivid in the early stages after the loss of the limb, and it tends to fade after some time (Frith et al. 2000 and Wegner 2002: 40-45). But the outlined model of the sense of agency can nevertheless explain why phantom limb patients experience some minimal sense of agency over their missing limb. We can assume that patients form a conscious intention to move (in response to the request from the experimenter). This should result in the formation of a movement prediction and in the release of a motor command. This, in turn, should lead to matches in the internal feed-forward comparison between the predicted state and the intended end state of the movement. In conjunction with the sense of movement initiation, which should be generated by the release of a motor command, this may well be sufficient for a minimal sense of agency.[10] In some cases, it might also be the case that there is some positive proprioceptive feedback from muscle contractions in the stump (Wegner 2002: 41). This should enhance the sense of agency, which is compatible with Wegner's description of the cases, according to which the sense of agency in phantom limb movements may be more or less vivid.[11]

[10] In contrast, it is difficult to see how Wegner's model (SCW) can explain the sense of agency here. The conditions of priority and consistency are both violated, because the intentions are not followed by actual movements. And it seems implausible to suggest that the satisfaction of exclusivity will by itself lead to a sense of agency.

[11] Frith et al. (2000: 1778-79) have argued that the comparator model can also explain why the sense of agency tends to fade after some time. Roughly, the suggestion is that the movement predictions will be modified and updated only after some period of time. This updating should lead to negative feedback signals in the internal feed-forward loop.

More importantly, the phenomenon of phantom limb movement does not raise any problems for RMC, as both the thesis and the model of RMC concern only cases in which an intention is in fact followed by a matching action. One could argue, even, that this provides some support for the model of RMC (and for the outlined model of the sense of agency), because it seems that a good explanation of why the patients experience a sense of movement initiation is provided by the assumption that their conscious intentions generate the relevant internal signals (the release of motor commands and the formation of movement predictions).

6.3. Alien limb movement

The third and final illusion of control that I shall consider stems from the perceived movement of alien limbs: someone else's or artificial limbs that are perceived to be moving in place of one's own. Wegner and colleagues have conducted interesting experiments on this phenomenon (Wegner et al. 2004). Subjects were watching themselves in a mirror while a confederate, who was standing right behind them, was moving her arms in accordance with instructions. The subjects' own arms remained at their sides and under a smock. This generated an illusory sense of agency for the observed movements. In the most relevant trial (experiment 1), the subjects could hear the instructions that were given to the confederate. The results show that this match between the observed movements and the overheard instructions enhanced their sense of agency for those movements to a significant degree. One interesting point here is that there is no reason to think that subjects formed intentions in accord with the overheard instructions. It was clear that the instructions were directed at the confederate, and subjects themselves were explicitly instructed not to move (which is incompatible with intending to follow the overheard instructions). Given this, the experiment seems to show that matches between actions and mere representations or thoughts can contribute to the sense of agency. But it is clear that this type of case does not raise any problems for RMC. Subjects were instructed to keep their arms at their sides, and this is what they did. The evidence is not only compatible with RMC, but the thesis of RMC provides a good explanation of why the subjects did what they were told to do: they kept their arms at their sides because they intended to do so in accord with the instructions.

How can we explain the sense of agency here? It should be no surprise that subjects have some sense of agency, because they successfully execute the instructions that were given to them (not to move their own arms). This explains why they have some sense of agency. But it does not

---

As a result, the motor control system should eventually cease to generate motor commands for the missing limb. This would explain why patients lose a sense of movement initiation only after some period of time.

explain a sense of agency for the confederate's movements.[12] Note, first of all, that there is ambiguous feedback at the personal and sub-personal levels. Subjects intend not to move. Internal feedback signals should confirm that this intention is executed. But these internal signals are in conflict with sensory feedback from the observed movements and with the content of the mental representations that are induced by the overheard instructions, while the sensory feedback from the observed movements is *in accord* with the mental representations of the overheard instructions. Secondly, as Wegner and colleagues note, classic EMG studies have shown that instructions to merely think about certain movements induce corresponding muscle potentials without resulting in overt movements (ibid.: 847). Given this, we may speculate that overhearing instructions to move induces activity in the motor control system that is automatically inhibited in accord with the agent's intention not to move (we will return to this inhibition mechanism below). Further, sub-threshold muscle potentials might provide more ambiguous feedback: proprioceptive feedback that is consistent with both the observation of arm movements and with the overheard instructions, but that does not match with the agent's intention not to move. Taken together, this can explain why subjects experience some sense of agency for the confederate's arm movements. But the model would predict that the sense of agency is minimal or weak. Subjects have no intention to move, and there is, presumably, no sense of movement initiation, as no motor commands are released. This appears to be compatible with the results of the experiment (ibid.: 841). The mean rating of perceived agency in the experimental condition (M = 3.00) was significantly greater than in the control condition (M = 2.05), but still rather low on a scale from 1 ("not at all") to 7 ("very much").

7. Automatisms

7.1. Utilization behavior

The pathology known as "utilization behavior" occurs in patients with frontal lobe lesions, and it consists in the automatic execution of stimulus driven actions (Lhermitte 1983, Frith et al. 2000, and Archibald et al. 2001). Typically, utilization behaviors are instrumentally adequate, but they are not intended. For instance, placing a toothbrush in front of the patient induces tooth brushing behavior, placing a banana in front of the patient results in peeling, and so on. In such cases, it seems clear that the patients do not have prior intentions to perform these actions, and usually

[12] A mere match between a mental representation and an observed movement is usually not sufficient for a sense of agency. For instance, a match between the expectation that "I'm going to sneeze" and sneezing does not result in a sense of agency.

they are unable to suppress the response. Interestingly, patients do not report any disturbance in their sense of agency. Utilization behavior could therefore be characterized as a *partial* automatism: it is action without conscious intention, but with an undisturbed sense of agency.

