Conscious will, reason-responsiveness, and moral responsibility

Markus E. Schlosser

Forthcoming in the *Journal of Ethics*

This is the manuscript (which may differ from the final print version in minor details)

Abstract: Empirical evidence challenges many of the assumptions that underlie traditional philosophical and commonsense conceptions of human agency. It has been suggested that this evidence threatens also to undermine free will and moral responsibility. In this paper, I will focus on the purported threat to moral responsibility. The evidence challenges assumptions concerning the ability to exercise conscious control and to act for reasons. This raises an apparent challenge to moral responsibility as these abilities appear to be necessary for morally responsible agency. I will argue that this challenge collapses once the underlying conditions on moral responsibility are specified in sufficient detail. I will argue, in other words, that the empirical evidence does not support a challenge to the assumption that we are, in general, morally responsible agents. In the final section, I will suggest that empirical research on human agency is nevertheless relevant to various questions about moral responsibility.

Keywords: Moral responsibility, conscious control, automaticity, reason-responsiveness, situationism, philosophy of cognitive science

1. Introduction

Over the past few decades, it has often been argued that common conceptions of human agency are untenable in the light of empirical evidence from psychology and cognitive neuroscience. It has also often been suggested that this evidence challenges the assumption that we are free and morally responsible agents. Intuitively, it is not difficult to see why the evidence seems to threaten free will and moral responsibility. This is obviously an important issue, and a great amount of attention has been devoted to the empirical challenge to free will. In contrast, the challenge to moral responsibility has not received much detailed analysis. I will focus here on this empirical challenge to moral responsibility, and I will argue that the challenge collapses once the underlying necessary conditions on moral responsibility are specified in sufficient detail.

First, I will provide summary accounts of the empirical challenges (section 2). Then I will explain why we can set the issue of free will aside and why it makes sense to focus on the empirical challenge to moral responsibility (section 3). In the main part, I will

distinguish between two lines of argument (section 4). According to the first, the empirical evidence undermines moral responsibility, because it shows that conscious will is an illusion. According to the second, the evidence undermines moral responsibility, because it shows that we are not responsive to the right reasons. Both arguments are general and largely theory-independent insofar as they are based on widely accepted conditions on moral responsibility, and insofar as they do not depend on the claims and commitments of particular theories of moral responsibility. I will argue that neither of the two arguments withstands closer inspection (sections 5 and 6). We will see, moreover, that there is empirical evidence in support of the assumption that we are able to exercise conscious control and that we are reason-responsive. We will see, in other words, that there is empirical evidence *in support* of the assumption that we are, in general, morally responsible agents. In section 7, I will say more about which theories of moral responsibility are covered by this general response, and I will explain why the response applies to criminal responsibility as well. Further, I will turn to theories that are not covered by the main response, and I will argue that the empirical evidence does not raise a challenge to these theories either. In the final part (section 8), I will suggest that empirical evidence on human agency can nevertheless be relevant to various practical questions on moral responsibility and that it provides valuable theoretical knowledge concerning the mechanisms that underlie morally responsible agency. It will be admitted, however, that some of the evidence shows that we—or most of us—are less reason-responsive than we would like to think.

2. Empirical challenges: A brief summary

In order to introduce the challenge to free will and moral responsibility, let me begin with brief summary accounts of the main conclusions that have been drawn from the evidence that is most relevant to the issue at hand. Representative examples of the evidence for these claims will be discussed in due course. We can distinguish between the following four empirical claims (EC):

> (EC1) *Actions are not consciously initiated*. Empirical evidence supports the view that our actions are initiated by unconscious processes rather than conscious

intentions: some actions are preceded by conscious intentions, but they are not initiated by them (Libet 1985; Wegner 2002; Soon et al. 2008).

(EC2) *Most actions are not consciously controlled*. Empirical evidence supports the view that most actions are *automatic*: most actions are controlled by stimulus-driven mechanisms that operate outside conscious awareness and without conscious guidance (Kihlstrom 1987; Bargh and Chartrand 1999; Custers and Aarts 2010).[1]

(EC3) *Reason explanations are not causal explanations*. Empirical evidence supports the view that commonsense reason explanations of actions are based on interpretation or post-hoc rationalization rather than on access to the causes of our actions: what we take to be our reasons are interpretive or theoretical constructs rather than the real causes of our actions (Nisbett and Wilson 1977; Gazzaniga and LeDoux 1978; Wegner 2002; Johansson et al. 2006).

(EC4) *Character traits do not explain behavior*. Empirical evidence supports the view that behavior is to be explained primarily in terms of situational features, which suggests that commonsense explanations of behavior in terms of character traits (such as honesty, kindness, and courage) are systematically mistaken (Ross and Nisbett 1991; Harman 1999; Doris 2002).

EC1 – 4 challenge common assumptions about human agency: assumptions that underlie traditional philosophical and commonsense conceptions of human agency. Intuitively, they also challenge the assumption that we have free will and that we are, in general, justified in holding each other morally responsible—they challenge free will and moral responsibility, for short. EC1 is most frequently mentioned in this respect, but EC2 – 4 are also relevant here, as I will explain in due course (see, most prominently, Libet 1999; Wegner 2002; see also Doris 2002; Nelkin 2005; Roskies 2006; Pockett et al. 2006; Baer et al. 2008; Sinnott-Armstrong and Nadel 2011).

[1] Obviously, EC1 and EC2 are closely related and there is a potential overlap as the conscious control of action can plausibly be taken to involve the conscious initiation of action. Nevertheless, the two claims can be distinguished from each other insofar as they are based on distinct strands of empirical research.

3. Free will and moral responsibility

Anyone who is interested in the traditional philosophical and in the empirical debate about free will and moral responsibility will have noticed that these two debates are largely disconnected. The main question in the traditional philosophical debate has been whether or not free will and moral responsibility are compatible with determinism. In the empirical debate, the main issue has not been *whether* our choices and actions are causally determined, but *how* they are caused. In particular, there has been much debate about the role of consciousness in the control of action and about the influence of situational factors—issues that have not been at the forefront in the philosophical debate.

There has also been much philosophical debate about the relationship between free will and moral responsibility. According to a traditional view, free will is necessary for moral responsibility insofar as the ability to do otherwise is necessary for moral responsibility. This view has been challenged more recently. In particular, both free will and moral responsibility have been disassociated from the ability to do otherwise, and it has been argued that they are independent in the sense that moral responsibility may be compatible with determinism even if free will is not (see Frankfurt 1969; Fischer 1994; Fischer and Ravizza 1998; Clarke 2003).

In contrast, in the empirical debate the relationship between free will and moral responsibility has not been discussed, and it has remained unclear whether the evidence is supposed to challenge free will *and* moral responsibility, or whether it is supposed to challenge moral responsibility *because* it challenges free will. However, in light of the evidence, the former interpretation is clearly more plausible. To see this, note that none of the empirical claims (EC1 – 4) challenges free will and moral responsibility directly. Rather, they challenge certain features of human agency that are presumed to be necessary for free will and moral responsibility. In particular, EC1 and EC2 challenge commonsense assumptions about conscious control, EC3 challenges assumptions about the ability to act for reasons, and EC4 challenges an assumption about the role of character traits. Given this, the empirical claims seem to challenge free will insofar as the mentioned abilities and features of human agency are necessary for free will. The same holds for moral responsibility. The empirical claims seem to challenge moral responsibility insofar as the mentioned abilities and features of human agency are necessary for moral responsibility.

This means that we can set the empirical challenge to free will aside and focus on the challenge to moral responsibility, as the latter does not depend on the former. Moreover, I think that there are good reasons to do just that.

First, it is sometimes remarked that we care about free will only insofar as it is necessary for moral responsibility. Even if this claim is too strong, as I think, it seems safe to say that we care about moral responsibility at least as much as we care about free will. Despite this, the empirical challenge to moral responsibility has not received much detailed analysis in comparison to the challenge to free will.[2] Second, I suspect that most of the attention has been devoted to the challenge to free will because it has been assumed that one can address the challenge to moral responsibility by addressing the challenge to free will. However, as mentioned, in philosophy the relationship between free will and moral responsibility has been controversial. It may be, then, that the strategy of addressing the challenge to moral responsibility by addressing the challenge to free will is based on false assumptions concerning the relationship between free will and moral responsibility. Given this, it makes sense to investigate whether the empirical evidence raises a challenge to moral responsibility independently of whether or not it raises a challenge to free will. Third, the nature of free will has itself been controversial. Recently, it has become common to bypass this issue by defining "free will" as whatever kind of agency or control that is required for moral responsibility (see, for instance, Pereboom 2001; Vargas 2007; Nelkin 2011). On this view, it makes not only sense to focus on the challenge to moral responsibility. Rather, we have to investigate whether the evidence challenges moral responsibility in order to see whether it challenges free will, because free will is defined in terms of moral responsibility. Given all this, we have more than ample reason to focus on the empirical challenge to moral responsibility.

---

[2] The empirical challenge to moral responsibility has received much *attention*, but not much *detailed analysis*. It has usually been addressed by way of addressing the challenge to free will, and the two challenges have often not been clearly distinguished. See Libet 1999; Wegner 2002; Greene and Cohen 2004; Roskies 2006; and many of the contributions to Pockett et al. 2006; Baer et al. 2008; Sinnott-Armstrong and Nadel 2011. Doris 2002 and Nelkin 2005 focus on the challenge to moral responsibility, but their discussion is restricted to the challenge from EC4.

