Rev.Phil.Psych. (2015) 6:815–825
DOI 10.1007/s13164-014-0190-7

Mental Agency as Self-Regulation

Leon de Bruin · Fleur Jongepier · Derek Strijbos

Published online: 22 June 2014

Abstract The article proposes a novel approach to mental agency that is inspired by Victoria McGeer's work on self-regulation. The basic idea is that certain mental acts (e.g., judging that p) leave further work to be done for an agent to be considered an authoritative self-ascriber of corresponding dispositional mental states (e.g., believing that p). First, we discuss Richard Moran's account of avowals, which grounds first-person authority in deliberative, self-directed agency. Although this view is promising, we argue that it ultimately fails to confront the empirical gap between occurrent judgments and dispositional beliefs. Second, we show how Victoria McGeer's account of self-regulation allows us to bridge this gap by emphasizing that avowals are only reliable and authoritative insofar as we take certain steps to live up to the commitments inherent in our self-ascriptions. Third, we address the question whether and to what extent self-regulation can be seen as a form of mental agency. Unlike the 'pure' deliberative form of mental agency advocated by Moran, which is direct, conscious and intra-personal, we follow McGeer and argue for a notion of mental agency as an (often) indirect, unconscious and inter-personal process of self-regulation.

1 Introduction

In the debate on mental agency there is a widespread agreement that mental acts, such as judging, thinking, choosing or deciding are not intentional actions. Although we can judge that *p*, we cannot intend to judge that *p*; although we can decide that *p*, we cannot intend to decide that *p*. Opinions diverge considerably, however, about what this implies for the significance of mental agency in our conscious mental lives. Some philosophers, most notably Galen Strawson (2003), conclude that most of our thoughts 'just happen' - intentions have little or nothing to do with their occurrence. Others, such as Peacocke, O'Shaughnessy and Mele are more optimistic. Although they agree with Strawson that it does not make sense to suggest that we can intend to think a particular

L. de Bruin (✉)
VU University, Amsterdam, Netherlands
e-mail: lcdebruin@gmail.com

F. Jongepier · D. Strijbos
Radboud University, Nijmegen, The Netherlands

thought, they do maintain that intentions play an important role in directed thinking. According to Peacocke (1999, 209), for example, directed as opposed to idle thought involves 'the intention to think a thought which stands in a certain relation to other thoughts or contents'. Similarly, O'Shaughnessy (2000, 89 and 211) claims that the intentions involved when one is engaged in such activity select 'the content of the governing enterprise', 'stir one's mental machinery' and constrain, 'under definite description', the advance of one's thinking. Finally, Mele (2009) proposes a distinction between 'trying to Φ' and 'trying to bring it about that one Φs'. Although one's Φ-ing need not be an action, the occurrence of the Φ-ings may still be explained, in part, by the fact that one has engaged in the mental action of trying to bring it about that one Φs. For example, Mele argues that, although one cannot strictly speaking try to remember something, one can try to bring it about that one remembers.

All these authors are concerned with the question whether the process that leads to an occurrent judgment, thought, choice or decision involves some kind of mental agency. In the present article, by contrast, we focus on another target for mental agency. The basic idea is that certain mental acts (e.g., judging that *p*) leave further work to be done for an agent to be considered an authoritative self-ascriber of corresponding dispositional mental states (e.g., believing that *p*). The main question we want to investigate is how this 'further work' should be spelled out, and to what extent it qualifies as mental agency.