According to a traditional explanation (Shallice 1989 and Archibald et al. 2001), cases of utilization behavior support the view that external stimuli can automatically induce overlearned responses by activating stored motor schemas (routines or programs). It assumes that activated motor schemas are in healthy subjects released only if the actions are in accord with the agent's intentions and long-terms plans, and that they are automatically inhibited otherwise. Evidence suggests that the SMA is the primary neural correlate of this inhibition mechanism (Goldberg 1985, Ball et al. 1999, and Sumner et al. 2007). This explains why the responses are enacted in patients with frontal lobe lesions, provided that this involves damage to the SMA. There are, however, two problems with this explanation. Firstly, it is implausible to explain all utilization behaviors as externally triggered reflexes or overlearned routines, because the execution of many utilization behaviors is controlled and fine-tuned by the particular features of the situation (and target objects). Secondly, there is no explanation of why patients do not report a disturbance in their sense of agency. Both shortcomings can be overcome if we supplement the traditional explanation with the comparator model of motor control. On this view (Frith et al. 2000), external stimuli automatically induce activity in the motor control system, bypassing the formation of consciously accessible intentions. In line with the traditional view, it assumes that in healthy subjects motor commands are released only if the actions are in accord with the agent's intentions and plans, and that they are automatically inhibited otherwise. In patients with frontal lobe lesions, this inhibition mechanism is damaged, and so stimulus driven motor commands are released even if the agent has no corresponding intention. The subsequent processing of feed-forward and feedback signals explains why the resulting movements are controlled and fine-tuned in accord with the features of the particular situation. Further, a lack of error signals in the sub-personal comparator system and the release of motor commands explain why patients have a basic sense of agency (including a sense of movement initiation).[13]

It is clear that utilization behavior does not raise a problem for RMC, because neither the thesis nor the model of RMC requires that all actions must be preceded and caused by conscious

[13] Utilization behavior raises similar problems as phantom limb movements for Wegner's original account of the sense of agency (SCW). The conditions of priority and consistency are violated (trivially, as there is no conscious intention). But it is implausible to suggest that the condition of exclusivity alone can explain the fact that the sense of agency is undisturbed.

intentions. In particular, utilization behaviors cannot be counterexamples to the thesis of RMC, given that the patients have no conscious intentions to perform the actions in question, and the model of RMC is compatible with the assumption that motor commands can be generated in response to external stimuli.

7.2. The anarchic hand syndrome

Perhaps the most striking example of an automatism is the "anarchic hand syndrome" (Goldberg et al. 1981 and Frith et al. 2000, for instance). Patients with this neurological disorder report, typically, that one of their hands is moving "on its own", and sometimes they attribute the actions to some alien force or agent. Anarchic hand movements are coordinated actions. Like in cases of utilization behavior, the movements are goal-directed and unintended. But unlike in cases of utilization behavior, the agent's sense of agency is disturbed and often the patient vigorously denies initiation and control of the movement. Many anarchic hand movements are stimulus driven. Some are highly routine actions, but not obviously stimulus-driven (unbuttoning of the patient's own shirt, for instance). Others are neither stimulus-driven nor routine (attempt at self-strangulation, for instance).

It is no surprise that patients do not have a full and vivid sense of agency, because they have no conscious intentions to perform the actions. But reports suggest that patients lack even a minimal sense of agency. This is puzzling in the light of what has been suggested about utilization behavior. Many instances of the anarchic hand movement are stimulus-driven, like cases of utilization behavior. And like in utilization behavior, this can be explained in terms of damage to the medial frontal regions (SMA, in particular) that are associated with the automatic inhibition of unintended actions (Goldberg 1981, Goldberg & Bloom 1990, and Frith et al. 2000). But unlike in utilization behavior, the sense of agency is disrupted. One possible explanation for the difference is that utilization behaviors are merely unintended, whereas anarchic hand movements are in conflict with some of the patient's intentions, plans, or moral commitments. This results in a glaring mismatch at the personal level, which should disrupt and undermine their sense of agency (and which may lead them to confabulate explanations in terms of control by alien forces or agents).

Utilization behavior, the anarchic hand syndrome, and the differences between them pose difficult challenges for every account of the sense of agency. The proposed explanation is tentative and it raises further questions. But, as before, the main point is that there is no problem

for RMC. Utilization behaviors are unproblematic, because the patients have no relevant intentions. In cases of anarchic hand movements, patients perceive also a discrepancy with standing intentions or commitments. They want to suppress the action, but they have no intentional control over it, due to the damage to the inhibition mechanism. This breakdown in the motor control system clearly violates the *ceteris paribus* clause in the thesis of RMC—anarchic hand movements are clearly cases where other things are *not* being equal. Given this, and given what has been said on utilization behavior, it is also clear that anarchic hand movements are compatible with the model of RMC.

7.3. Spiritualist phenomena

Finally, let us briefly consider spiritualist "experiments", such as table turning, automatic writing (with Ouija boards), pendulum divining, dowsing, and hypnosis (see Wegner 2002). Cases of this kind are notoriously difficult to interpret, partly because there is reason to think that reports from participants are strongly biased. But it is not difficult to see that these cases are compatible with RMC. In cases of Ouija board writing, for instance, participants are asked to move the board slowly in circles, in cases of table turning, participants are asked to put their hands in a certain position onto the table, and so forth. Subjects act intentionally in accord with such instructions, but they lack the more specific intention to move the board in a certain way, or they lack the additional intention to exert pressure on the table in a certain way, and so on. Something like this holds for all the mentioned cases. Given, then, that the agent lacks the specific intention that matches with the behavior in question, there is no problem for RMC. Again, one could even argue that these cases provide some support for RMC, because subjects intend and act in accord with the instructions.

As mentioned, it is not clear what to make of participants' reports concerning their sense of agency. In particular, it is unclear to what extent such post-act judgments are based on veridical memories concerning their sense of agency. In any case, it would be difficult to provide a full explanation of each and every case. But as this is not required for the defense of RMC, I shall restrict the discussion of the sense of agency here to a few general remarks. Two things are in need of explanation. Why do participants perform the particular or additional action (moving the Ouija board in a certain way, exerting a certain pressure on the table, and so on)? And why do they not experience a sense of agency for that? Traditional ideomotor theory provides a plausible starting point for an answer to the first question (see Kunf et al. 2001). Roughly, the idea here is

that unconscious representations, which may be induced by subtle suggestions or by the subject's unconscious wishes, can automatically generate associated actions. We can combine this explanation with the comparator model of motor control, if we drop the assumption that the activation of representations results directly in matching actions (see Jansson et al. 2007). On this view, the activation of unconscious representations leads to matching actions by way of activating the motor control system (by generating motor commands, forward models, and so on). The causal structure is here basically the same as for stimulus driven actions. The difference is only that actions are now generated by subtle suggestions or unconscious desires (see Kirsch & Lynn 1998).