4. The empirical challenge to moral responsibility

Let us turn, then, to the question of how the empirical challenges might generate a challenge to moral responsibility. My main aim here is to provide a diagnosis that is based on widely accepted assumptions about moral responsibility and that is largely independent from the claims and commitments of particular theories of moral responsibility. This will enable me to provide a general response that is largely theory-independent in this sense. (We will return to this and to theories that are not covered by this strategy in section 7.)

As mentioned, EC1 and EC2 challenge commonsense assumptions about conscious control, EC3 challenges assumptions about our ability to act for reasons, and EC4 challenges an assumption about the role of character traits. It is widely agreed that the ability to act for reasons is necessary for moral responsibility, and it seems only plausible to assume that moral responsibility requires the ability to exercise conscious control over one's actions. Given this, we can identify, to a first approximation, two conditions on moral responsibility that would seem to generate an empirical challenge on the basis of EC1 – 3:

(CC) Moral responsibility requires conscious control.

(AR) Moral responsibility requires the ability to act for reasons.

EC4, which is commonly known as situationism, would seem to raise an obvious challenge to character-based accounts of moral responsibility. Such accounts are rather uncommon, as it seems clear that acting in accord with character traits is neither necessary nor sufficient for moral responsibility (see Doris 2002, 128–132; Nelkin 2005). Moreover, traditional character-based accounts are motivated by the idea that an agent can be held responsible only if the behavior in question can be attributed to an *enduring feature* of the agent (Hume 1960/1740, 411). This does not require that the agent possesses character traits in a robust and traditional philosophical sense (such as honesty, kindness, or courage). It requires only the possession of enduring behavioral dispositions, and it is generally acknowledged that this is compatible with the evidence for situationism (see Doris 2002 and Nelkin 2005). Given this, we can safely ignore the challenge from EC4 to character-based accounts. However, the evidence for EC4 may also be taken to challenge the assumption that we are responsive to the relevant reasons, and thereby to generate a challenge on the basis of AR, as I will explain below (section 6.2).

CC and AR are plausible conditions, but they are also extremely vague. In order to assess whether CC and AR do indeed generate a challenge to moral responsibility (in conjunction with EC1 – 4), it is necessary to specify them in more detail. In the following two sections, I will propose plausible specifications, and I will argue that sufficiently refined versions of CC and AR fail to support a general empirical challenge to moral responsibility.

5. Moral responsibility and conscious control

Does moral responsibility require conscious control? All depends here on how this condition is understood more specifically, and there is a host of issues that one may wish to have clarified. At what point must the agent have or exercise conscious control? Over which action? Does the agent have to be aware of performing an action under a certain description? What is the underlying notion of control? What kind of consciousness is required? And so on.

If we gave answers to all these questions, we would obtain a highly specific condition. This would, presumably, commit us to specific views on moral responsibility, and it would exclude others—something that I would rather avoid. But we can make progress here if we turn the approach around: instead of reflecting on CC itself, we can begin by asking how EC1 and EC2 are supposed to undermine moral responsibility by undermining the condition on conscious control in CC. Before we turn to that, let me add some general qualifications.

First, I will assume that the required kind of consciousness is *access* consciousness, not phenomenal consciousness: to require conscious control is to require that conscious access to the relevant mental representations plays the right role in the control of one's actions. This is a plausible qualification that does not beg any questions, as the empirical evidence for EC1 and EC2 concerns only the role of reportable (that is, accessible) representations. Second, one might construe CC as a condition on being a morally responsible *agent*. This would seem to be uncontroversial, but also uninteresting. The interesting question is how we can turn CC into a more specific and plausible condition on moral responsibility for particular actions (and omissions). Third, if we construe CC as a condition on moral responsibility for actions, then we have to restrict it to *direct* (or

underived) responsibility. To see why, consider the case of drunk driving, which is a standard example of indirect (or derived) responsibility. One may be held morally responsible for the consequences of drunk driving even if one does, at the time, not have sufficient conscious control over one's actions. The reason that is usually given here is, roughly, that the responsibility derives from earlier actions over which one had conscious control and which were known, at that time, to incur considerable risks. The responsibility for drinking is direct, and it grounds the indirect responsibility for the consequences of drunk driving.[3] Examples of this kind show that it would clearly be implausible to impose a condition along the lines of CC on cases of indirect responsibility. Accordingly, the discussion will be restricted to direct moral responsibility (and I shall omit this qualification from now on).

As indicated, I propose to make CC more specific not by reflecting further on this condition, but by turning to the empirical challenges. How do we have to specify CC so that EC1 and EC2 raise a challenge to moral responsibility? The most obvious and the most important point to note here is that the empirical evidence in support of EC1 and EC2 concerns exclusively the role of conscious states that either proximally precede or accompany the execution of overt behavior—it concerns, as I shall say, the role of consciousness in the *proximal* and *online* control of action. Taking this into account, one may specify CC as follows:

> An agent is morally responsible for an action (or omission) only if the agent exercises proximal and online conscious control over it.

However, this condition is clearly too strong. Obvious counterexamples are provided by habitual actions and by cases that involve unintentional omissions. Consider the following. Suppose that you were driving on the motorway and that you forgot to indicate before you pulled over to overtake another car. Suppose that this resulted in an accident. It seems highly implausible to suggest that you are responsible for causing the accident only if you exercised proximal and online conscious control over not indicating. Most would agree, here, that you may well be fully responsible even if you did not exercise proximal and

[3] Cases that involve severe addiction are more complicated, but we can ignore this issue here.

online conscious control over not indicating (depending on further details of the case). Note that this case is complex, in the sense that it is neither a pure action nor a pure omission. It involves an action, an omission, and an outcome (the accident). We assume that the outcome is to be explained in terms of both the action and the omission. To say that you may be fully responsible is to say, then, that you may be fully responsible for the outcome because you may be responsible for both the action and the omission.

It seems, however, that we can turn CC into a condition that can accommodate such cases if we turn it into a claim about the *ability* to exercise proximal and online conscious control:

> (ACC) An agent is morally responsible for performing (or failing to perform) an action only if the agent is able, at the time, to exercise proximal and online conscious control over that action.

Return to the example. We assume that you caused the accident partly because you did not indicate. According to ACC, you are responsible for this omission only if you were able, at the time, to exercise proximal and online conscious control over indicating. For all we know, this condition holds. Everyone, we may assume, who is able to drive a car is able to exercise conscious control over whether or not to indicate (other things, such as the mechanics and electronics of the car, being equal).

With some adjustments, the example suggests also that ACC can accommodate responsibility for habitual actions and omissions. Suppose that you did not indicate because you have the bad habit of not indicating before you overtake. This does not change anything. Given that the omission was habitual, you probably did not exercise proximal and online conscious control over not indicating. Yet, there is no reason to think that you were unable to exercise conscious control over whether or not to indicate, and you may well be fully responsible.[4] The same would seem to hold for the manifestations of a good habit to indicate (when you are supposed to). There is no reason to think that you are unable to

[4] Note the difference between this case and cases of indirect (or derived) responsibility, such as drunk driving. Indirect responsibility is grounded in earlier acts for which one is directly responsible. This cannot be said about the present case, for it is not clear that the omission of not indicating can be traced back to earlier acts. We assume that you have a habit of not indicating, but you may not have *done* anything to acquire this habit.

exercise conscious control over whether or not to indicate, and you may be fully responsible for manifesting the habit (and the consequences thereof). Note that this condition on omissions and habitual actions is only plausible. It is plausible to hold that one can be held responsible for an omission only if one could have done intentionally and consciously what one failed to do, and it is plausible to hold that one can be held responsible for a habitual action (or omission) only if one could have done intentionally and consciously what one did (or failed to do) out of habit.

ACC appears to be a plausible condition and it appears to generate challenges to moral responsibility on the basis of EC1 and EC2. In the following two sections, I will assume ACC for the sake of argument, and I will respond to the challenges from EC1 and EC2 in turn. It is important to note, here, that should ACC be too strong or untenable for other reasons, then so much the worse for those who think that the empirical evidence on conscious control challenges moral responsibility. ACC is the best candidate for establishing a plausible connection between moral responsibility and conscious control—if ACC is untenable, then so are the empirical challenges from EC1 and EC2.

5.1. Conscious control and EC1

According to EC1, consciousness plays no role in the initiation (or proximal control) of action. I will argue now that EC1 is compatible with the control condition in ACC, because the exercise of proximal conscious control can be *conditioned*. In order to explain this, it will be helpful to consider Benjamin Libet's well-known neuroscientific challenge to free will. We will see here that the focus on responsibility enables a response that goes beyond the standard reply to this challenge.

In the Libet experiment, participants were instructed to initiate a certain movement when they felt like doing so. During this, the occurrence of the readiness potential (RP) was recorded. The RP is an unconscious brain potential that was known to precede voluntary movements. Libet found that RPs precede the conscious intentions to move by about 350ms, and he drew two main conclusions from this: movements are not consciously initiated and we do not have free will in the sense we commonly think we do (Libet 1985 and 1999). Both conclusions have been challenged for various reasons and the issue has remained controversial to this day (Pockett et al. 2006; Mele 2009; Sinnott-Armstrong and Nadel

2011; Schlosser 2012a). I will argue here only that the evidence is compatible with the assumption that participants possessed and exercised a kind of proximal conscious control that is sufficient for moral responsibility, and I will set aside the questions of whether this kind of control is the same as genuine conscious initiation and sufficient for free will.