The article is structured as follows. In the next section, we start with a discussion of Moran's account of avowals. According to Moran, we express first-person authority over our mental states by avowing them from a committed, endorsing, first-personal stance through deliberation on the objects of our mental states. In section three, we argue that Moran's account fails to properly account for the unfortunate (empirical) fact that our words and actions often do not mesh. In such cases, a 'gap' emerges between our occurrent judgments and our dispositional beliefs, which makes that we can no longer be considered authoritative self-ascribers. In section four we turn to McGeer's (2007) account of self-regulation, which provides the right resources to deal with this problem. According to McGeer, our avowals are only reliable and authoritative insofar as we take certain steps to live up to the commitments inherent in our self-ascriptions. Thus, performing a mental act such as judging that *p* implies that one intentionally self-regulates oneself into the disposition of believing that *p*. We discuss several examples, ranging from everyday life to psychopathology, in order to shed light on the link between self-regulation and first-person authority. In section five we address the question whether and to what extent self-regulation can be seen as a form of mental agency. Unlike the 'pure' deliberative form of mental agency advocated by Moran, which is direct, conscious and intra-personal, we follow McGeer and argue for a notion of mental agency as an (often) indirect, unconscious and inter-personal process of self-regulation.

2 Moran on Mental Agency and First-Person Authority

According to Moran, mental agency is crucial because without it we cannot be *self*-knowers in any robust sense of the term. Knowledge of one's own mind involves being able to express first-person authority, which has to do with the authority of a person to make up her mind and speak for her beliefs. Moran claims that first-person authority cannot simply rest upon some kind of observational privilege. Imagine a creature with

complete information, complete accuracy and complete reliability regarding its own mental states.[1] Yet its relation to its mental states is entirely passive: it is utterly unable to affect what comes by in the passing show of consciousness, and its attitude to its own mind is purely spectatorial and third personal. Moran argues that despite its superior epistemic capacities, such a creature would lack the kind of first-person authority that defines human self-knowledge. First-person authority, he claims, requires not (just) being able to adopt an epistemic stance towards our own mental states, but an agentive one:

> "One is an agent with respect to one's attitudes insofar as one orients oneself toward the question of one's beliefs by reflecting on what's true, or orients oneself toward the question of one's desires by reflecting on what's worthwhile or diverting or satisfying […] There is a role for the agent here insofar as we may speak of a person's responsibility for his attitudes" (2001, 64).

Central to Moran's account is the idea that self-knowledge manifests itself primarily in the form of avowals of one's attitudes from a committed, endorsing, first personal stance. An avowal is a declaration of one's belief that obeys the so-called 'Transparency Condition': "With respect to belief, the claim of transparency is that from within the first-person perspective, I treat the question of my belief about P as equivalent to the question of the truth of P" (2001, 62–63). Thus, in asking oneself, "Do I believe that p?" one reflects directly on the object of one's belief, weighing reasons for and against p being the case, settling one's belief according to the outcome of this deliberative process. Importantly, the ability to avow our beliefs in conformity to the Transparency Condition presupposes a capacity for mental agency: "[O]nly if I can see my own belief as somehow 'up to me' will it make sense for me to answer a question as to what I believe about something by reflecting exclusively on that very thing, the object of my belief" (2001, 66–67). There is no 'extra' work to be done to form our beliefs on Moran's account: deliberative avowal of the object of one's belief constitutes one's belief. As a result, judgment and belief cannot come apart - judging or avowing that *p* just is the determination of one's belief (see also Hamilton 2000, 23; Hieronymi 2009).

Although avowals have a special constitutive role in so far as they provide what we might call maker's knowledge, they do not of course literally guarantee that actions implied by the corresponding beliefs, intentions, etc. are brought about. This is the famous condition of the literary character Oblomov: every day he decides to get up and dress himself, yet again and again he fails to bring his avowals into accordance with his actions.[2] The discrepancy between avowal and action is, of course, not only the condition of Oblomov; it is something that we are all too familiar with in everyday life. Apart from the obvious cases of lying and cheating, we can sometimes behave and

[1] Strawson (2003), for example, claims that there is nothing incoherent in the idea of a 'Pure Observer": a reasoning, thinking and judging creature that has a full and vivid sense of itself as an observer although it has no capacity for any sort of intentional action. Although Strawson admits that it is 'excessively' unlikely that any such creature could evolve naturally, he argues that this does not alter the fact that Pure Observers are conceptually possible.

