Ethic Theory Moral Prac (2015) 18:297–307
DOI 10.1007/s10677-015-9579-3

Self-Interpretation as First-Person Mindshaping: Implications for Confabulation Research

Derek Strijbos · Leon de Bruin

Accepted: 24 February 2015 / Published online: 17 March 2015

Abstract It is generally acknowledged that confabulation undermines the authority of self-attribution of mental states. But why? The mainstream answer is that confabulation misrepresents the actual state of one's mind at some relevant time prior to the confabulatory response. This construal, we argue, rests on an understanding of self-attribution as first-person *mindreading*. Recent developments in the literature on folk psychology, however, suggest that mental state attribution also plays an important role in regulating or shaping future behaviour in conformity with normative expectations. We explore an analogue understanding of self-attribution of mental states in terms of first-person *mindshaping*. The main aim of this paper is to explore how this insight alters the implications of empirical confabulation studies on first-person authority. We also indicate how this sheds new light on the phenomenon of confabulation itself.

Keywords Self-attribution · Confabulation · Self-regulation · Mindshaping · First-person authority

1 Introduction

We often attribute beliefs, desires, emotions and intentions to other people. For example, we readily understand that Sarah takes the exit because she *wants* to avoid the traffic jam ahead, and that John carries an umbrella because he believes it is going to rain soon. We also frequently attribute mental states to *ourselves*. Why are you dressed up like a rabbit? Because I believed this was going to be a costume party (and apparently I was wrong about that). How are these capacities for first-person and third person mental state attribution related?

It is a common belief, deeply embedded in our folk psychological practice, that there is an important asymmetry between first-person and third-person ways of knowing minds. Barring

D. Strijbos (✉)
Department of Philosophy, Radboud University Nijmegen, Erasmusplein 1, 6500 HD Nijmegen, The Netherlands
e-mail: d.strijbos@ftr.ru.nl

L. de Bruin
Department of Philosophy, VU University Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands

obvious cases of self-deception or folk psychological incompetence, people are normally considered to be in a privileged position to make claims about their own mental states. That is, self-attributions of mental states are generally considered to have a special authoritative status compared to the attribution of mental states to others.

However, findings in the empirical sciences seem to undermine this special kind of first-person authority. In the wake of the seminal study by Nisbett and Wilson (1977), several experiments have been taken to show that human beings are prone to confabulation - we are often not aware of the actual causes of our actions and just make them up as we go along (cf. Gazzaniga 1998; Wilson 2002; Wheatley 2009). Nisbett and Wilson asked people in a shopping mall which pair of identical pantyhose on a display table they preferred and why. They all pointed to some attribute of the preferred pair, such as its superior knit, sheerness, or elasticity. However, with the exception of a single student who was taking psychology courses, nobody mentioned what according to the study design was to be considered the actual cause of their preference – the preferred pair's relative position – even when specifically asked whether their choice had been influenced by position. On the basis of these and other findings, it has been argued that to a significant extent, self-attribution of mental states should be understood as a form of confabulation or 'intention invention'. (Wegner 2002)

There is much controversy to what extent these experimental findings generalize. However, many seem to agree that *if* they do, that is, if confabulation turns out to be a widespread phenomenon in everyday social practice, this would seriously undermine first-person authority of mental state attribution. Leaving aside the empirical question regarding the validity and generalization of mentioned findings, we want to focus on the conceptual question *how* confabulation is supposed to undermine first-person authority (if these findings would indeed generalize). In other words, our question is how to understand the implications of confabulation on the authoritative status of mental state self-attribution.

The mainstream answer to this question is that confabulation undermines first-person authority because it misrepresents the actual state of one's mind at some relevant time in the past. This construal, we argue, depends on a conception of confabulation (and self-interpretation in general) as a form of first-person *mindreading*. In this paper, we want to explore the implications of a complementary view of self-interpretation as first-person *mindshaping*. On this view the (in)adequacy of self-interpretation is not merely determined by the mental states instantiated at some previous time, but also by one's future behaviour, i.e., to what extent one is able to live up to the interpretation one has given. We argue that this view partly mitigates the undermining effect of confabulation on first-person authority. At the same time, however, it points towards important future-directed aspects of confabulation that have so far been neglected in philosophical discussions.