In the main experiment, participants were instructed to initiate a certain movement when the "urge to act" arises (Libet 1985). According to Libet, some participants reported that "during some of the trials a recallable conscious urge to act appeared but was 'aborted' or somehow suppressed". They "waited for another urge to appear" which was then executed as instructed (ibid., 538). Let us call this the *initiation* experiment and let us refer to the mentioned trials as *spontaneous veto* trials. In order to investigate this further, Libet conducted a version of the experiment in which participants were instructed to inhibit the movement when the urge to move arises. Let us call this the *instructed veto* experiment.

We may assume that participants in both experiments formed *distal* intentions at the beginning of the experiment in accord with the instructions: an intention to perform the movement when the urge arises or an intention to inhibit the movement when the urge arises. The reports from the spontaneous veto trials suggest that the former do not determine the execution of the movement *whenever* an urge to move arises. But it seems nevertheless clear that the distal intentions were causally efficacious in the initiation experiment, because all participants did eventually what they intended to do (Zhu 2003, for instance). The same can be said about the instructed veto experiment. Presumably, participants intended to comply with the instruction to inhibit the movement, and that is what they did. In all cases, the distal intentions *conditioned* the proximal control, as I shall say: they conditioned whether the agent initiated or inhibited the movement.

Libet argued that the instructed veto experiment supports the claim that we have "conscious veto control" (Libet 1999). In particular, it seems to support the claim that all participants, including the ones in the initiation experiment, were *able* to consciously inhibit the movement. It has been noted that this argument is highly problematic, as the veto experiment is highly problematic. The veto experiment seems to require something impossible: that one intends to move and intends not to move (Mele 2009, for instance). What has not been noted, however, is that the comparison between the regular initiation and the spontaneous veto trials supports the same conclusion: given that some participants in

the initiation experiment inhibited the movement in response to becoming consciously aware of the urge to move, and given that nothing in the instructions or in the circumstances can account for this, we have reason to think that all participants in the initiation experiment were able to inhibit the movement in response to becoming consciously aware of the urge to move. This alone suffices to satisfy ACC, which requires only an ability to exercise conscious control. However, one might insist, plausibly, that it must also be shown here that the participants *exercised* conscious control, as they paid close attention and formed conscious intentions.[5]

According to a standard response to Libet's challenge, the evidence is at least compatible with the interpretation that participants exercised proximal conscious control (see Zhu 2003; Mele 2009). This response emphasizes the point that motivational states (such as desires and urges) must be distinguished from executive states (such as intentions and decisions), and it suggests that participants first became consciously aware of an urge to move *in response to which* they decided to execute the movement. This interpretation is compatible with the evidence and it suggests that participants *exercised* proximal conscious control—it suggests this especially in light of the assumption that they were able to inhibit the movement in response to becoming consciously aware of the urge to move.

Given what has been said above, we can identify a further role for consciousness in the initiation of the movements, but we also have to qualify the standard reply concerning the agent's exercise of proximal control. As pointed out, participants formed distal intentions that conditioned whether they initiated or inhibited the movement—intentions that conditioned their proximal control. For all we know, the distal intentions were formed consciously. Given this, it seems clear that consciousness played also a distal role in the initiation of the movements. However, this also means that the agent's exercise of proximal control was conditioned by the distal intentions. Here is where the focus on responsibility

[5] I shall assume here for the sake of argument that the conscious events that proximately preceded the movements were conscious *intentions*. I should note, however, that this assumption has also been questioned. For instance, Keller and Heckhausen (1990, 359) suggested that the conscious events in question were the result of the "selective attention" to look for an urge to move, which was, in turn, induced by the artificial setup and the instructions of the experiment. They suggested, in other words, that the conscious events in question were neither intentions nor conscious events that precede ordinary actions.

becomes relevant. One might think that an action is performed with free will only if it is not conditioned in this sense, perhaps because one thinks that free will requires that the agent is the genuine and spontaneous initiator of the action (whatever that means, exactly).[6] In contrast, the exercise of unconditioned conscious control is clearly not required for moral responsibility. To see this, take any two instances of a wrongful act and assume that the initiation of the action was conditioned by a distal intention only in one of the two cases. There is absolutely no reason to think that the agent who executed a distal intention is any less responsible for the action (other things being equal). To the contrary, the fact that the agent executed a distal intention suggests that the action was planned or premeditated, which would, if anything, seem to increase the degree of responsibility.

Note that it does not matter that the distal intentions did not condition the *precise* timing of the action. The Libet experiment required initiation (or inhibition) within a certain timeframe, and so we may assume that participants formed distal intentions to initiate (or inhibit) the action *within* that timeframe, which clearly conditioned the timing of their proximal control. More importantly, the precise timing of an action is usually irrelevant to moral responsibility. We are typically responsible for performing (or failing to perform) certain *types* of action, not for the precise timing of their execution—and if the timing matters, it matters usually only insofar as performing a certain movement at a certain time is constitutive of being a particular type of action. Further, the timing of everyday actions is usually conditioned or constrained by situational factors, and successful execution usually depends on this kind of sensitivity. The fact that our actions are conditioned in this way does not render us less responsible, especially if they are conditioned in accord with distal intentions. And the vast majority of our everyday actions are conditioned by distal intentions and plans. In fact, it is difficult to think of individual acts that are entirely unconditioned—or truly spontaneous—as even spontaneous acts are usually constrained by some of our long-term plans, beliefs, or commitments.

---

[6] Libet and his followers assumed such a strong conception of free will (Libet 1999; Wegner 2002). Some philosophers think that free choices can be conditioned by antecedent states and events only if the conditioning is not deterministic (Kane 1998; O'Connor 2000; Pereboom 2001; Clarke 2003). Others, including myself, think that free will is compatible with deterministic causation as well. In any case, the point here is merely that it seems at least *prima facie* plausible to require unconditioned control for free will.

Note, further, that most experiments on voluntary action and decision-making provide indirect support for the claim that we are able to exercise proximal conscious control by way of forming distal intentions. In most experiments, participants are asked, in effect, to form distal intentions about how to exercise proximal control during the experiment. If everything goes well, they do what they are asked to do, usually in response to becoming aware of certain stimuli or circumstances. Moreover, there is also plenty of direct evidence from the research on implementation intentions for this. Implementation intentions are distal intentions to perform a certain action when certain circumstances obtain. The evidence shows that forming implementation intentions has a medium-to-strong effect on subsequent behavior (for an extensive meta-analysis see Gollwitzer and Sheeran 2006). This is good evidence for the kind of control described above: consciously formed intentions that lead to the exercise of proximal control in response to becoming aware that certain conditions or circumstances obtain.

To conclude, then, the exercise of unconditioned conscious control is not required for moral responsibility, and the claim that we can consciously initiate actions by way of exercising conditioned conscious control is compatible with EC1 and supported by empirical evidence. Note that this response is largely independent from the particular details of the Libet experiment, as it is based primarily on considerations concerning the role of distal intentions in the conditioning of proximal conscious control. It is, in other words, a perfectly general response to the challenge to moral responsibility from EC1.[7]

5.2. Conscious control and EC2

According to EC2, most of our actions are not consciously controlled. Much of the empirical evidence for EC2 stems from priming studies in which behavior is manipulated in ways that bypasses awareness and conscious control. Two main conclusions have been drawn from this research. First, the evidence shows that the control of actions does not require consciousness—that control can be *automatic*. Second, the evidence suggests that much of our everyday behavior is automatic.

[7] Elsewhere I have responded to the closely related neuroscientific challenge to free will and to Daniel Wegner's empirical argument for the claim that conscious will is an illusion. See Schlosser 2012a and 2012b.

There is a vast amount of research on automaticity. In order to illustrate why the evidence is supposed to support the two claims it should suffice to consider three representative types of automaticity: over-learned motor skills, automatic stereotype activation, and automatic imitation. Much of the claimed automaticity in everyday life consists in over-learned motor skills. When one learns a new motor skill, one first has to pay attention to the individual steps. After some practice, the execution becomes automatic and conscious guidance is not required any more. This is thought to be a good thing, as it frees up attention for other tasks. Priming of stereotypes or trait concepts has been shown to increase the tendency to act in accord with the stereotype or trait. For instance, priming with words that are associated with the elderly induces temporarily a slower walking pace and priming with words related to rudeness leads to more interruptions during conversation than in the control group. Further, experiments have shown that we tend to imitate without awareness or conscious control. In particular, given certain goals and circumstances, we automatically imitate the posture of interlocutors. Given this—and many other studies on automaticity—, one can see why researchers have been drawn to the conclusion that much of our behavior is automatic (for more on all this see Greenwald and Banaji 1995; Bargh and Chartrand 1999; Custers and Aarts 2010).