[2] On Moran's view, one's mental act (e.g., one's decision to get out of bed, or to stop gambling) is only as strong as one's hold on one's practical reasons (2003, 82). If it turns out, in Oblomov's case for example, that one's resolutions turn out to be poor indicators of the future, then on Moran's view we should say that Oblomov did not, in fact, have sufficient *reason* to get out of bed. In other words, Oblomov had not properly made up is mind, because he could not accept the commitments that were implied by his resolution.

act in ways that conflict with our consciously held beliefs (cf. Schwitzgebel 2010). The question is how we should think of first-person authority in such cases.

3 Putting Our Money Where Our Mouth is

The chasm between the first-person perspective, from which the agent asks: "what shall I—qua agent—believe?", and the third-person perspective, from which the agent stands back from herself and poses the question, "what do I—qua subject—believe?", takes a central place in Moran's work (see esp. his discussion of Sartre's example of the akratic gambler, e.g. Moran 2003, 78–83). According to Moran, first-person authority can only be granted from the first-person perspective – the 'transcendental perspective of agency'. By contrast, an agent who can only resort to the third-person perspective, what Moran calls the 'empirical perspective of psychological facticity', becomes 'alienated' from himself: he ceases to function as a rational being, and is no longer able to adopt the stance from which he is able to 'declare the authority of reason over his beliefs and his actions' (2001, 127). Given that on Moran's account the empirical self-perspective ultimately undermines one's authority, and given that the only type of mental agency that can truly secure first person authority is a deliberative kind of agency, the mismatch between avowal and action can only be solved by a further instance of deliberation; by (re) considering one's practical reasons for a certain belief, desire or intention. Someone might for instance come to reconsider his decision to take up a prestigious job with a much better salary after having carefully deliberated about what his new life would be like, and the kind of person he might turn into. Now this might indeed be one way to overcome the gap between saying and doing, but it is certainly not the only way.[3] In fact, there are serious problems with the idea that first-person authority can be solely grounded in rational deliberation.

First of all, we sometimes deny someone authority over her own mental states in spite of being a rational agent, capable of adopting a deliberative stance. Take Sarah, for instance, who recently took the Implicit Association Test (IAT).[4] The IAT is a computer-based test that measures people's 'implicit biases' - the positive or negative attitudes towards a person, thing or group that they hold at an unconscious level (in contrast to their explicit biases, i.e. the attitudes that they are consciously aware of having). In a conversation about ethnic biases Sarah might arrive at the explicit judgment that all races are of equal intelligence, but her scores on the IA might proof otherwise and show that she implicitly believes that black people are less intelligent than white people.[5] We might rebuke Sarah for not living up to the implications and commitments of her judgment, and for not really believing what she says she believes.

[3] Note, though, that deliberative instances like these are often not a 'private' or individual matter, but something we do together with others: deliberation is often an *interpersonal* affair. As we shall argue in section 4, we might understand this type of interpersonal deliberation (e.g. calling someone to ask his/her opinion on taking up the new job) as a form of self-regulation.

[4] The IAT requires people to complete several tasks where they are asked to quickly pair two concepts together. For example, you might be asked to pair "women" with "math" or "women" with "liberal arts." Scoring of the IAT assumes that the more closely you associate two concepts in your mind, the faster you will be able to pair them together on the task. The IAT measures your reaction times and calculates a score accordingly. See: https://implicit.harvard.edu/implicit/demo/background/index.jsp

[5] Since its development in 1997, over 4.5 million people haven taken the IAT online. The collected data strongly suggests that many of these people hold implicit biases towards members of particular groups.

In such a case, we deny Sarah first-person authority, i.e. we do not consider her to be an authoritative self-ascriber of the belief that all races are of equal intelligence - despite the fact that Sarah is a rational agent who is capable of adopting a deliberative stance toward her own mental states.[6]

The reverse is true as well: we sometimes grant an agent authority in spite of being irrational. This happens for instance in cases of psychopathology, where patients report on their irrational emotions or feelings. It is a fact of our everyday folk psychological practice, and a crucial assumption of diagnostic interviewing and clinical treatment, that even when someone has outweighing reasons against holding a certain emotion or feeling, and can thus be characterized as irrational, we do not question that the person speaks her mind authoritatively. One might even argue that many of our emotions and feelings are not 'rational' or 'irrational' in the first place.[7] We sometimes know we are angry or sad, before knowing the reasons as to why we are angry or sad, or whether we ought to feel that way. In spite of the fact that we sometimes have no reasons for or against holding an emotion, we nonetheless seem to enjoy first-person authority in our expressions of them. Thus, even though Moran's deliberative model seems to works rather well for beliefs, it is not clear whether it can explain first-person authority for other mental states.