Our paper is structured as follows. In the next two sections, we introduce the notion of confabulation (section two) and the distinction between mindreading and mindshaping (section three). We go on to explore two accounts of self-attribution of mental states in terms of first-person mindshaping. First, we discuss Richard Moran's (2001) account of self-knowledge, which explains first-person authority in terms of our capacity to deliberatively make up our minds (section four). Although this capacity for rational self-determination arguably does play a role in upholding our status as authoritative self-ascribers, we argue that it has little resources to mitigate the authority-undermining subliminal influences on thinking and decision-making processes that the confabulation studies are designed to tease out (section five). In section six, we turn to Victoria McGeer's (2008) account of self-regulation. We argue that this account of first-person mindshaping is better suited to account for authoritative self-attribution of mental states in light of the confabulation studies. In section seven, we turn the tables and explore

what this self-regulatory aspect of first-person authority implies for a fuller account of confabulation.

2 Confabulation as Failed Self-Interpretation

There is no clear-cut definition of the term 'confabulation'. It was first used by the German neurologists Bonhoeffer, Pick and Wernicke over a century ago as a technical term to indicate false memory reports made by patients who suffered what came to be known as Korsakoff's amnesia (Hirstein 2009). It was later applied to other syndromes in neurology (e.g., Geschwind 1965; Gazzaniga 1998) and psychiatry (e.g., Kraepelin 1913; Wing et al. 1974), as well as in the fields of social psychology (e.g., Nisbett and Wilson 1977), moral psychology (e.g., Haidt 2001) and many, many more. As its extension expanded, controversy about the definition of the concept grew. In the traditional sense, 'confabulation' designated the phenomenon of people unintentionally making false reports about their memory. It was later applied to cases that didn't involve explicit reports and came to cover other mental states and processes than memory, such as intentions and actions, emotions and perceptions (for an overview see Hirstein 2009).

For our purposes, we will take the provisional definition given by Coltheart and Turner (2009, 180) as our point of departure. They state that:

> When a person does not know or does not have access to the answer to a question addressed to that person (typically the question might be a request for explanation of why a person behaved in a certain way or else a question asking why the person holds a particular belief), but when asked the question responds by offering an answer to it rather that saying 'I don't know', and if this is done with no intention to deceive the questioner, then that response counts as a confabulation.

How to explain confabulation? In their seminal article, Nisbett and Wilson (1977, 248) suggested that "When people are asked to report how a particular stimulus influenced a particular response, they do not do so by consulting a memory of the mediating process, but by applying or generating causal theories about the effects of that type of stimulus on that type of response." (for discussion see Coltheart and Turner 2009) On this view, self-attribution is a form of self-*interpretation*. When giving answers to questions about our reasons for action, we do not introspect on the mental states that caused our action, but rather come up with a folk-psychological story that makes it plausible why the type of action we performed is a reasonable response to the type of situation we faced. Accordingly, cases of confabulation are understood as *failed* attempts at self-interpretation.

In what follows, we will simply assume that: 1) Nisbett and Wilson's explanation of the experimental data is roughly correct, and 2) that it also applies to related experimental setups triggering confabulation about other kinds of mental states, such as emotions, beliefs, moral judgments, and so on. Our question is: how is ill-grounded self-interpretation supposed to undermine first-person authority?

3 Self-Interpretation: Mindreading or Mindshaping?

The view that self-attribution of mental states is the result of interpretation rather than introspection, has grown more popular over the years. (e.g., Dennett 1987, Gazzaniga 1998, Carruthers 2011). On this view, there is no principled psychological distinction between self-

attribution of mental states and attribution of mental states to others. First-person and third-person attribution differ in their interpretative targets, but make use of one and the same underlying folk psychological capacity. Interestingly, the confabulation data feature prominently in interpretationist accounts of self-attribution (e.g., Gazzaniga 1998, Carruthers 2011). Following the suggestion of Nisbett and Wilson (1977) mentioned above, the idea is that cases of confabulation are best explained as self-interpretation *gone wrong*, implying that adequate forms of everyday self-attribution are best understood as self-interpretation gone right.