Is any of this incompatible with the control condition in ACC? Note, first of all, that most instances of automaticity are either basic act units that are in the service of conscious intentions or mere modulations of ongoing activity. For instance, the automatic actions that one performs while driving a car usually serve the conscious plan to drive from A to B, and the priming of stereotypes merely modulates how one walks, how one interacts with others, and so on. More importantly, the claim that the control of action does *not require* consciousness is fully compatible with the control condition in ACC. This condition requires the *ability* to exercise conscious control, not the actual exercise of conscious control. Given this, the claim that much of our behavior is not consciously controlled is clearly compatible with the claim that we are able to exercise conscious control over much of our behavior. To illustrate, return to our previous example and assume now that you have a good and reliable habit to indicate—that you tend to indicate automatically when you are supposed to. As mentioned, there is absolutely no reason to think that you are not able to exercise conscious control over your indicating actions. The same holds, to take

only one other example, for the tendency to imitate another person's posture. Nothing in the evidence suggests that this behavior cannot be consciously controlled. If it is brought to one's attention, one could stop imitating, and if one wants to imitate, one could consciously do so.

One might object here that this does not hold for the automatic performance of highly skilled actions. Many actions become merely routine or over-learned with repetition. But some actions become also more skillful. For such actions, conscious control is not only not required, but it is detrimental to a successful performance. In other words, for some actions, control has to become automatic and unconscious for it to reach a certain level of skill. Common examples are the highly skilled movements of, say, athletes or dancers. Given this, one can imagine cases in which an agent is unable to exercise conscious control over an action for which the agent is, intuitively, responsible. Imagine, for instance, a sharpshooter who can hit a moving target only by getting into a state of high responsiveness in which he pulls the trigger automatically, and without conscious control, during a swift move of aiming ahead of the target's trajectory.

However, far from showing that EC2 raises a challenge to moral responsibility, this shows only that there are exceptional cases. As it stands, ACC requires the ability to exercise proximal *and* online control. This makes sense for the vast majority of our actions, and there is no good reason to think that it does not hold for the vast majority of our actions. It holds even for most cases that involve highly skilled movements. Consider, for instance, the movements of a skilled tennis player. Arguably, the player must get into a certain flow of automatic actions and reactions in order to perform well. However, what is automatic here is only the fine-tuning of the player's online control and there is absolutely no reason to think that the player cannot consciously initiate, influence, and abort his play. There is no evidence that would suggest this, and the example strongly suggests that it would simply be crazy to think so.

Let us return, then, to the sharpshooter example. We assume that a successful execution of the shot cannot be under his conscious control. This concerns only the fine-tuning of his *online* control over when to pull the trigger *during* the move of aiming ahead of the target. But for all we know, he has the intention to shoot, and he has conscious *proximal* control over whether or not to initiate the whole sequence. This is why he appears

to be responsible. The lack of online conscious control does not exempt him from responsibility, because he has proximal conscious control over the initiation of the sequence that allows him to exercise automatic control over when to pull the trigger.

We could refine ACC further so as to accommodate such cases. But there is no need to do so, as such cases are highly unusual—ACC remains a very plausible condition that holds for the vast majority of our actions. The important point is that neither EC1 nor EC2 give rise to a general challenge to moral responsibility on the basis of ACC. We have seen, moreover, that there is evidence which provides indirect support for the assumption that we are morally responsible agents, because it provides support for the claim that we are able to exercise conscious control over our actions.

6. Moral responsibility and reason-responsiveness

According to AR, moral responsibility requires the ability to act for reasons. This condition is widely accepted. But like CC, it requires further clarification. Note, first of all, that the condition requires the *ability* to act for reasons. This is only plausible as one may be responsible for an action independently of the reasons one acted for. In particular, one may be responsible for failing to act for certain reasons. It is now common to express this point by the claim that one must be *reason-responsive*: it is required that one is able to respond to reasons (see, for instance, Wallace 1994; Fischer and Ravizza 1998; Vargas 2007). Second, some philosophers have argued that acting for reasons does not require that the reasons play a causal role in the performance of the action. However, the widely held standard view is that an agent acts for reasons only if the reasons—or the mental states that represent the reasons—are causes of the action (Davidson 1963; Goldman 1970; Mele 2003; Schlosser 2010). Third, it seems clear that one may be *morally* responsible only if one is able to act for moral reasons. I shall express this by saying that the agent must be able to respond—or be responsive—to the *relevant* reasons. I take it that this has an epistemic component. That is to say, it is not sufficient that one is responsive in the sense that the reasons can trigger the right response. Rather, the agent must be able to respond to the relevant reasons *qua* reasons—in virtue of knowing and comprehending that they are reasons for the action. Given all this, condition AR can be reformulated as follows:

(ARR) An agent is morally responsible for performing (or failing to perform) an action only if the agent is, at the time, able to respond to the relevant reasons (*qua* reasons).

This is another plausible condition on moral responsibility that appears to give rise to a challenge in connection with the empirical evidence. The most obvious threat here consists in the evidence for EC3.

6.1. Reason-responsiveness and EC3

According to EC3, evidence shows that ordinary and commonsense reason explanations of actions are not genuine causal explanations, because it shows that the reasons that we give are based on self-interpretation, theorizing, or post-hoc rationalization, rather than on direct or introspective access to the causes of our actions. Moreover, processes of self-interpretation are thought to be subject to various biases, and so the reasons that we give in commonsense explanations are likely to be biased too.

Consider, first, two observations. Strictly speaking, the evidence for EC3 cannot show that we are not reason-responsive, as it shows, at best, that the reasons that we take ourselves to be acting for are not the causes of our actions. This is compatible, in principle, with the claim that we are responsive to the relevant reasons and it is compatible with the claim that we are responsive to the reasons that we take ourselves to be acting for, because the fact that one does not respond to certain reasons (*qua* reasons) does not show that one is not able to respond to them. Although this is correct, strictly speaking and in principle, I think that we should not make much of it. The only kind of evidence for reason-responsiveness that we can ever have stems from cases in which we act for reasons (*qua* reasons). If it were true that the reasons that we take ourselves to be acting for are never the real causes of our actions, then our reason-responsiveness would indeed be in doubt. A second point to note is that the evidence stems primarily from post-hoc interviews and questionnaires. This provides only indirect evidence about what reasons the participants took themselves to be acting for *during* the task. It may be, for instance, that perceived pressures to appear reasonable or implicit biases to maintain a certain self-image easily override one's memory. Or we may simply be rather poor in remembering and reporting our reasons—especially if they are reasons for the rather insignificant choices that

participants usually have to make in experimental situations. It is important to bear these points in mind. However, I think that we should not make much of this either. Some reports may be skewed by the implicit pressures of the situation, self-serving biases, or bad memory functions, but it is unlikely that this can explain away all the evidence in support of EC3. Whether we like it or not, we tend to confabulate reason explanations, at least in certain circumstances or with respect to certain actions.

A more significant point is that most of the evidence stems from cases in which there are either *no good* or *no salient* reasons to choose one option rather than another. Consider, for instance, the famous position effect, which is perhaps the most frequently mentioned piece of evidence in support of EC3 (Nisbett and Wilson 1977, 243–44). Participants were asked to evaluate articles of clothing and to select the best quality product. In one study, they were shown four different night gowns. In a second study, they were shown four identical nylon stockings. In both cases, participants showed a pronounced tendency to rate the article positioned at the far right highest. They were entirely unaware of this position effect, and when they were asked to explain their choices, they referred to differences in quality. It is worth to note two points here, which are hardly ever mentioned. First, it would have been very odd if participants had given the position as a reason, because the position is clearly not a good reason to prefer any one item over the others. Second, it is unclear whether there were any salient reasons to prefer one gown over the others in the first study, and it is clear that there was no reason to prefer one of the four products in the second study. Given this, it should really be not so surprising that participants were very poor in giving their reasons. If there are no salient reasons, or if there are no good reasons at all, how could one do a good job in giving good reasons? In other words, it is not so surprising that participants confabulated reason explanations, because they had to. If there are no good or salient reasons, and if the experiment requires you to give reasons, what else can you do?

This helps us to see why the evidence for EC3 should not be taken to raise doubts about our reason-responsiveness. Suppose that you failed to do what you had reason to do: there was a good reason, R, to perform a certain action, but you failed to respond to R. What evidence would count for (or against) the claim that you were reason-responsive, given that you did not respond? Suppose there are similar circumstances in which you act for the reason R, because R is more salient on these occasions. This would be evidence for

the claim that you would have acted for R in the original scenario, had R been more salient. This, in turn, would lend support to the claim that you were responsive to R in the original scenario, even though you did not respond. It would, even, provide some reason to assume that you are, in general, responsive to reasons of this kind.[8] Similarly, consistent failure on your part to respond to reasons of a certain kind—no matter how salient the reasons and no matter how high the stakes—, would provide evidence in support of the claim that you are not responsive to reasons of that kind. The point here is, then, that the empirical evidence in support of EC3, including the mentioned position effect, fails to show that we are not reason-responsive, because it does not provide evidence of this kind.[9]

There is, in fact, plenty of evidence in support of the assumption that we are reason-responsive. In a meta-analysis of evidence concerning the efficacy of intentions, Webb and Sheeran (2006) analyzed forty-nine studies of cases in which participants are given good reasons for significant real-life choices, such as reasons to take physical exercise, to wear a seatbelt, engage in parent-child communication, to have regular health checkups, to practice safe sex, to quit smoking, and so on. They found that the evidence supports the hypothesis that interventions on intentions by way of giving good reasons engenders the corresponding changes in intentions and actions. In particular, their meta-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. Note, here, that the modest effect size on changes in behavior is perfectly in line with reasonable expectations concerning the efficacy of intentions. We know, from everyday experience and observation, that intentions are not very effective when they are up against strong habits or addictions. The important point is that there is a robust effect from reason-giving all the way to changes in behavior across a wide range of real-life situations. Further, note

[8] According to Fischer and Ravizza (1998), reason-responsiveness can be *analyzed* in terms of counterfactual conditionals. I am sympathetic to this view, but here I presume only the weaker claim that the truth of certain counterfactuals counts as *evidence* for reason-responsiveness.