Another problem is that Moran, while putting a lot of emphasis on the importance of rational deliberation, seems to underestimate what can be achieved by adopting an instrumental third-person perspective towards oneself. Take the case of Sarah, for instance. When she has to find a new applicant for a job, we would probably advice her to use a blind evaluation procedure in order to neutralize her implicit biases, instead of asking her to deliberate about her feelings towards black people, given that her beliefs about black people are precisely opaque to her, and hence not available for transparent deliberation.[8] If Sarah were to start systematically evaluating her own behavior and making attempts to change her implicit beliefs via changing her environment or job application procedures, we would not be led to think that she is somehow alienated toward herself, acting in 'bad faith' or avoiding responsibility—quite the contrary. On Moran's broadly Kantian view, however, there does not seem to be much room for such instrumental, non-deliberative type of agency, at least not in so far as this type

[6] Bilgrami (2006, 2010) argues that there is an ambiguity in the notion of intentional states such as belief. On the one hand beliefs are understood as normative states, i.e. commitments, on the other hand they are viewed as mere dispositions. He thinks we should keep these conceptions strictly separate and then goes on to argue that we always have first-person authority over our intentional states, but only when conceived as commitments. This enables him to deny that first-person authority is undermined in cases of self-deception, such as Sarah's. He argues that self-deception is not a case of having a false meta-belief about the existence of the relevant first-order belief (which would undermine first-person authority), but rather a true meta-belief about or avowal of a first-order *commitment* that is incompatible with another first-order dispositional belief state. We agree with Bilgrami that intentional states are inherently normative, but we are hesitant about the distinction he proposes. For one, we think there is no such sharp divide in folk psychology. However, also on Bilgrami's account there is the question how one gets from making a commitment to being granted first-person authority regarding one's corresponding dispositional states. This is the question that concerns us in this paper.

[7] We would like to thank an anonymous referee for pointing this out.

[8] Notice that the case would not be any different if Sarah were aware of her implicit biases (and hence could deliberative about her states in a transparent fashion). It is a notorious fact about implicit biases that they persist even when people know they have them.

of agency is supposed to enhance a person's authority and autonomy, rather than to diminish it.

4 First-person Authority through Self-Regulation

Victoria McGeer (1996; 2007) offers a refreshing and more realistic take on how to think of first-person authority within the constitutive framework. Whereas Moran's model presents a 'classical' Kantian picture of mental agency as the activity of the individual to make up his mind through practical deliberation, McGeer offers a picture of agency that hinges on interpersonal relationships and allows for instrumental reasoning. As McGeer observes, our avowals are only authoritative in so far as we take certain steps to live up to the commitments underlying our self-ascriptions. We demand not only that people be able to come up with reasons for their beliefs, but also that they engage in self-regulative activities if necessary. In contrast to the pure deliberation of some belief or intention, self-regulation involves the question of "how to bring about causally what ought to arise spontaneously as the expressive outcome of deliberation itself" (McGeer 2007, 91; see also Moran 2001, 117–118). Though there is still a central place for mental agency and avowals in McGeer's picture of self-knowledge, the activity of avowing is not necessarily a matter of transparent deliberation, but includes avowing in the service of empirical self-regulation. For instance, a person's decision to attend AA meetings is not (just) a transparent decision about what to do or believe (which is fallible), but crucially involves a commitment to regulate one's states in such ways as to make it the case that one, in fact, attends these meetings. This may include trying to be a better deliberative agent, but more often involves engaging in extra-deliberative activities, such as making notes to oneself, calling a friend if in doubt, or even changing one's (social) environment.[9]