How is confabulation on this interpretationist view supposed to undermine the authority of mental state self-attributions? From an interpretationist viewpoint, how one should answer this question depends on one's understanding of *third-person* mental state attribution, i.e., on one's understanding of folk psychology.

Traditionally the literature on folk psychology has been dominated by two mainstream accounts: the Theory Theory and the Simulation Theory. Roughly, the Theory Theory claims that our capacity for mindreading depends on a psychological theory that specifies how mental states are related to environmental conditions, observable behavior and other mental states. The Simulation Theory, by contrast, argues that mindreading involves "putting ourselves in the others' shoes" by simulating the beliefs, desires and intentions we would have in their situation. Despite their differences, both theories (and their various spin-offs) revolve around the same idea, namely, that folk psychology centers on the capacity for *reading other people's mental states*, with the primary function of predicting and explaining their behavior (Hutto and Rattcliffe 2007; Gallagher 2012; Andrews 2012). We may label this the mindreading approach to folk psychology. It assumes that proper use of folk psychology consists in making correct judgments about other people's mental states and their behavior from a more-or-less observational stance.

On the mindreading approach, the adequacy of third-person interpretation is measured by the extent to which the mental states attributed match the states that caused the behavior that was the target of interpretation. Applying this rendering of folk psychology to the first-person, self-interpretation becomes a matter of "turning our mindreading capacities upon ourselves" (Carruthers 2009, 123). Analogous to the third-person case, confabulation is expected to result in incorrect judgements and therefore as lacking first-person authority. On the assumption of widespread confabulation in everyday social practice, this account of self-interpretation would thus seriously undermine the commonly held belief that self-ascriptions of mental states have special first-person authority.

Over the last decade, however, the mindreading approach to folk psychology has been challenged. It has been argued that mental state attribution is also about *mindshaping* (Zawidzki 2013). We do not just passively read the mental states of other agents in order to predict or explain their behavior. Rather, we often *make* their behavior readable to us. That is, we modify other minds so as to match more closely with our normative expectations. This happens through a variety of practices, behaviors and mechanisms - including imitation, pedagogy, norm cognition and enforcement, and language based regulative frameworks, like self- and group-constituting narratives (McGeer 2007; Hutto 2008; Gallagher 2012; Zawidzki 2013). From a developmental perspective, this mindshaping view holds that folk psychological explications serve the primary function of teaching children what to think and how to act under certain circumstances, thereby regulating their thoughts and behaviour so as to match socio-cultural norms. Correspondingly, everyday use of folk psychology has the important social function of correcting, normalizing or justifying behaviour that is seemingly 'out of line', thereby re-enforcing the norms by which we judge each other's conduct. The adequacy of third-person mental state attribution is therefore not only dependent on whether the interpretative target in fact instantiated these states at some relevant time in the past, but also, and

under certain circumstances (e.g., education) most importantly, on whether the target will adjust her behaviour accordingly in the future.

How should we interpret the experimental findings on confabulation when we apply the idea of mindshaping to self-attribution of mental states? That is, how does confabulation affect first-person authority if self-attribution is regarded as a capacity for 'self-shaping'? In order to answer this question, we will examine two different accounts of self-attribution found in the philosophical literature: Moran's (2001) understanding of first-person authority as 'deliberative avowal' and McGeer's (2008) account of self-regulation.