[9] Note that it would not seem to be particularly difficult to gather this kind of evidence. With respect to the position effect, for instance, it would be interesting to see what happens in cases where there are clear qualitative differences. This would provide a contrast class that may tell us something interesting about our responsiveness to the relevant reasons.

that some of the cases have clearly an ethical dimension. Given this, the meta-analysis provides at least some empirical support for the assumption that we are responsive to moral reasons (*qua* reasons).

Finally, the received view, as expressed in EC3, is considerably stronger than the view that Nisbett and Wilson proposed in their seminal article. Their main claim was that verbal reports of mental states are based on self-interpretation, theorizing, or post-hoc rationalization, rather than on direct or introspective access. They noted, however, that this is perfectly compatible with the possibility that we are "often right about the causes of our judgments and behavior" (1977, 253). So, the reasons that we give in ordinary explanations of actions can be the real causes of our actions, and it is possible that they often are the real causes of our actions. This just means that, on their view, we can be responsive to the relevant reasons, and it is perfectly possible that we often are responsive to them. As far as I can see, this point holds for all the evidence that is usually given in support of EC3. This evidence supports, at best, an *epistemic* view about how we form beliefs about our reasons for actions, and it is, as such, perfectly compatible with the claim that the reasons that we take ourselves to be acting for can, and often are, the real causes of our actions (*qua* reasons). Moreover, the mentioned meta-analysis (Webb and Sheeran 2006) even casts doubt on the epistemic view that ordinary reason explanations are always based on self-interpretation (theorizing or post-hoc rationalization). If one is given explicit reasons for a certain action, then it seems unlikely that one will confabulate a post-hoc rationalization, simply because one will not have to confabulate a reason explanation—unlike in cases where there are no good or no salient reasons. Reasons that were made explicit are relatively easy to remember and it is relatively easy to reproduce them, and so it seems likely that one will give those reasons when one is asked to explain one's action.

6.2. Reason-responsiveness and EC4

According to EC4, commonsense explanations of behavior in terms of character traits are systematically mistaken, because behavior is to be explained primarily in terms of situational features. As indicated, the empirical evidence for this view may be taken to challenge the assumption that we are responsive to the relevant reasons, and hence to challenge moral responsibility in conjunction with ARR (see Nelkin 2005). In order to

explain and discuss this, I shall consider two representative examples: the Milgram experiment and the evidence concerning the effects of mood on helping behavior.[10]

In the Milgram experiment (Milgram 1963), participants showed a willingness to administer high levels of electric shocks to a person sitting in an adjacent room as punishment for mistakes in a memory test. The procedure started with shocks that were described as "slight", ending with "danger: severe shock", and "XXX". During this, more and more (prerecorded and fake) signs of distress and pain became audible through a speaker. The majority of participants (26 out of 40) went on to punish with the highest level of shock. Milgram interpreted the results in terms of obedience, as participants were told that they had to continue, as required by the experiment, when they hesitated or questioned the procedure. These orders were issued by an experimenter, who was present in the room, overlooking the proceedings.

There is a large amount of evidence which shows that morally insignificant factors in the environment can affect the willingness to help by way of inducing positive or negative changes in the agent's mood (see, for instance, Carlson et al. 1988; Miller 2009). What is particularly surprising here is that both being in a positive mood and being in a negative mood leads to an increase in helping behavior relative to mood-neutral control groups. Various explanations have been proposed, for instance in terms of mood management: it may be that a good mood increases helping because helping behavior sustains the good mood and that a bad mood also increases helping because helping someone relieves the bad mood. There are problems with this and there are explanations that do not appeal to mood management, but these issues need not concern us here. Whatever the best explanation, it is surprising and counterintuitive that we are, apparently, least willing to help when we are in a neutral mood.

Does evidence of this kind show that we are not responsive to the relevant reasons? In the Milgram experiment, the majority failed to respond to salient moral reasons. They responded, rather, to less important reasons to obey authority. The evidence on mood and helping behavior may be taken to show that we do not respond to moral reasons even when

[10] As Nelkin (2005) points out, the challenge here does not stem from the thesis of situationism (EC4) itself, but from the empirical evidence for this view.

we engage in moral behavior. Under closer inspection one can see, however, that the evidence is not only compatible with the assumption that we are responsive to moral reasons, but that it provides some support for it.

In the Milgram experiment, many of the participants who were willing to administer high shocks showed clear signs of reluctance and distress. Arguably, this is a sign of responsiveness to moral reasons. The assumption that they recognized a conflict between their behavior and what morality requires provides at least a good explanation of their reluctance and distress. More importantly, further experiments have shown that the results vary in significant ways with the level of "personalization" (Bandura 1999):

> Milgram's (1974) research on obedient aggression is widely cited as evidence that good people can be talked into performing cruel deeds. What is rarely noted is the equally striking evidence that most people refuse to behave cruelly, even under unrelenting authoritarian commands, if the situation is personalized by having them inflict pain by direct personal action rather than remotely and if they see the suffering they cause […]. (202)

As explained in the previous section, such variability in responses to reasons is not only compatible with the assumption that we are reason-responsive. Rather, the fact that the likelihood of an appropriate response varies in proportion to the salience (or "personalization") of the relevant reasons is evidence for the claim that we are responsive to the relevant reasons, even in those cases where we fail to respond. Concerning the influence of mood on helping behavior, it is crucial to note that the evidence shows only that being in a good or bad mood *increases* or *augments* the tendency to help. It does not show, or even suggest, that participants help *only* because they are in a good or bad mood (see Miller 2009). For all we know, they help partly because helping is a good thing to do and partly because their mood enhances or augments the motivation to act for that reason. Again, the fact that variations in the circumstances and the subsequent changes in mood tend to increase helping behavior suggests that we are, in general, responsive to the relevant moral reasons, even in those cases where we fail to help. To illustrate, consider any participant, P, who was assigned to a mood-neutral condition and who did not help. The evidence suggests that the right manipulation of P's mood would have made P's helping

likely: had P been assigned to a mood-manipulation condition, P would probably have helped. And, for all we know, if P had helped, P would have helped partly because helping is a good thing to do. This may be taken to suggest that P is responsive to the relevant moral reasons, even though he did not respond to them. Similar points hold, *mutatis mutandis*, for most of the evidence for EC4, and so I conclude that the evidence for EC4 fails also to generate a general challenge to moral responsibility.

7. Theories of moral responsibility and criminal responsibility

This completes my main response to the empirical challenges to moral responsibility. As pointed out, this response is general and largely theory-independent insofar as it applies to the most widely accepted theories of moral responsibility, and insofar as it does not depend on the claims and commitments of particular theories. In this section, I will first say more about which theories are covered by this main response, and I will argue that this response applies also to criminal responsibility. Then I will turn to theories that are not covered by the main response.

7.1. The scope of the main response

I proposed a diagnosis of the empirical challenge that distinguishes between a condition on conscious control (ACC) and a condition on reason-responsiveness (ARR). All theories of moral responsibility require that mental states, such as desires, beliefs, and intentions, have an influence on our choices and actions. Few theories make explicit claims concerning the role of consciousness in the initiation and control of action, but there is no theory according to which conscious control is not required at all. In particular, no theory requires the *exercise* of conscious control as a general necessary condition on moral responsibility, and no theory explicitly denies a condition along the lines of ACC—which requires the *ability* to exercise conscious control. Given this, it seems that the offered responses to the

empirical challenges from conscious control are perfectly general: they would seem to apply to all theories of moral responsibility.[11]

Theories in which the notion of reason-responsiveness takes center stage have been very prominent in the recent debate. They are, as far as I can tell, currently the most widely accepted theories of moral responsibility. The most influential version of this view is the theory developed by Fischer 1994 and Fischer and Ravizza 1998. Other theories that appeal explicitly to reason-responsiveness include Wallace 1994; Arpaly 2003; Vargas 2007; and Nelkin 2011. Further, there are numerous other theories that do not explicitly employ the notion of reason-responsiveness, but that assume a condition along the lines of ARR—a condition that requires the ability to act for the relevant reasons. They include compatibilist theories, such as Dennett 1984; Wolf 1990; Mele 1995; and incompatibilist theories, such as van Inwagen 1983; Kane 1998; O'Connor 2000. Given all this, we can see that the offered responses to the empirical challenges from ARR cover the great majority of prominent theories of moral responsibility, including standard compatibilist and incompatibilist positions.

7.2. Criminal responsibility

So far, the focus has been on moral responsibility. I will show now, in brief, that the main response applies also to criminal responsibility by showing that the proposed conditions on conscious control and reason-responsiveness are in line with the conditions on criminal responsibility. I will base my considerations here on the Model Penal Code (American Law Institute 1985).