But why do participants report that they are not initiating and controlling the actions in question? The subjects lack the relevant conscious intentions. But the same is true of many habitual and routine actions, which are accompanied by a minimal sense of agency. So, why are automatic routine actions accompanied by a sense of agency, but not the spiritualist automatisms? Note, first of all, that automatic routine actions are usually familiar and overlearned actions that serve the pursuit of conscious goals or long-term plans. The automatisms in spiritualist experiments are neither familiar routines, nor are they based on conscious intentions and plans. Secondly, in spiritualist experiments a facilitator often primes the expectation that someone or something else will take control of the action, or the subject already has the expectation that something abnormal is bound to happen. This may induce or strengthen a bias towards the interpretation that one's actions are controlled by someone or something else. Thirdly, participants in spiritualist experiments do often not initiate the action from a state of inaction. Rather, the action often emerges in continuation of an intentional action (moving the Ouija board in circles turns into "writing", for instance). For this reason, subjects might lack an important component of the sense of agency for the action in question. They might lack the sense of movement initiation. Finally, in some of the mentioned cases, subjects perform actions that are usually not considered to be part of a normal subject's act repertoire. In cases of pendulum divining and dowsing, for instance, the movements of the pendulum or the rod are influenced by the smallest of hand movements and muscle contractions. The resulting act consequences are surprising and puzzling, because we tend to assume that we are unable to exercise such a fine-grained kind of control. Given all this, we can see why the sense of agency is disrupted in spiritualist automatisms, but not in automatic routine actions. Recall, here, that the basic sense of agency is thought to be phenomenologically thin, which would suggest that it can be easily

obscured and outweighed by opposing biases and subtle suggestions—particularly if the agent has no corresponding conscious intention and no sense of movement initiation.

8. The unconscious precursors of conscious intentions

Benjamin Libet's well-known experiments concerning the role of consciousness in the initiation of action seem to show that proximal conscious intentions are *preceded* by "specific cerebral processes that mediate the act", on average by 350ms (1985: 529). Libet concluded from this that the "initiation of a spontaneous voluntary act begins unconsciously" (ibid.). A first thing to note here is that it remains unclear whether the conscious events in question are genuine intentions. Libet himself variously used the terms "urge", "wish", "choice", and "intention" interchangeably. It could be that the event in question is an awareness of movement initiation—an awareness of "being about to move"—, rather than a genuine intention. Or it could be, as Keller & Heckhausen (1990) have argued, that the instructions and the experimental setup induced a type of awareness that does normally not precede the execution of movements. Subjects were instructed to perform a certain type of movement whenever they felt like doing so. They were instructed, in particular, "to let the urge come on its own" (Libet 1985: 531). Libet's aim was to isolate and study self-generated and *spontaneous* actions that are not triggered in response to external stimuli. But the experimental setup created a highly unusual context and it is questionable that the instructions resulted in spontaneous actions. It seems, rather, that subjects were instructed to act in response to *internal* stimuli: the "perceived urge to move can be interpreted as an internal stimulus which triggered the release of a predefined motor act" (Keller & Heckhausen 1990: 352). Moreover, the instructed and highly unusual "selective attention" to look for an urge of "wanting to move" may have resulted in the awareness of a process that is normally unconscious (ibid.: 359).

More importantly, the experiment fails to support a model of AMC even if we assume that the subjects reported the awareness of proximal intentions. Under this interpretation, the experiment shows that proximal intentions have unconscious precursors, and it suggests that proximal intentions do not initiate the intended act. But this is perfectly compatible with RMC, as causation by conscious intentions requires neither the absence of unconscious precursors nor that conscious intentions *initiate* the act (in the sense of being a first or uncaused cause). In particular, RMC is compatible with the claim that, at the sub-personal level, the neural correlates of intentions are parts or segments in the causal chains that culminate in movements (see Mele 2009). In connection with this it is important to note that unconscious precursors are

unproblematic as long as the subsequent formation of the proximal intention is in accord with a distal intention or with some of the agent's other desires, beliefs, or commitments. This was clearly the case in the Libet experiment, because subjects had the distal intention to perform the predefined type of movement. The fact that distal intentions lead to the formation of more specific proximal intentions via mediating unconscious neural processes is not particularly surprising, and it seems unproblematic as long as there is reason to think that these unconscious processes have their source in accessible mental states that render the choice intelligible from the agent's point of view (similar considerations apply to the Libet-style experiment conducted by Soon et al. 2008; for more on this see Schlosser *forthcoming*).

9. The post-act confabulation of reasons and intentions

It has been argued that the empirical evidence concerning the confabulation of reason explanations supports the view that ordinary reason explanations of our own actions are in general based on biased processes of self-interpretation and post-act rationalization. Evidence suggests, for instance, that we tend to give reasons that are in line with our self-conception or with an ideal of rational agency, or that we tend to give socially accepted reasons when we are asked to justify our actions (Nisbett & Wilson 1977, Gazzaniga & LeDoux 1978, Haidt 2001, Wilson 2002, and Wegner 2002). We already assume, with Wegner, that *judgments* about our own agency are subject to various biases, and the evidence on confabulation strongly suggests that the self-ascriptions of *some* reasons and intentions are based on self-interpretation (rather than recollection).[14] But the evidence, I will now argue, does not support the generalization of this claim, and it does not support the model of AMC.

Firstly, the most striking cases of confabulation stem from split-brain patients (subjects with a surgically severed corpus callosum). In a particularly straightforward example, the instruction to "take a walk" was presented to the patient's left visual field (processed by the right hemisphere). In response, the patient got up and headed for the door. When asked "Why are you doing that?" the patient replied "Oh, I need to get a drink"—a clear case of confabulation (Gazzaniga 1998: 133). It is traditionally thought that in most subjects the left hemisphere is dominant in the processing of language and speech. In split-brain patients, the left hemisphere does not have access to what is presented to the right hemisphere. This, in conjunction with the evidence on confabulation, has lead to the suggestion that the left hemisphere hosts a cognitive

[14] *Confabulated* explanations are manifestly false, whereas post-act *rationalizations* may be true (see Nisbett & Wilson 1977: 253).

module that is dedicated to the interpretation of our actions—the left-brain "interpreter" (Gazzaniga & LeDoux 1978 and Gazzaniga 1998). Whatever the truth on this matter, the evidence does not support a model of AMC. It does not even support a model of AMC for split-brain patients. Return to the example. Either the patient formed a conscious intention in response to the instruction, or he did not. If he did not, then the case does not raise a problem for RMC. If he did form a conscious intention (in response to the instruction), then it might well be that this intention was causally efficacious in the initiation and guidance of the action. This assumption of mental causation is compatible with the evidence, and it provides, once more, a good explanation of why the patient acted the way he did. The fact that the patient confabulated a post-act explanation does nothing to undermine this. It suggests only that the postulated interpreter module could not access a memory of having acted intentionally in response to the instruction. Either way, the evidence is compatible with RMC, and it fails to support the model of AMC.