Self-regulation does not meet Moran's transparency condition, as it involves being focused on the nature of mental states *as* mental states, rather than their content or the 'object' of our states (see also Vierkant 2012). Given that we have learned that our intentions may change once we grow older (e.g. one might lose one's passion for philosophy, or come to adopt radically different political views), we may take certain steps to ensure that our 'future selves' will be bound by our present intentions, e.g. one might decide to give all one's money away to some political party or charity organization to avoid a change of heart in one's future life. On a more daily basis, we are usually well aware that alcohol may change the way we think about certain people or events; that hunger may cause frustration; that going out for a walk when one is stuck during some task is sometimes better than trying to finish it. We may, accordingly, take steps in order to bring about some change in our mental states not by reflection on some belief or intention, but by changing our environments, engaging in different sorts of behaviour, making bets with friends, and so on.

[9] On Moran's account, people suffering from weakness of will strictly speaking lack self-knowledge - as they have failed to properly make up their minds. On McGeer's view, however, avowals might not just express one's resolutions, but also issue an 'invitation' to start engaging in self-regulative practices. A person's avowal to attend an AA-meeting, for instance, might express self-knowledge in spite of not being sure-fire, because it expresses a commitment to bring it about that one attends the meeting.

McGeer's account thus reminds us of the fact that in everyday life, merely making up one's mind about something, avowing or committing oneself to some proposition in deliberative spirit is often not enough to determine one's future mental states accordingly. But there is a further, more radical implication. We need self-regulation to guard us against the more subversive inclinations of our rational faculty. The literature on self-deception is full of examples in which people lure themselves into believing or wanting something despite deep convictions, desires or feelings to the contrary.[10] Given the fact that these deeply engrained dispositional states will almost always win out on the long run (otherwise it would probably not count as a case of self-deception), rational deliberation may very well enlarge the gap between our openly avowed commitments and our deep-seated mental dispositions and subsequent actions. This can also happen when our capacity for rational deliberation is under influence of our current moods, needs, appetites, etc. A deliberative judgment may seem perfectly rational when, for example, feeling elated just after having achieved a success. Yet taking on just this one more assignment despite one's cramped schedule, may turn out to be disappointingly unrealistic when moods have calmed down after a good night sleep. The point here is not that it is impossible to question or reflect on our avowals on Moran's account, since self-reflection on one's reasons, on Moran's Kantian view, is precisely what is at stake (see e.g. Moran 2001, 138–148). The point, rather, is that reflection on one's transparent avowals can only be achieved by yet a further instance of transparent deliberation. As McGeer shows, deliberation is not always the golden route to reflect, change and assess one's transparent judgments. More importantly, deliberation is only possible and can be effective only in so far as certain mental states are transparent to us from the first-person perspective, but this leaves reflection on our opaque states, such as implicit beliefs, unaccounted for.[11]

Self-regulative authoritative agency implies that one is continually ready to take a step back from one's first-person, rationalizing inclinations in order to reflect on them from a second- or third-person point of view on self, thereby making a more comprehensive assessment of one's own situation and the relevant factors (contextual, psychological, pharmacological, etc.) that shape one's reasoning and one's capacity to stay true to one's commitments. To earn the status of reliable, authoritative self-ascribers, we thus not only need to be able to program our minds and (social) surroundings such that we stay on the path set out by our commitments, we also need to carefully monitor the deliberative processes that motivate them. As McGeer puts it:

> (…) our best protection— indeed, our only protection—against an ego-driven corruption of reason is to cultivate an allocentric capacity to see ourselves as we see others—namely, as empirical subjects whose psychological states are responding to a variety of influences that are largely invisible from a naïvely egocentric first-person point of view. (2007, 101)

[10] See McGeer (2007, 92): "developing deliberative autonomy or 'spontaneity' in Moran's purist sense can be a sign of real psychic disease, indicating a capacity to manipulate oneself through the power of one's own reason into a condition of deep self-deception." See also Lear (2004).