Consider, first, the condition on conscious control (ACC). The Model Penal Code requires that a culpable act is voluntary and it defines this negatively: it gives examples of involuntary acts (such as reflexes, sleepwalking, hypnosis) and it states that any bodily movement is involuntary "that is otherwise not a product of the effort or determination of the actor, either conscious or habitual" (§ 2.01). According to the commentary, this is

[11] As pointed out in section 5, ACC appears to be the best candidate for establishing a plausible connection between moral responsibility and conscious control. Should ACC be untenable, then so are the empirical challenges to moral responsibility from conscious control.

supposed to capture "conduct that is not within the control of the actor" (215). Arguably, to require that something is "within" the control of the agent is to require that the agent is able to exercise control over it. Further, the qualification that the control may be "either conscious or habitual" strongly suggests that criminal responsibility does not require the exercise of conscious control. This shows that the conditions on criminal responsibility are in line with ACC, which requires the ability to exercise conscious control.

According to ARR, moral responsibility requires the ability to respond to the relevant reasons (*qua* reasons). A good indication that this is in line with the conditions on criminal responsibility is provided by the Model Penal Code's section on "mental disease or defect excluding responsibility" (§ 4.01). This section states that an agent is not responsible for criminal conduct if the agent "lacks substantial capacity either to appreciate the criminality [wrongfulness] of his conduct or to conform his conduct to the requirements of the law" (ibid., brackets in original). Although this passage refers to "conduct as a result of mental disease or defect", it clearly highlights the underlying presupposition that one must be able to understand the reasons why certain acts are wrong and to adjust one's behavior accordingly—a presupposition that is in line with ARR. See also Morse (2006, 37), who points out that "legally responsible or legally competent agents are people who have the general capacity to grasp and be guided by good reason".

Given, then, that ACC and ARR are in line with the conditions on criminal responsibility, and given that the empirical evidence does not undermine AAC and ARR, as I have argued, we can conclude that the empirical challenges to common conceptions of human agency do not raise a general challenge to criminal responsibility either.

7.3. Self-disclosure and consequentialist theories

In section 4, I already addressed one type of view that is not covered by the main response: character-based accounts. My approach was twofold. First, I noted that character-based accounts are unpromising and widely rejected, and then I pointed out that the empirical evidence would not undermine such views anyway. There are two more types of theories that are apparently not covered by the main response: consequentialist theories (for instance Schlick 1939; Smart 1963) and, what I shall call, self-disclosure theories. The latter come in two main versions. Hierarchical self-disclosure theories require, roughly, that the agent

identifies (or would identify) with the desire that motivates the behavior in question (Frankfurt 1971; Dworkin 1988). Non-hierarchical views require that the behavior in question is an expression of the agent's deep concerns and values (or the agent's "deep self") without imposing conditions concerning the agent's identification with desires. Some versions of non-hierarchical views require reason-responsiveness (Watson 1975; Arpaly 2003), whereas others require only that the agent is *in principle* reason-responsive—that an *ideal* agent would respond (Scanlon 1998; Smith 2005).

My approach here is again twofold. First, I think it is fair to say that self-disclosure and consequentialist views are not nearly as widely endorsed as reason-responsiveness theories. In the case of consequentialist theories, this is an understatement. They are widely considered to be untenable (see, for instance, Strawson 1962; Wallace 1994; Scanlon 1986). Hierarchical theories are plagued by internal regress problems (Watson 1975) and they are, perhaps, better interpreted as accounts of self-governance rather than moral responsibility (Bratman 2004, for instance). They are, in any case, not particularly promising as accounts of moral responsibility. The most obvious problem stems from cases of weak-willed actions for which the agent is intuitively responsible, but which are motivated by desires with which the agent neither does identify nor would identify (see, for instance, Fischer 2012). Non-hierarchical self-disclosure views are free from the mentioned regress problems, but it is not clear that they can account for responsibility for weak-willed actions.

In any case, it is not difficult to see that the empirical challenges to common conceptions of agency (EC1 – EC4) do not raise general challenges for those views either. Let us consider first non-hierarchical self-disclosure views. As mentioned, some versions of this view require reason-responsiveness. They are covered, in part, by the main response—they are covered insofar as the empirical challenges do not undermine the assumption that we are reason-responsive, as I have argued. Does the evidence challenge the assumption that our actions often express our concerns and values? The evidence for EC3 shows that our beliefs about which attitudes motivate our actions are, in certain circumstances, prone to errors. The evidence for EC4 shows that we tend to underestimate the influence of situational factors, and that we tend to overestimate the extent to which our actions express our attitudes. Further, the evidence for EC2 includes the evidence on the automatic influence of implicit biases, such as implicit gender or racial stereotypes (Greenwald and

Banaji 1995, for instance). This raises difficult questions for non-hierarchical self-disclosure views. Are implicit attitudes ever part of one's "true self"? Are they, perhaps, part of one's true self if they are in line with one's explicit attitudes? And so on. However, despite raising difficult questions, the evidence does clearly not show that our actions never express our attitudes (including our deep concerns and values). It does not even show that our actions rarely express our attitudes. For all I can see, the evidence is compatible with the assumption that our actions often express our concerns and values, partly because the expression of attitudes does not depend on reliable self-knowledge; partly because I see no reason to assume that most intentional actions are driven by implicit biases or situational factors; and partly because the influence of implicit biases and situational factors does not rule out the influence of the agent's concerns and values. Moreover, we have seen that there is positive and direct evidence for the assumption in question. The meta-analysis concerning the efficacy of intentions (Webb and Sheeran 2006) shows that changes in the agent's explicit attitudes tend to bring about the corresponding changes behavior, which suggests, among other things, that our actions tend to express our concerns and values.

Let us turn, then, to hierarchical versions of the self-disclosure view. One might worry here that some of the evidence for EC2 and EC4 raises a challenge for such views (concerning the latter see Doris 2002, 140–146). If one is influenced by implicit biases or by situational factors that one is not aware of, then it seems that one is not in a position to identify with the attitudes that motivate one's behavior. Moreover, one probably would not identify with them, as implicit biases and the motives that drive responses to situational cues are often in conflict with one's explicit attitudes (by which I mean, roughly, attitudes that are consciously accessible and reflectively endorsed). Yet it is not obvious that one is not morally responsible for the resulting behavior (more on this in section 8). My main response here is that there is no reason to suppose that this worry generates a *general* challenge to moral responsibility—a challenge to the assumption that we are, in general, morally responsible agents. There are two points to be made here. First, even if the influence of implicit biases and situational factors is more widespread than we would like to think, I see no reason to suppose that they influence most of the actions for which we tend to hold each other responsible. Second, it is questionable that agents would never identify with the influence of implicit biases or the motives that drive responses to

situational cues. Some implicit influences may be in line with one's explicit attitudes. Further, on many occasions, it may well be that we would identify with implicit attitudes even if they are not in line with our explicit attitudes. Proponents of hierarchical theories often assume that to identify with a motive is the same as endorsing that motive in accordance with one's explicit evaluative attitudes. However, there is a plausible weaker notion of identification. One may be said to identify with a motivational attitude by accepting that the attitude is "one's own", and one may accept this even if that attitude is not in line with one's explicit evaluative attitudes. Suppose you learn that an action of yours was motivated by an implicit attitude in a way that conflicts with your explicit evaluative attitudes. You may not identify with the attitude in any sense, and you may, thereby, refuse to take responsibility for the action. However, I see no reason to suppose that we would always take such a stance. There may well be many cases in which we would accept such new self-knowledge as a discovery that tells us something about ourselves and in which we would, thereby, take responsibility for the behavior in question. Finally, even though it is not obvious that one is not morally responsible for actions that are influenced by implicit biases or situational cues, it is also not obvious that one always is responsible in such cases (we will return to this in section 8). All in all, we can conclude that this worry does not have the potential to generate a general challenge to moral responsibility.

This brings us, finally, to consequentialist theories. Very roughly, such views justify the practices of praise and blame in terms of their social utility—their good consequences for society as a whole. Usually, they require freedom from coercion and constraint, and they acknowledge that certain physical and mental impairments can excuse or exempt an agent from moral responsibility. But they do not impose further general conditions concerning the agent's freedom or control. Rather, they presuppose that the practices of praise and blame are effective in maximizing utility and regulating society. It seems clear that none of the empirical challenges to common conceptions of agency (EC1 – EC4) raises a challenge to this assumption. Some of the evidence may be taken to provide useful information on how the practices of praise and blame can be improved so as to maximize

their efficacy in the regulation of behavior. But there is nothing in the evidence that suggests that such practices are altogether ineffective.[12]

8. The relevance of empirical research

I have argued that the empirical evidence does not support a general challenge to responsibility. I did not mean to suggest, however, that empirical evidence on human agency is irrelevant to questions concerning responsibility. To the contrary, I think it is quite obvious that this research can be highly relevant in several ways. A detailed discussion of this issue is beyond the scope of this paper. But before I conclude, I shall add a few observations and remarks to substantiate this point.

Most obviously, perhaps, scientific evidence can be of relevance in individual cases by helping to establish that excusing or mitigating conditions obtain. For instance, it has been suggested that a teacher's obsession with child pornography was entirely due to a tumor in the frontal lobe, which rendered him also unable to control the urges to acquire pornographic material (Burns and Swerdlow 2003). Whether or not evidence of this kind does excuse or mitigate depends, of course, on the details of the particular case. But it seems clear that evidence of this kind can be relevant in individual cases.