Further, it is far from obvious that evidence from split-brain patients supports any interesting generalizations concerning the agency of healthy subjects. But even if it did, it would not support a general model of AMC, because it does not even support a model of AMC for split-brain patients (as I have just argued). If generalized, the evidence suggests, at best, that we tend to confabulate *under certain circumstances*—in particular, when we cannot access the intentions and reasons we acted for. In some cases, this inaccessibility may be permanent and systematic (as in split-brain patients). In other cases, it might be due to the fact that the intentions and reasons were simply not stored in memory. Suppose, for instance, that you cannot recall whether or not you locked the door when you left home this morning. It may well be that you locked the door consciously and intentionally. But perhaps you have no memory of this, because this routine action was not significant or vivid enough to enter long-term memory. In any case, none of this raises a problem for RMC. If the agent has no conscious intention, there is no problem for RMC. And if the agent has a conscious intention, there is no reason to think that the intention was not efficacious in the initiation and guidance of the matching action. In other words, all of this is perfectly compatible with RMC, and it fails to provide any support for the model of AMC.

Secondly, in many of the cases where subjects tend to confabulate post-act reason explanations, there are either no salient reasons or there are no reasons at all. Consider, for instance, the famous position effect (Nisbett & Wilson 1977). Subjects were asked to evaluate articles of clothing and to select the best quality product. In one study, subjects were shown four different night gowns. In a second study, subjects were shown four identical nylon stockings. In

both cases, subjects showed a strong tendency to rate the article positioned at the far right highest. Subjects were entirely unaware of this position effect, and when they were asked to explain their choices, they referred to perceived differences in quality (in both studies). What is never mentioned, however, is that it would be very odd if they had given the position as a reason, because the position is not a good reason to prefer any one item over the others. Here it is crucial to be aware of the following ambiguity. There is a sense in which all causes of actions are reasons, simply because a "cause" is the "reason why" something occurs. But there is also a sense in which not all causes are reasons, because not all causes of actions provide rational grounds that favor those actions. When we seek a reason explanation of an action, we seek usually reasons in the second sense—we seek an explanation in terms of rational grounds. Return to the two studies. In the second study, there are no good reasons that one could give, because the items are qualitatively identical. In the first study, it is unclear whether or not there are any salient reasons. Perhaps it was not difficult to spot some differences. But there is no reason to assume that these differences supported a clear qualitative ranking. Given this, it is not so surprising that the subjects confabulated reason explanations. Probably they wanted to (or felt the need to) give answers in order to comply with the experiment, or in order to preserve the self-concept of being a rational consumer (or something along those lines). In the first study, there were no salient reasons, and in the second study, there were no good reasons at all. And so the only way in which subjects could possibly give reasons is by making them up.[15] Again, this kind of evidence shows only that we tend to confabulate under certain circumstances, and we should not draw any general conclusions—we should, in particular, not draw conclusions about cases in which there are salient or explicit reasons.

Thirdly, the majority of the evidence concerns the confabulation of reason explanations. Wegner suggests that this amounts to evidence for the confabulation of intentions (2002: 171-81). But the *reasons* for which we act can be distinguished from the *intentions* with which we act. Usually, the reasons that favor an action are, *ipso facto*, the reasons that favor the choice of that action (they favor the formation of an intention to perform that action). In other words, intentions are usually based on reasons, which suggests that intentions are distinct from reasons. In order to see why this distinction is relevant here, it will be helpful to consider the following thought

[15] Malle (2006) suggests the following interesting interpretation. It is consistent with the evidence that the position did not influence the choices directly. Rather, the position might have inducted the (non-veridical) representation of a qualitative difference. If that is correct, it may well be that subjects gave the reason on the basis of which they made their choices.

experiment. Suppose that Paula wants to buy a new mouse for her computer and that she has narrowed her search down to a particular product, which is available in white and gray. Assume that she chooses the white version, because she thinks that it matches better with the color of her computer. What if we asked her why she wants the white version? It is unlikely, I think, that Paula would confabulate an explanation, provided that she has a clear preference. But even if she would confabulate a reason explanation, it does not follow that she would confabulate her intention. It is one thing to be wrong about the reasons for making a certain choice. But it is something else to be wrong about the intention itself. Paula has made up her mind. What if we asked her which version of the mouse she intends to buy? She could communicate this simply by pointing at the white version ("this one")—she could report her intention without reporting any reasons. But if we asked her to give reasons for either option, she could also report reasons without reporting an intention (or choice). She could say, for instance, that the match with the color of her computer is a reason for her to choose the white version. It does not follow from this that she intends to get the white one. This shows that we cannot simply infer the confabulation of intentions from the confabulation of reasons.[16] Moreover, it shows that the evidence concerning the confabulation of reasons is compatible with RMC. Whenever an agent confabulates a reason explanation for an action, the agent may have acted with the conscious intention to perform that action, and the intention may have caused the action. And if the agent did not act with a conscious intention, then there is no problem for RMC either.

Finally, there is plenty of evidence for the claim that providing subjects with good reasons for actions has an effect on their intentions and actions. Most obviously, most experiments in psychology and cognitive neuroscience show that giving subjects instructions, which provide usually good reasons within the context of scientific experimentation, has a strong and reliable effect on their intentions and actions (we will return to this below). But there is also plenty of direct evidence for the claim that reason-giving and conscious intending is causally efficacious. In a meta-analysis, Webb and Sheeran (2006) have collected and analyzed 47 studies in which subjects are given good reasons for significant real-life choices. They have found that the evidence supports the thesis that interventions on an agent's intentions by way of giving good reasons engender the corresponding changes in intentions and actions. In particular, their meta-

[16] This is easily obscured, because reasons often enter into the contents of intentions. Consider, for instance, the explanation that "I did A in order to bring about B". If the reason ("in order to bring about B") is confabulated, then the self-attribution that "I intended to do A in order to bring about B" would be *partly* confabulated. It does not follow, however, that the narrower self-attribution that "I intended to do A" would be confabulated as well.