[11] Moran allows that there can be what he calls "psychic givens", but stresses that such "givens" must be reflected upon, only then can they be "understood in terms of the person's responsibilities, and hence as implying either "endorsement, permission, or disapproval," or simply passive allowance" (2001, 148). The point here is that for opaque states, even passive allowance is not available, as these states are not available for reflection in the first place. And even if they were, pure awareness often does not do the work.

The tight connection between the capacity for self-regulation and first-person authority is clearly borne out in cases where the former is severely deficient, such as in cases of psychopathology. Consider the case of Peter, who has avowed to stop drinking once and for all. Peter's family and friends, however, know better: they have been through this process of avowal before, and time and time again, he has failed to live up to them. Peter is an alcoholic, who, due to his addiction, experiences major problems in programming his moods, appetites, activities, etc. such that he will abstain from drinking. His avowal of sobriety has become a hollow phrase to those around him, and there may come a point when people start doubting the content of the avowal itself. After so many relapses, even Peter himself might start to question what it really is he values or desires most when he avows to stop drinking once again. The occurrent avowal may feel as genuine and wholeheartedly as any, yet when almost all ties to his future dispositions are cut, it becomes unsure what it still means.

What this case from psychopathology has in common with everyday life examples (e.g., Sarah's case of implicit bias) is that both feature a failure of self-regulation. But there are also important differences. In everyday life, we assume that people have the capacity to bring what they say in line with their actions. Suppose your spouse asks you whether you truly believe that family is more important than work. There are two ways in which you could meet this challenge. First, you might try to get your priorities straight, work less and spend more time with your family. If you succeed, he or she will (hopefully) consider you to be an authoritative self-ascriber of the belief that you value family over work. Second, you can come to the conclusion that your spouse is actually right: you wish you would have the opportunity to work less and spend more time with your family, but unfortunately this is impossible (perhaps you're a workaholic, or perhaps the both of you need the money badly). However, in that case, you should know better than to tell him or her that you value family over work. In other words, in everyday life we assume that people still have sufficient capacity for self-regulation that enables them 'to do better', either by changing their words or by changing their actions. In cases of psychopathology, this assumption may become problematic. Peter fails to live up to his avowal to stop drinking once and for all, and at some point, we give up hope and no longer require and/or expect him to bring his dispositional beliefs and desires in line with his judgments and decisions relating to his drinking. The issue whether he still has first-person authority with respect to these dispositional states no longer seems to apply. We no longer consider him to be the right target to respond to our regulative practice of correction and encouragement.

5 But is it Mental Agency?

In the previous sections, we have investigated what it takes for an agent engaged in mental acts as judging that *p* or deciding to *a*, to be considered an authoritative self-ascriber of the corresponding dispositional states of believing that *p* or having the desire to *a*. We have argued that Moran's account is promising, but fails to explain how we can take responsibility for mental states that are not transparent to us, and does not offer us the kind of agency that is required to 'speak for' our dispositional states. Also, we have shown how McGeer's (2007) account is able to overcome this problem by emphasizing the importance of engaging in self-regulative activities. We will now turn

to the question whether and to what extent self-regulation can be seen as a form of mental agency.

As we already alluded to in the introduction, there is a striking difference between mental agency and bodily agency. Whereas I can 'directly' bring it about that there is light in the room (by flipping a switch) or that the letter 'h' appears on my screen (by hitting the 'h' key), I cannot similarly bring about the belief that the earth is triangular. This is not only because we cannot simply believe at will. It also seems the case that, in order to exercise agency over our thoughts at all (such as considering whether *p* is true, or using *p* in inferences), the thought-content must have *already* made its appearance in our consciousness (see also Strawson 2003, 235–238). In other words, the presence of certain thought contents cannot be a matter of agency at all.