More generally, empirical evidence suggests that damage to certain brain areas and certain kinds of substance abuse can diminish or undermine the ability to inhibit urges and impulses (Burns and Bechara 2007). This may diminish or undermine the agent's ability to exercise conscious control and it may diminish or undermine the agent's ability to act for reasons. Given this, empirical evidence can and should inform also our views on certain *types* of cases, and this point is not restricted to cases that involve pathologies or addictions. Take, for instance, the evidence on the influence of implicit biases. Such biases are widespread among individuals with different ethnic, social, and economic backgrounds (Greenwald and Krieger 2006) and their influence may bypass awareness in various ways. One may be unaware of having a certain bias, one may be unaware of its activation on particular occasions, and one may be unaware of the way in which it affects one's

[12] I am setting aside here the difficult question of whether imprisonment and other forms of legal punishment are effective means for the regulation of behavior and society—this is a separate issue.

judgments and actions. It has been controversial whether or not one can ever rid oneself of implicit biases, but it is generally agreed that their influence can be reduced and controlled (Blair 2002; Devine et al. 2002). Despite this, we face an epistemological problem here, as we are usually unaware of their activation and of the way in which they influence our judgments and actions. So, even if we are, in principle, able to control or counteract the influence of implicit biases, conscious control is severely hampered by the fact that it is difficult to know when and how to exercise it.

Something similar can be said about some of the evidence for EC3 and EC4. Usually, the influence of situational factors does not render us unable to exercise conscious control, and it does not undermine our responsiveness to reasons. But some of the evidence for EC3 and EC4 raises questions concerning the scope or the degree of our responsibility in certain situations, as it shows that irrelevant factors can influence our judgments and actions in ways that bypass conscious awareness. As before, we face the epistemological problem that it is difficult to know when and how one should try to control the influence of such factors, as it is difficult know when and how they influence our choices and actions.

Further, it cannot be denied that some of the evidence shows also that we—or most of us—are less responsive to reasons than we would like to think. Most find the results of the Milgram experiment and of other infamous studies, such as the Stanford prison experiment (Haney et al. 1973), disturbing and contrary to prior expectations. The evidence on mood enhanced helping behavior is less shocking, as a lot less seems to be at stake. But it is also surprising, in some respects, and contrary to common sense—although what is contrary to common sense here is not so much the degree of reason-responsiveness, but the way in which it is modulated.

I do not want to suggest that any of this shows that we are always excused in certain types of cases, and I doubt it shows that responsibility is always diminished in certain types of cases. First, the fact that some circumstances decrease our reason-responsiveness is not sufficient to excuse or mitigate. To me, it seems that the degree of reason-responsiveness that is compatible with the evidence is sufficient for the ascription of full moral (and criminal) responsibility—given, of course, that no other excusing or mitigating conditions

obtain.[13] Second, it is questionable that responsibility requires that one can become aware of the influence of certain factors if it is possible to acquire indirect (theoretical or inferential) knowledge about when and how such factors are operative. It is even unclear that responsibility requires that one has such indirect knowledge. Arguably, in some cases it is sufficient that one could have acquired the relevant knowledge—cases in which one "should have known". This suggests that the assessment of such cases depends on epistemic conditions. Unfortunately, there is no widely accepted view concerning the relevant epistemic conditions on responsibility, and it is clearly beyond the scope of this paper to propose and defend such a view.

However, even if the evidence does not establish excusing or mitigating conditions for certain types of cases, it is nevertheless relevant. It raises important questions, it provides important knowledge about the causes of our actions, and it can thereby help us to better our agency. Further, it can be argued that some of the evidence generates new responsibilities for institutions and organizations—responsibilities to provide training and to require the use of procedures that are known to reduce the influence of implicit biases and certain situational factors (see Jolls and Sunstein 2006, for instance).

Finally, empirical evidence can and should inform theoretical questions concerning the capacities that underlie responsible agency. In philosophy, it has been common to assume that reason-responsiveness is of one piece, as it were—that there is one single mechanism or faculty of practical reason that is responsive to reasons for action. In particular, it has been common to assume that one and the same mechanism is operative during offline (or reflective) deliberation and online decision-making. In contrast, in the cognitive sciences it is now widely assumed that reasoning and decision-making is implemented by two distinct systems (or types of processes): system 1 is automatic, effortless, and heuristics-based and system 2 is conscious, deliberate, and rule-based. This framework has been successfully deployed in many areas of research (for overviews see Sloman 1996; Evans 2008). System 2 is plausibly associated with the ability to engage in

[13] According to Fischer and Ravizza (1998), reason-responsiveness requires that there is an understandable pattern of counterfactual scenarios in which the agent recognizes the relevant reason and that there is at least one possible world in which the agent acts on that recognition (69–81). This account would seem to support my claim that the degree of reason-responsiveness is sufficiently robust in the cases in question.

practical reasoning, and hence with reason-responsiveness. However, some of the evidence suggests that system 1 is reason-responsive as well, as it seems to show that system 1 leads in some tasks to better choices than system 2 (Dijksterhuis 2004, for instance). The implicit claim here is that system 1 is *by itself* reason-responsive. This may be challenged and there is evidence to suggest that reason-responsive processing depends on an interplay between the two systems (Baumeister et al. 2011, for instance). However, if it turns out that both system 2 and system 1 are reason-responsive mechanisms, then we should abandon the assumption that there is one single faculty of practical reason.

Let me close with two more remarks on this. First, dual-system theory is not universally accepted (for critical reviews see Osman 2004; Keren and Schul 2009). But even a dual-system account of practical reason would seem to *vindicate* the assumption that we are reason-responsive. It would, however, make questions and answers concerning our reason-responsiveness more complex. For instance, either one of the two systems may be reliably responsive to the relevant reasons in some circumstances, but unreliable and unresponsive in others. The two systems may interact in most situations, they may interfere with each other in some cases, there may be cases in which processing switches between the two systems, and so on. Second, it seems clear that empirical research on dual-system theories of reasoning and decision-making is highly relevant to the question of *how* reason-responsive the agent is in different circumstances and with respect to different kinds of reasons (De Neys 2006, for instance).

9. Conclusion

I have argued that the empirical challenges to common conceptions of human agency fail to generate a general challenge to moral (and criminal) responsibility. They fail to show that we lack the required abilities, and so they fail to show that we are, in general and unless specific excusing or mitigating condition obtain, responsible agents. Despite this, empirical evidence is clearly relevant. It can and should inform our theoretical understanding of the mechanisms that underlie responsible agency, and it may change normative judgments in individual cases or with respect to certain types of cases.

We have seen that the evidence on human agency is not only compatible with the relevant conditions on conscious control and reason-responsiveness, but that there is plenty

of empirical evidence in support of both conscious control and reason-responsiveness. Given this, we can conclude that the empirical evidence on human agency is not only compatible with the assumption that we are, in general, responsible agents, but that it provides, in fact, support for it.

I have addressed here only the relevance of *existing* empirical evidence on human agency for the obvious reason that one can only speculate about the relevance of future findings. I think it may well be that evidence from psychology and cognitive neuroscience will become more and more relevant in the future as the science of human agency progresses. Is there any reason to think that we will face more credible general challenges to responsibility from future findings on human agency? Perhaps the right answer here is simply that we have to wait and see—or that this issue is beyond the scope of this paper. Let me point out, however, that some of the considerations that I have presented here may be taken to suggest that it is *difficult to see* how future findings could possibly generate a more credible challenge. In particular, it is difficult to see how future experiments could show that we are unable to exercise conscious control or that we lack the ability to respond to reasons, as controlled experiments on human agency seem to presuppose that we have those abilities. They presume that participants are able to exercise control in response to becoming aware of certain cues and in accord with instructions, which provide them with reasons for action within the context of scientific experimentation. Given this, it seems reasonable to think that the relevance of empirical evidence to moral (and criminal) responsibility will remain restricted to individual cases and types of cases, and to our theoretical understanding of the underlying mechanisms.

Acknowledgments

The research for this article was funded by a grant from the Netherlands Organization for Scientific Research (NWO). Earlier versions were presented at the 38th Conference on Value Inquiry (Salem State University), a workshop on Responsibility and Neuroscience (Institute of Philosophy, London), and a research meeting at the University of Leiden. I would like to thank the participants at these meetings for their helpful comments, and I am especially grateful to Neil Levy for very helpful written comments on an earlier draft.

References

American Law Institute. 1985. *Model Penal Code and Commentaries*. Philadelphia, PA: The American Law Institute.

Arpaly, Nomy. 2003. *Unprincipled virtue*. Oxford: Oxford University Press.

Bandura, Albert. 1999. Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review* 3: 193–209.

Baer, John, James C. Kaufman, and Roy F. Baumeister. (eds.). 2008. *Are we free? Psychology and free will*. Oxford: Oxford University Press.

Bargh, John A., and Tanya L. Chartrand. 1999. The unbearable automaticity of being. *American Psychologist* 54: 462–479.

Baumeister, Roy F., E. J. Masicampo, and Kathleen D. Vohs. 2011. Do conscious thoughts cause behavior? *Annual Review of Psychology* 62: 331–361.

Blair, Irene V. 2002. The malleability of automatic stereotypes and prejudice. *Personality and Social Psychology Review* 6: 242–261.

Bratman, Michael E. 2004. Planning agency, autonomous agency. In *Personal autonomy*, ed. James Stacey Taylor, 33–57. Cambridge: Cambridge University Press.

Burns, Kelly, and Antoine Bechara. 2007. Decision making and free will: A neuroscience perspective. *Behavioral Sciences and the Law* 25: 263–280.