analysis shows that the intervention of reason-giving has a medium-to-large effect on changes in intentions, and that changes in intentions have a small-to-medium effect on changes in behavior. One might think that this result is problematic, because the effect size of changes in intentions is only small-to-medium. But I think that the analysis only confirms reasonable and honest expectations concerning the efficacy of intentions. Most of the studies concern difficult changes in behavior, such as taking physical exercise, wearing a seatbelt, regular use of contraceptives, quitting smoking, and so on. We know, from experience, that long-term plans and distal intentions are often not very effective when they concern changes in habitual behaviors or when they are up against addictions. The important point is that there is an effect from reason-giving all the way to changes in behavior across a wide range of real-life situations. This supports the claim that reason-giving and intending are causally efficacious (ibid.: 260).[17] But it also casts serious doubt on the claim that all reason explanations are based on mere self-interpretation or rationalization. When subjects are given explicit reasons for a certain type of behavior, it is rather unlikely that they will engage in self-interpretation, simply because they do not have to interpret their own actions. If the reasons have been made explicit, it will be relatively easy to remember them, and so it will be relatively easy to give them in reason explanations. It will, in any case, be a lot easier to give one's reasons than in cases where there were no salient reasons (position effect); where reasons and intentions were not stored in memory (highly routine or insignificant actions); or where reasons and intentions are pathologically inaccessible (split-brain cases).

To summarize, the empirical evidence on the confabulation of reason explanations shows that judgments about our own agency are subject to biases and that reason explanations are in some cases based on mere self-interpretation. But there is good reason to resist generalization, and we should not infer the confabulation of intentions from the confabulation of reasons. In any case, the evidence is compatible with RMC, and it does not provide any direct support for the model of AMC.

10. RMC: Arguments, evidence, and conclusions

I have argued that neither the Libet experiment nor the empirical evidence on confabulation provides direct support for the model of AMC, and we have seen that the model of RMC can accommodate paradigm examples of automatisms and illusions of control. Does this mean that the model of RMC is as good as the model of AMC in explaining automatisms and illusions of

[17] This claim is based on an interventionist model of causation, which is a standard model for causal inferences in the empirical sciences (Woodward 2003, for instance).

control? The model of AMC posits one mechanism of behavior causation. The model of RMC assumes that there are two causal pathways that can lead to behavior output. Given this, one might think that the model of RMC is less parsimonious, if not *ad hoc*. But note, first of all, that the model of RMC stipulates only two distinct *input* pathways to the motor control system. It seems quite clear, in fact, that it stipulates also only one mechanism of behavior causation: the sub-personal comparator mechanism of motor control. Moreover, this is a case where the empirical evidence silences the theoretical virtue of parsimony. Empirical evidence suggests that motor and pre-motor areas receive inputs from two distinct pathways (Goldberg 1985, Jahanshahi & Frith 1998, and Haggard 2008, for instance). This is consistent with findings from brain stimulation experiments, which show that one can bring about coordinated movements without generating conscious intentions, and it supports the assumption that inputs from two distinct pathways into the motor control system can lead to the formation of motor commands (and the performance of actions). We can conclude, then, that the model of RMC is at least as good as the model of AMC in explaining automatisms and illusions of control.[18]

There are, however, good reasons to prefer the model of RMC. As noted earlier, most experiments in psychology and cognitive neuroscience suggest that conscious intentions are causally efficacious. It will be worthwhile to elaborate on this. Most experiments feature more than one condition, and at the beginning of an experiment subjects usually agree voluntarily to follow the instruction for the condition they have been selected for. Presumably, this results in the conscious formation of the relevant intentions, which is then usually followed by the performance of matching actions. This supports important counterfactual claims. Consider a randomly selected subject S who takes part in a well-designed experiment with two conditions (the experimental and control condition). Suppose that S is randomly assigned to the experimental condition. Presumably, S will follow the instructions for this condition (*ceteris paribus*), and had S been assigned to the control condition, S would act differently and in accordance with the instructions for this condition (*ceteris paribus*). It seems clear that something along those lines holds for the vast majority of subjects in most experiments. This reflection on the experimental method supports not only the claim that conscious intentions are frequently followed by matching action. But it supports counterfactual claims concerning the co-variation of

[18] It seems, even, that the model of RMC provides a *better* explanation, because it provides a mechanism that shows how automatisms and illusions of control are generated. But the explanatory power of RMC derives to a large extent from the explanatory power of the comparator model. This model of motor control appears to be compatible with the model of AMC, which is why I conclude only that the model of RMC is at least as good.

conscious intentions and matching actions. And this, in turn, supports the claim that conscious intentions are causally efficacious, because it supports the claim that changes in an agent's conscious intentions tend to bring about matching changes in the agent's actions.[19]

Further, it seems clear that something similar holds for a vast amount of everyday actions as well. One example will suffice to illustrate and substantiate this point. Suppose that you are at a friend's house and that you are offered something to drink: tea or coffee, perhaps? You ask for a cup of coffee and your friend heads off to the kitchen, presumably with the intention to get you some coffee. After a couple of minutes she reappears with your coffee. What if you had asked for tea? Common experience and knowledge suggests that your friend would have intended and acted in accord with your request (*ceteris paribus* and with the relevant background conditions in place). Again, it seems clear that something along those lines holds for a vast amount of everyday choices and actions. This provides further support for the thesis that changes in conscious intentions tend to bring about matching changes in behavior.[20]

But there is also plenty of empirical evidence in support of RMC. The mentioned meta-analysis of 47 studies on the efficacy of intentions (Webb & Sheeran 2006) shows that interventions on intentions by means of reason-giving tend to bring about matching changes in intentions and actions. In order "to ensure that changes in intention were responsible for the impact of the interventions on behavior", Webb and Sheraan conducted a mediation analysis from 15 studies where the correlation between intention and behavior could be retrieved. This analysis confirmed the mediating role of intentions (ibid.: 256). This shows that changes in intentions are not mere by-products or epiphenomena—as suggested by the model of AMC—, and it confirms the thesis that intentions are efficacious in the causation of behavior. Another set of evidence stems from research on the efficacy of implementation intentions ("If the circumstances C arise, then I will perform the action A"). A meta-analysis of 94 studies has shown that the formation of implementation intentions has a medium-to-strong effect on subsequent performance and goal-attainment (Gollwitzer & Sheeran 2006). Again, this supports the claim that conscious prior intentions are causally efficacious in the guidance of action, because it shows that changes in conscious (implementation) intentions tend to bring about matching changes in behavior.

[19] Again, this inference is based on an interventionist theory of causal explanation. See note 17.
[20] Note that this argument makes no appeal to the first-person experience of agency. Rather, the epistemic reasons to believe in the relevant counterfactual claims concerning the co-variation of conscious intentions and actions stem from extensive observational and inductive knowledge about intentional action and interaction—knowledge that has been acquired in many years of interacting with the world and with other agents.