However, we also briefly touched upon Mele's (2009) distinction between 'trying to Φ' (which Strawson and others deny is possible) and '*trying to bring it about* that one Φs'. Though most people find it impossible to *try to sneeze*, it is quite possible to *try to bring it about* that one sneezes (e.g. by looking directly into a bright light or sniffing freshly grained pepper). Likewise, it may be possible to try to bring it about that one believes *p* or decides *q*, leaving room for the idea that thinking, deciding, and so forth, are mental actions after all. We might refer to this distinction as the distinction between 'direct' and 'indirect' forms of mental agency (see also Dorsch 2009 who discusses the notion of 'mediated' agency).

This distinction is helpful for understanding the relation between self-regulation and mental agency. Self-regulation falls under the 'indirect' category of mental action: it involves managing one's life and organizing one's environment (including one's social environment) *such that* one is likely to entertain different or new mental states, or likely to make different or new decisions. For instance, we might not be able to exercise any kind of direct mental agency over deciding to start a relationship with X (since, strictly speaking, there is no such thing as '*trying* to decide'), we nonetheless can take steps to *bring it about* that we decide to start a relationship with X (e.g. by making a list of all the good qualities of X, by asking other people what they think of X, by going out with X, by confessing your love to X, by kissing X etc.) Thus, self-regulation is not a matter of directly bringing certain thought-contents to mind, but rather of creating a situation in which they would (or could). Importantly, on this line of reasoning, self-regulation is compatible with Strawson's claim that our thought contents are not intentionally brought about, and that the genuine role of action in thought is merely 'catalytic' (Strawson 2003, 231).

At this point, there are two questions that we need to consider. First, should we indeed take self-regulation as a form of mental agency? Second, is self-regulation as an indirect form of mental agency as insignificant as Strawson suggests?

When discussing cases of self-regulation, we find that the notion of 'mental agency' is ambiguous between the claim that it concerns agency over 'mental objects', such as thoughts and intentions, and the claim that it concerns a type of mental activity, i.e. that the type of agency exercised must be performed 'in the head'. Self-regulation is a form of mental agency in the first, but not (necessarily) the second sense. This is because the distinction between bodily and mental agency starts to break down once we consider self-regulative activities in more detail. Consider, for instance, going for a walk to clear one's head, taking painkillers to alleviate pain, eating a snack to get out of a grumpy mood, breathing slowly to reduce one's nerves, or perhaps even forcing a smile to cheer

oneself up. These cases are 'direct' instances of agency insofar as they concern direct bodily actions; they are 'indirect' instances of mental agency, however, insofar as these bodily actions are performed with the specific intention of changing or bringing about certain mental states in oneself. There is a sense, then, in which self-regulation might involve a direct form of agency after all, though not 'mental' in the classic (and hence, in our view, overly narrow) sense of the term.

Regarding the second question, we want to suggest that the 'catalytic', indirect dimension of mental agency is much, much more substantial than Strawson takes it to be. Strawson makes it quite clear that he does not expect too much of indirect mental agency:

> No doubt there are other such preparatory, ground-setting, tuning, retuning, shepherding, active moves or intentional initiations. But action, in thinking, really goes no further than this. The rest is waiting, seeing if anything happens, waiting for content to come to mind (…). (Strawson 2003, 232)

Strawson would probably allow for first-person authority even in the absence of any form of mental agency. That is, even if one is in no way responsible for a certain thought content springing to mind—suppose someone utters "pink elephant!" in your direct vicinity, leading you to think or visualize a pink elephant—this presumably changes nothing to the fact that you are nonetheless able to report, with full authority, "I am thinking of a pink elephant". Now this might work for occurrent or 'episodic' states (but see our discussion of Moran's arguments against the idea that first-person authority can be granted in virtue of some kind of observational privilege in section 2). However, it is not clear how passive self-observation could grant us first-person authority over dispositional states, such as the belief that you value family more than work, or that you hold no biases with respect to people's race or gender.

As we have shown, self-regulation provides a straightforward answer to this question. We take people's word for what they say because we can justifiably assume—save the exceptional case—that they are in the position to ensure that their words line up with their actions, and it is this assumption that grounds authority. And, similarly, other people consider us to be authoritative self-ascribers of the beliefs we express, as long as we are able to regulate ourselves into having the dispositions these beliefs imply.