Burns, Jeffrey M., and Russell H. Swerdlow. 2003. Right orbitofrontal tumor with pedophilia symptom and constructional apraxia sign. *Archives of Neurology* 60: 437–440.

Carlson, Michael, Ventura Charlin, and Norman Miller. 1988. Positive mood and helping behavior: A test of six hypotheses. *Journal of Personality and Social Psychology* 55: 211–229.

Custers, Ruud, and Henk Aarts. 2010. The unconscious will: How the pursuit of goals operates outside of conscious awareness. *Science* 329: 47–50.

Clarke, Randolph. 2003. *Libertarian accounts of free will*. Oxford: Oxford University Press.

Davidson, Donald. 1963. Actions, reasons, and causes. *Journal of Philosophy*, 60, 685–700.

De Neys, Wim. 2006. Dual processing in reasoning: Two systems but one reasoner. *Psychological Science* 17: 428–433.

Dennett, Daniel C. 1984. *Elbow room: The varieties of free will worth wanting*. Cambridge: MIT Press.

Devine, Patricia G., E. Ashby Plant, David M. Amodio, Eddie Harmon-Jones, E., and Stephanie L. Vance. 2002. The regulation of explicit and implicit race bias: The role of motivations to respond without prejudice. *Journal of Personality and Social Psychology* 82: 835–848

Dijksterhuis, Ap. 2004. Think different: The merits of unconscious thought in preference development and decision making. *Journal of Personality and Social Psychology* 87: 586–598.

Doris, John M. 2002. *Lack of character: Personality and moral behavior*. Cambridge: Cambridge University Press.

Dworkin, Gerald. 1988. *The theory and practice of autonomy*. Cambridge: Cambridge University Press.

Evans, Jonathan St. B. T. 2008. Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology* 59: 255–78.

Fischer, John M. 1994. *The metaphysics of free will*. Oxford: Blackwell.

—. 2012. Semicompatibilism and its rivals. *Journal of Ethics* 16: 117–143.

Fischer, John M., and Mark Ravizza 1998. *Responsibility and control: A theory of moral* responsibility. Cambridge: Cambridge University Press.

Frankfurt, Harry G. 1969. Alternate possibilities and moral responsibility. *Journal of Philosophy* 66: 829–39.

—. 1971. Freedom of the will and the concept of a person. *Journal of Philosophy* 68: 5–20.

Gazzaniga, Michael S., and Joseph E. LeDoux. 1978. *The integrated mind*. New York: Plenum.

Goldman, Alvin. 1970. *A theory of human action*. New York: Prentice-Hall.

Gollwitzer, Peter M., and Paschal Sheeran. 2006. Implementation intentions and goal achievement: A meta-analysis of effects and processes. *Advances in Experimental Social Psychology* 38: 69–119.

Greene, Joshua and Jonathan Cohen. 2004. For the law, neuroscience changes nothing and everything. *Philosophical Transactions of the Royal Society of London* B 359: 1775–1785.

Greenwald, Anthony G., and Mahzarin R. Banaji. 1995. Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review* 102: 4–27.

Greenwald, Anthony G., and Linda H. Krieger. 2006. Implicit bias: Scientific foundations. *California Law Review* 94: 945–967.

Haney, Craig, Curtis Banks, and Philip Zimbardo. 1973. Study of prisoners and guards in a simulated prison. *Naval Research Reviews* 9: 1–17.

Harman, Gilbert. 1999. Moral philosophy meets social psychology: Virtue ethics and the fundamental attribution error. *Proceedings of the Aristotelian Society* 99: 315–331.

Hume, David. 1960. *A treatise of human nature*, ed. L. A. Selby-Bigge. Oxford: Clarendon Press, originally published 1740.

Johansson, Petter, Lars Hall, Sverker Sikström, Betty Tärning, and Andreas Lind. 2006. How something can be said about telling more than we can know: On choice blindness and introspection. *Consciousness and Cognition* 15: 673–692.

Jolls, Christine, and Cass R. Sunstein. 2006. The Law of implicit bias. *California Law Review* 94: 969–996.

Kane, Robert. 1998. *The significance of free will*. Oxford: Oxford University Press.

Keller, I., and H. Heckhausen. 1990. Readiness potentials preceding spontaneous motor acts: Voluntary vs. involuntary control. *Electroencephalography and Clinical Neurophysiology*, 76, 351–61.

Keren, Gideon, and Yaacov Schul. 2009. Two is not always better than one: A critical evaluation of two-system theories. *Perspectives on Psychological Science* 4: 533–550.

Kihlstrom, John F. 1987. The cognitive unconscious. *Science* 237: 1445–1452.

Libet, Benjamin. 1985. Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Science* 8: 529–66.

—. 1999. Do we have free will? *Journal of Consciousness Studies* 6: 47–57.

Mele, Alfred. 1995. *Autonomous agents*. Oxford: Oxford University Press.

—. 2003. *Motivation and agency*. Oxford: Oxford University Press.

—. 2009. *Efficacious intentions: The power of conscious will*. Oxford: Oxford University Press.

Milgram, Stanley. 1963. Behavioral study of obedience. *Journal of Abnormal and Social Psychology* 67: 371–8.

—. 1974. *Obedience to authority: An experimental view*. New York: Harper and Row.

Miller, Christian. 2009. Social psychology, mood, and helping: mixed results for virtue ethics. *Journal of Ethics* 13:145–173.

Morse, Stephen J. 2006. Moral and legal responsibility and the new neuroscience. In *Neuroethics*, ed. Judy Illes, 33–50. Oxford: Oxford University Press.

Nelkin, Dana K. 2005. Freedom, responsibility and the challenge of situationism. *Midwest Studies in Philosophy* 29: 181–206.

—. 2011. Making sense of freedom and responsibility. Oxford: Oxford University Press.

Nisbett, Richard E., and Timothy D. Wilson 1977. Telling more than we can know: Verbal reports on mental processes. *Psychological Review* 84: 231–259.

O'Connor, Timothy. 2000. *Persons and causes: The metaphysics of free will*. Oxford: Oxford University Press.

Osman, Magda. 2004. An evaluation of dual-process theories of reasoning. *Psychonomic Bulletin and Review* 11: 988–1010.

Pockett, Susan, William P. Banks, and Shaun Gallagher. (eds.). 2006. *Does consciousness cause behavior?* Cambridge, MA: MIT Press.

Pereboom, Derk. 2001. *Living without free will*. Cambridge: Cambridge University Press.

Roskies, Adina. 2006. Neuroscientific challenges to free will and responsibility. *Trends in Cognitive Sciences* 10: 419–23.

Ross, Lee, and Richard E. Nisbett. 1991. *The person and the situation*. Philadelphia: Temple University Press.

Scanlon, T. M. 1986. The significance of choice. *The Tanner Lectures on Human Values* 7: 149–216.

—. 1998. *What we owe to each other*. Cambridge: Harvard University Press.

Schlick, Moritz. 1939. When is a man responsible? In *Problems of ethics*, Moritz Schlick, 141–158. New York: Prentice-Hall.

Schlosser, Markus E. 2010. Agency, ownership, and the standard theory. In *New Waves in Philosophy of Action*, ed. Jesus Aguilar, Andrei A. Buckareff, and Keith Frankish, 13–31. Basingstoke: Palgrave Macmillan.

—. 2012a. Free will and the unconscious precursors of choice. *Philosophical Psychology* 25: 365–384.

—. 2012b. Causally efficacious intentions and the sense of agency: In defense of real mental causation. *Journal of Theoretical and Philosophical Psychology* 32: 135–160.

Sinnott-Armstrong, Walter, and Lynn Nadel. (eds.). 2011. *Conscious will and responsibility: A tribute to Benjamin Libet*. New York: Oxford University Press.

Sloman, Steven A. 1996. The empirical case for two systems of reasoning. *Psychological Bulletin* 119: 3–22.

Smart, J. J. C. 1961. Free will, praise, and blame. *Mind* 70: 291–306.

Smith, Angela M. 2005. Responsibility for attitudes: Activity and passivity in mental life. *Ethics* 115: 236–71.

Soon, Chun Siong, Marcel Brass, Hans-Jochen Heinze, and John-Dylan Haynes. 2008. Unconscious determinants of free decisions in the human brain. *Nature Neuroscience* 11: 543–545.

Strawson, Peter F. 1962. Freedom and resentment. *Proceedings of the British Academy* 48: 1–25.

van Inwagen, Peter. 1983. *An essay on free will*. Oxford: Clarendon Press.

Vargas, Manuel. 2007. Revisionism. In *Four views on free will*, ed. John M. Fischer, Robert Kane, Derk Pereboom, and Manuel Vargas, 126–165. Oxford: Blackwell Publishing.

Wallace, R. Jay. 1994. *Responsibility and the moral sentiments*. Cambridge: Harvard University Press.

Watson, Gary. 1975. Free agency. *Journal of Philosophy* 72: 205–20.

Webb, Thomas L., and Paschal Sheeran. 2006. Does changing behavioral intentions engender behavior change? A meta-analysis of the experimental evidence. *Psychological Bulletin* 132: 249–268.

Wegner, Daniel M. 2002. *The illusion of conscious will*. Cambridge: MIT Press.

Wolf, Susan. 1990. *Freedom within reason*. Oxford: Oxford University Press.

Zhu, Jing. 2003. Reclaiming volition: An alternative interpretation of Libet's experiment. *Journal of Consciousness Studies* 10: 61–77.