It might be objected here that this evidence supports only the claim that *distal* intentions are causally efficacious, whereas Wegner's challenge concerns the efficacy of *proximal* intentions. There are a number of points to note here. First, it is correct that most of the mentioned empirical evidence concerns distal intentions, and the evidence on implementation intentions shows also that the execution of distal if-then intentions can be automatic—it shows that the execution of distal intentions does not require conscious proximal intentions (Gollwitzer & Sheeran 2006). Nevertheless, nothing in the evidence suggests that intentions can be efficacious only if they are distal. In fact, in some of the experiments on implementation intentions, the intentions are executed with so little delay that they appear to be efficacious as proximal rather than distal intentions (Webb & Sheeran 2004, for instance). Second, the mentioned considerations on the efficacy of instructions provide some support for the claim that proximal intentions are efficacious. In many experiments, instructions are given right before the task and sometimes also between or during tasks. Presumably, this leads to the formation of proximal intentions which are executed with little or no delay in accordance with the instructions. Third, common experience suggests that we can change our mind and retract a previously formed intention at the last minute, as it were. In such cases, we abandon a prior intention by forming an opposing proximal intention. But usually we do this in accordance with some of our other desires, plans, or commitments. The assumption of this kind of "veto control" is compatible with the results of the Libet experiment (see Libet 1985), and more recent research has identified distinct neural correlates of this ability to consciously inhibit previously planned actions (Brass & Haggard 2007).

All in all, we can conclude that there is ample reason to favor the model of RMC over the model of AMC, and I shall close now with two further remarks concerning the scope of the presented argument. First, the main concern has been to defend the assumption that conscious intentions are causally efficacious in the initiation and guidance of actions. It should have become clear, however, that some of the arguments and some of the evidence support also the claim that conscious reasoning and reason-giving can be causally efficacious in the guidance of behavior. Second, one might think that my response to Wegner's challenge neglects a central question—namely, the question of whether or not intentions are ever causally efficacious *in virtue of being conscious*.

Recall what the fundamental disagreement between AMC and RMC consists in. According to AMC, the real causes of our actions are inaccessible to consciousness. RMC holds that the real causes of many actions are *accessible* to consciousness. In addition, we may ask whether or not

being consciously *accessed* plays ever a causal role in the initiation and guidance of behavior. This is an important and difficult question. But it is important to note that it is a *further* question. Empirical evidence on automatic goal activation suggests that the initiation and guidance of some goal-directed actions can be unconscious (Chartrand & Bargh 1996, Bargh & Chartrand 1999, Hassin et al. 2005). This shows that consciousness is not necessary for the initiation and guidance of *some* goal-directed actions. If one assumes that these actions are initiated and guided by unconscious *intentions*, then one may infer that not all intentions need to be conscious in order to cause goal-directed actions. However, the evidence does not show that the initiation and guidance of *all* actions could be unconscious. The masked priming of complex semantic contents appears to be impossible (Baars 2002), and empirical evidence suggests that consciousness is required for the integration and strategic use of certain types of information (Dehaene & Naccache 2001); for certain types of conflict resolution (Morsella 2005); and for the planning of future actions by means of mental time travel (Suddendorf & Corballis 2007 and Baumeister & Masicampo 2010). This, of course, does not settle the question concerning the role of consciousness—a task that would require more than another article. But it suggests that the formation of intentions with certain complex contents cannot be unconscious and that consciousness plays an important and necessary role in the planning, initiation, and guidance of some actions—it suggests, in other words, that at least some conscious intentions are causally efficacious in virtue of being conscious.

References

Aarts, H., Custers, R., & Wegner, D. M. (2005). On the inference of personal authorship: Enhancing experienced agency by priming effect information. *Consciousness and Cognition*, 14, 439–458.

Ajzen, I. (1985). From intentions to action: A theory of planned behavior. In J. Kuhl & J. Beckmann (Eds.). *Action-control: From cognition to behavior*. New York: Springer, pp. 11–39.

Andersen, R. A., & Buneo, C. A. (2002). Intentional maps in posterior parietal cortex. *Annual Review of Neuroscience*, 25, 189–220.

Andersen, R. A., & Cui, H. (2009). Intention, action planning, and decision making in parietal-frontal circuits. *Neuron*, 63, 568–583.

Anscombe, G.E.M. (1957). *Intention*. Oxford: Basil Blackwell.

Archibald, S. J., Mateer, C.A. & Kerns, K. A. (2001). Utilization behavior: Clinical manifestations and neurological mechanisms, *Neuropsychology Review*, 11, 117–130.

Austin, J. J., & Vancouver, J. B. (1996). Goal constructs in psychology: Structure, process, and content. *Psychological Bulletin*, 120, 338–375.

Baars, B. J. (2002). The conscious access hypothesis: Origins and recent evidence. *Trends in Cognitive Sciences*, 6, 47–52.

Baer, J., Kaufman, J. C., & Baumeister, R. F. (Eds.) (2008). *Are we free? Psychology and free will*. Oxford: Oxford University Press.