Much of our self-regulation is performed more or less automatically and with very little reflection. Self-regulation often becomes part of our habits or routines (e.g. not going to bed too late, having a decent breakfast, turning off the music when starting to work, etc.). There are also various forms of self-regulation that are more effortful and require deliberation and careful reflection. We might reflect on whether some of our beliefs should in fact be abandoned, and we might take steps so that we are more likely to abandon them. We might also pay special attention to our belief-forming habits, and closely monitor the way in which we reach certain beliefs, checking our habits of inference and guarding against malfunction (Pettit and McGeer 2002, 289; McGeer 2008, 89). The point is not that we (must) engage, constantly, in all these types of self-regulation, but rather that we have the capacity to do so were the need to arise, and that we may justly demand from others that they use this capacity in situations that commits them to do so. Self-regulation is a form of mental agency that is deeply embedded in our social practice. Without it, we would lose much of our status as authoritative self-ascribers of our dispositional mental states.

Acknowledgements We would like to thank Naomi Kloosterboer and two anonymous referees for their valuable comments and suggestions. During the writing of this article, Leon de Bruin's research was supported by a grant from the Templeton World Charity Foundation. The opinions expressed in this publication are his own and do not necessarily reflect the views of Templeton World Charity Foundation. Fleur Jongepier's research was supported by The Netherlands Organisation for Scientific Research (research project 322-20-003).

References

Bilgrami, A. 2006. *Self-Knowledge and Resentment*. Cambridge: Harvard University Press.

Bilgrami, A. 2010. Précis of Self-Knowledge and Resentment. *Philosophy and Phenomenological Research* 81(3): 749–765.

Dorsch, F. 2009. Judging and the Scope of Mental Agency. In Lucy O'Brien & Matthew Soteriou (eds.), *Mental Actions*. Oxford: OUP, 38–71.

Hamilton, A. 2000. The authority of avowals and the concept of belief. *European Journal of Philosophy* 17: 20–39.

Hieronymi, P. 2009. Two kinds of agency. In Lucy O'Brien & Matthew Soteriou (eds.), *Mental Actions* Oxford: OUP, 138–162.

Lear, J. 2004. Avowal and Unfreedom. *Philosophy and Phenomenological Research* 69: 448–454.

McGeer, V. 1996. Is 'Self-Knowledge' an Empirical Problem? Renegotiating the Space of Philosophical Explanation. *Journal of Philosophy* 93(10): 483–515.

McGeer, V. 2007. The regulative dimension of folk psychology. In Daniel Hutto & Matthew Ratcliffe (eds.), *Folk-Psychology Re-Assessed*. Dordrecht: Springer, 138–156.

McGeer, V. 2008. The moral development of first-person authority. *European Journal of Philosophy* 16(1): 81–108.

Mele, A. 2009. Mental Action: A Case Study. In Lucy O'Brien & Matthew Soteriou (eds.), *Mental Actions*. Oxford: OUP, 17–37.

Moran, R. 2001. *Authority and Estrangement*. Princeton: Princeton University Press.

O'Shaughnessy. 2000. *Consciousness and the World*. Oxford: OUP.

Peacocke, C. 1999. *Being Known*. Oxford: Oxford University Press.

Pettit, P., and V. McGeer. 2002. The Self-Regulating Mind. *Language and Communication* 22(3): 281–299.

Schwitzgebel, E. 2010. Acting contrary to our professed beliefs or the gulf between occurrent judgment and dispositional belief. *Pacific Philosophical Quaterly* 91: 531–553.

Strawson, G. 2003. Mental ballistics or the involuntariness of spontaneity. *The Proceedings of the Aristotelian Society* 77: 227–256.

Vierkant, T. 2012. What Metarepresentation is For. In Michael Beran, Johannes Brandl, Josef Perner, & Joëlle Proust (eds.), *Foundations of Metacognition*. Oxford: OUP, 279–288.