Ball, T., Schreiber, A., Feige, B., Wagner, M., Lücking, C. M., & Kristeva-Feige, R. (1999). The role of higher-order motor areas in voluntary movement as revealed by high-resolution EEG and fMRI. *Neuroimage*, 10, 682–694.
Bargh, J. A. (1994). The four horsemen of automaticity: Awareness, efficiency, intention, and control in social cognition. In R. S. Wyer, Jr., & T. K. Srull (Eds.). *Handbook of social cognition* (2nd ed.). Hillsdale: Erlbaum, pp. 1–40.
Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist*, 54, 462-479.
Baumeister, R. F., & Masicampo, E. J. (2010). Conscious thought is for facilitating social and cultural interactions: How mental simulations serve the animal–culture interface. *Psychological Review*, 117, 945–971.
Bayne, T. (2006). Phenomenology and the feeling of doing: Wegner on the conscious will. In S. Pockett et al. (Eds.), pp. 169–186.
Bayne, T., & Pacherie E. (2007) Narrators and comparators: The architecture of agentive self-awareness. *Synthese*, 159, 475–491.
Berti, A., Bottini, G., Gandola, M., Pia, L., Smania, N., Stracciari, A., Castiglioni, I., Vallar, G., & Paulesu, E. (2005). Shared cortical anatomy for motor awareness and motor control. *Science*, 309, 488–491.
Brass, M., & Haggard, P. (2007). To do or not to do: The neural signature of self-control. *Journal of Neuroscience*, 27, 9141–9145.
Bratman, M. E. (1987). *Intentions, plans, and practical reason*. Cambridge, MA: Harvard University Press.
Chalmers, D. (1996). *The conscious mind: In search of a fundamental theory*. Oxford: Oxford University Press.
Chartrand, T. L., & Bargh, J. A. (1996). Automatic activation of impression formation and memorization goals: Nonconscious goal priming reproduces effects of explicit task instructions. *Journal of Personality and Social Psychology*, 71, 464–478.
Christensen, M. S., Lundbye-Jensen, J., Geertsen, S. S., Petersen, T. H., Paulson, O. B., & Nielsen, J. B. (2007). Premotor cortex modulates somatosensory cortex during voluntary movements without proprioceptive feedback. *Nature Neuroscience*, 10, 417–419.
Crane, T. (1995). The mental causation debate. *Proceedings of the Aristotelian Society*, Suppl. Vol. LXIX, 211–236.
Damasio, A. (2010). *Self comes to mind: Constructing the conscious brain*. New York: Random House.
D'Andrade, R. (1987). A folk model of the mind. In D. Holland & N. Quinn (Eds.). *Cultural models in language and thought*. Cambridge: Cambridge University Press, pp. 112–148.
Davidson, D. (1963). Actions, reasons, and causes. *Journal of Philosophy*, 60, 685–700.
Davies, P. S. (2009). *Subjects of the world: Darwin's rhetoric and the study of agency in nature*. Chicago: University of Chicago Press.
Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition*, 79, 1–37.
Dennett, D. (2003). *Freedom evolves*, London: Penguin Books.
— (2008). Some observations on the psychology of thinking about free will. In J. Baer et al. (Eds.), pp. 248–259.
Desmurget, M., Reilly, K. T., Richard, N., Szathmari, A., Mottolese, C., & Sirigu, A. (2009). Movement intention after parietal cortex stimulation in humans. *Science*, 324, 811–813.
Desmurget, M., & Sirigu, A. (2009). A parietal-premotor network for movement intention and motor awareness. *Trends in Cognitive Sciences*, 13, 411–419.
Elton, M. (2000). The personal/sub-personal distinction: An introduction. *Philosophical Explorations*, 3, 1–5.
Enç, B. (2003). *How we act: Causes, reasons, and intentions*. Oxford: Oxford University Press.
Engbert, K., Wohlschläger, A., & Haggard, P. (2008). Who is causing what? The sense of agency is relational and efferent-triggered. *Cognition*, 107, 693–704.
Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, 59, 255–78.

Farrer, C., Franck, N., Georgieff, N., Frith, C. D., Decety, J., & Jeannerod, M. (2003). Modulating the experience of agency: A positron emission tomography study. *Neuroimage*, 18, 324–333.
Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention and behavior: An introduction to theory and research.* Reading, MA: Addison Wesley.
Fried, I., Katz, A. McCarthy, G., Sass, K.J., Williamson, P., Spencer, S. S., & Spencer, D. D. (1991). Functional organisation of human supplementary motor cortex studies by electrical stimulation. *Journal of Neuroscience*, 11, 3656–3666.
Frith, C. D. (2007). *Making up the mind: How the brain creates our mental world.* Oxford: Blackwell.
Frith, C. D., Blakemore, S., & Wolpert, D. M. (2000). Abnormalities in the awareness and control of action. *Philosophical Transactions of the Royal Society* B, 355, 1771–1788.
Gallagher, S. (2006). Where's the action? Epiphenomenalism and the problem of free will. In S. Pockett, et al. (Eds.), pp. 109–124.
—(2007). The natural philosophy of action. *Philosophy Compass*, 2, 1–11.
Gazzaniga, M. S. (1998). *The mind's past.* Berkeley: University of California Press.
Gazzaniga, M. S., & LeDoux, J. E. (1978). *The integrated mind.* New York: Plenum.
Gibbons, T. (2006). Mental causation without downward causation. *Philosophical Review*, 115, 79–103.
Goldberg, G. (1985). Supplementary motor area structure and function: Review and hypotheses. *Behavioral and Brain Sciences*, 8, 567–616.
Goldberg, G., & Bloom, K. K. (1990). The alien hand sign: Localization, lateralization and recovery. *American Journal of Physical Medicine and Rehabilitation*, 69, 228–238.
Goldberg, G., Mayer, N. H., & Toglia, J. U. (1981). Medial frontal cortex and the alien hand sign. *Archives of Neurology*, 38, 683–686.
Goldman, A. (1970). *A theory of human action.* New York: Prentice-Hall.
Gollwitzer, P. M. (1993). Goal achievement: The role of intentions. *European Review of Social Psychology*, 4, 141–185.
Gollwitzer, P. M., & Sheeran, P. (2006). Implementation intentions and goal achievement: A meta-analysis of effects and processes. *Advances in Experimental Social Psychology*, 38, 69–119.
Greenwood, J. D. (ed.) (1991). *The future of folk psychology: Intentionality and cognitive science*. Cambridge: Cambridge University Press.
Haggard, P. (2005). Conscious intention and motor cognition. *Trends in Cognitive Sciences*, 9, 290–295.
— (2008). Human volition: Towards a neuroscience of will. *Nature Reviews: Neuroscience*, 9, 934–946.
Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108, 814–834.
Hassin, R. R., Uleman, J. S., & Bargh, J. A. (Eds.) (2005). *The new unconscious*. Oxford: Oxford University Press.
Haynes, J-D., Sakai, K., Rees, G., Gilbert, S., Frith, C. D., & Passingham, R.E. (2007). Reading hidden intentions in the human brain. *Current Biology*, 17, 323–328.
Heckhausen, H. (1991). *Motivation and action*. Berlin: Springer-Verlag.
Horgan, T., Tienson, J., & Graham, G. (2003). The phenomenology of first-person agency. In S. Walter & H.D. Heckmann (Eds.) *Physicalism and mental causation: The metaphysics of mind and action.* Exeter: Imprint Academic, pp. 323–40.
Horgan, T., & Woodward, J. (1985). Folk psychology is here to stay. *Philosophical Review*, 94, 197–225.
Jahanshahi, M., & Frith, C. D. (1998). Willed action and its impairments. *Cognitive Neuropsychology*, 15, 483–533.
Jansson, E., Wilson, A. D., Williams, J. H., & Mon-Williams, M. (2007). Methodological problems undermine tests of the ideo-motor conjecture. *Experimental Brain Research*, 182, 549–558.
Keller, I., & Heckhausen, H. (1990). Readiness potentials preceding spontaneous motor acts: Voluntary vs. involuntary control. *Electroencephalography and Clinical Neurophysiology*, 76, 351–61.
Keren, G., & Schul, Y. (2009). Two is not always better than one: A critical evaluation of two-system theories. *Perspectives on Psychological Science*, 4, 533–550.
Kim, J. (1998). *Mind in a physical world*. Cambridge, MA: MIT Press.
Kirsch, I., & Lynn, S. J. (1998) Social-cognitive alternatives to dissociation theories of hypnotic involuntariness. *Review of General Psychology*, 2, 66–80.

Knuf, L., Aschersleben, G. & Prinz, W. (2001). An analysis of ideomotor action. *Journal of Experimental Psychology*, 130, 779-798.
Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences*, 8, 529–66.
Lhermitte, F. (1983). Utilization behavior and its relation to lesions of the frontal lobes. *Brain*, 106, 237–255.
Locke, E. A., & Latham, G. P. (1990). *A theory of goal setting and task performance*. Englewood Cliffs, NJ: Prentice-Hall.
Malle, F. B. (1999). How people explain behavior: A new theoretical framework. *Personality and Social Psychology Review*, 3, 23–48.
— (2004). *How the mind explains behavior: Folk explanations, meaning, and social interaction*, Cambridge, MA: MIT Press.
— (2006). Of windmills and straw men: Folk assumptions of mind and action. In S. Pockett, et al. (Eds.), pp. 207–231.
Marcel, A. (2003). The sense of agency: Awareness and ownership of action. In J. Roessler & N. Eilan (Eds.). *Agency and self-awareness*. Oxford: Oxford University Press, pp. 48–93.
Melden, A. I. (1961). *Free action*. London: Routledge and Kegan Paul.
Mele, A. (2009). *Efficacious intentions*. Oxford: Oxford University Press.
Miller, E. K., & Cohen, J. D. (2001) An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, 24, 167–202.
Moore, J. W., Wegner, D. M., & Haggard, P. (2009) Modulating the sense of agency with external cues. *Consciousness and Cognition*, 18, 1056–1064.
Morsella, E. (2005). The function of phenomenal states: Supramodular interaction theory. *Psychological Review*, 112, 1000–1021.
Nahmias, E. (2002). When consciousness matters: A critical review of Daniel Wegner's The illusion of conscious will. *Philosophical Psychology*, 15, 527–541.
Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84, 231–259.
Osman, M. (2005). An evaluation of dual-process theories of reasoning. *Psychonomic Bulletin & Review*, 11, 988–1010.
Pacherie, E. (2007). The sense of control and the sense of agency. *Psyche*, 13, 1–30.
Pacherie, E., & Haggard P. (2010) What are intentions? In W. Sinnott-Armstrong & L. Nadel (Eds.) *Conscious will and responsibility: A tribute to Benjamin Libet*. Oxford: Oxford University Press.
Penfield, W. (1975). *The mystery of mind*. Princeton: Princeton University Press.
Pockett, S., Banks, W.P., & Gallagher, S. (Eds.) (2006). *Does consciousness cause behavior?* Cambridge, MA: MIT Press.
Ross, D., Spurrett, D., Kincaid, H., & Stephens, G. L. (2007). *Distributed cognition and the will: Individual volition and social context*. Cambridge, MA: MIT Press.
Rowe J. B., Toni I., Josephs O., Frackowiak R. S., & Passingham R. E. (2000). The prefrontal cortex: Response selection or maintenance within working memory? *Science*, 288, 1656–1660.
Schlosser, M. E. (*forthcoming*). Free will and the unconscious precursors of choice. *Philosophical Psychology*.
Scott, S. H. (2004). Optimal feedback control and the neural basis of volitional motor control. *Nature Reviews: Neuroscience*, 5, 532–546.
Sehon, S. (2005). *Teleological realism: Mind, agency, and explanation*. Cambridge, MA: MIT Press.
Shallice, T., Burgess, P. W., Schon, F., & Baxter, D. M. (1989) The origins of utilization behaviour. *Brain*, 112, 1587–1598.
Shoemaker, S. (2007). *Physical realization*. Oxford: Oxford University Press.
Soon, C. S., Brass, M., Heinze H. J., & Haynes J. D. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*, 11, 543–545.
Suddendorf, T., & Corballis, M. C. (2007). The evolution of foresight: What is mental time travel, and is it unique to humans? *Behavioral and Brain Sciences*, 30, 299–351.

Sumner P., Nachev P., Morris P., Peters A.M., Jackson S. R., Kennard C., & Husain M. (2007). Human medial frontal cortex mediates unconscious inhibition of voluntary action. *Neuron*, 54, 697–711.
Synofzik, M., Vosgerau, G., & Newen, A. (2008) Beyond the comparator model: A multifactorial two-step account of agency. *Consciousness and Cognition*, 17, 219–239.
Triandis, H. C., (1977). *Interpersonal behavior*. Monterey, CA: Brooks/Cole.
Wakefield, J., & Dreyfus, H. (1991). Intentionality and the phenomenology of action. In E. Lepore and R. van Gulick (Eds.). *John Searle and his critics*. Oxford: Blackwell, pp. 259–70.
Webb, T.L., & Sheeran, P. (2004). Identifying good opportunities to act: Implementation intentions and cue discrimination. *European Journal of Social Psychology*, 34, 407–419.
— (2006). Does changing behavioral intentions engender behavior change? A meta-analysis of the experimental evidence. *Psychological Bulletin*, 132, 249–268.
Wegner, D. M. (2002). *The illusion of conscious will*. Cambridge: MIT Press.
— (2004). Précis of the illusion of conscious will. *Behavioral and Brain Sciences*, 27, 649–92.
— (2005). Who is the controller of controlled processes? In R. Hassin et al. (Eds.), pp. 19–36.
— (2008). Self is magic. In J. Baer et al. (Eds.), pp. 226–247.
Wegner, D. M., Sparrow, B., & Winerman, L. (2004). Vicarious agency: Experiencing control over the movements of others. *Journal of Personality and Social Psychology*, 86, 838–48.
Wegner, D. M., & Wheatley, T. P. (1999). Apparent mental causation: Sources of the experience of will. *American Psychologist*, 54, 480–92.
Wilson, T. (2002). *Strangers to ourselves: Discovering the adaptive unconscious*. Cambridge, MA: Harvard University Press.
Wolpert, D. M, & Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks*, 11, 1317–1329.
Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford: Oxford University Press.