4 First-Person Mindshaping as Deliberative Avowal

Moran (2001) claims that first-person authority requires not (just) being able to adopt an observational stance towards our own mental states, but (also) an agentive one, in the sense that we actively have to *make up our mind* about what we believe. Central to Moran's account is 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) Moran justifies this move by appealing to our capacity for rational agency. We have and form our beliefs and other attitudes in a way that is *responsive to reasons*. (cf. Hieronymi 2009) This means that we should play an active role in the determination of our beliefs by *avowing* them, i.e., by explicitly endorsing their content.[1]

On Moran's account, first-person authority rests on the authorial privilege to determine one's beliefs (and other propositional attitudes) by rational deliberation. The confabulation studies suggest that we do not always judge best which states and processes are psychologically active in us and have a causal impact on our behavior. However, Moran argues that this sort of epistemic accuracy is not really to the point:

> "The primary thought gaining expression in the idea of 'first-person authority' may not be that the person himself must always 'know best' what he thinks about something, but rather that it is *his business* what he thinks about something, that it is *up to him*. In declaring his belief, he does not express himself as an expert witness to a realm of psychological facts, so much as he expresses his rational authority over that realm" (Moran 2001, 123–4, italics added).

Thus, on Moran's view, first-person authority is explained in terms of self-directed rational agency. It is not primarily a matter of having epistemic authority over certain psychological facts that happen to be one's own, but rather of taking responsibility for one's mental states, by deliberating about the reasons for having them and thus shaping one's own mind through deliberative avowal. This is what according to Moran grounds the special authority of self-knowledge relative to knowledge of other minds.

On Moran's account, first-person mindshaping is restricted to what can be deliberatively avowed in full consciousness. From the perspective of deliberative avowal, the subliminal factors that influence our reasoning and decision-making cannot come into view, let alone

[1] This does not mean that we can adopt or reject beliefs at will (and therefore it should not be understood as a kind of doxastic voluntarism). What it means is that we need to take responsibility for our attitudes. When asked whether I believe that P, I take responsibility if I reflect on the reasons I take myself to have, and arrive at my own conclusion after deliberation. As a result, avowing the belief that P expresses my endorsement of P and my commitment to the truth of P. As soon as I reconsider and start doubting P's truth, the avowal ceases to exist (Moran 2001, 74–77, 80–82).

under control of the self-ascribing agent. Yet it is precisely vulnerability to such subliminal influences that trick subjects in confabulation studies into self-attributing confabulated mental states. In Nisbett and Wilson's original experiment, subjects were led to self-attribute a judgment about the pantyhose they chose (e.g., that it had a superior quality) that they did not make prior to their choice (cf. Carruthers 2011). Or consider the experiment by Schnall et al. (2008) where subjects who judged a vignette sitting behind a dirty, disgust-inducing desk gave significantly higher moral wrongness ratings than subjects who judged the same vignette sitting behind a clean, non-disgust-inducing desk. The state of the desk unconsciously influenced their moral reasoning and judgment.

These and other studies are designed to manipulate the process of deliberative avowal. If we understand confabulation (and self-interpretation in general) as a form of first-person mindreading, as we argued above, then confabulation can be said to undermine first-person authority because it misrepresents the actual state of one's mind at some relevant time in the past. However, on a mindshaping view, it could be argued that first-person authority need not be undermined once the relevant mental states have been avowed. This is because the adequacy of mental state attribution does not only depend on the correct representation of one's past mental state, but also on the extent to which it shapes one's future behaviour.. On a mindshaping view of self-attribution, first-person authority is partly determined by the self-ascriber's ability to align her deliberative judgments with the long-term dispositional states that cause her future behaviour. On Moran's account, however, the only authoritative way an agent is able bridge this gap between judgments and corresponding dispositional beliefs is by means of deliberative avowal. But in many everyday cases, this is not enough to keep one's dispositions in check.

Consider the following example. You deliberatively avow that it is more important to spend more time with your kids than to be promoted at work. Yet your daily habits turn out to be stronger than your judgment, to the effect that you keep making long hours at work. Your spouse starts to confront you with this. In response, you repeatedly go through the same deliberative process, reaching the same transparent and rational conclusion that you should spend more time with your kids. There comes a point, however, when your spouse stops listening to your avowals. S/he wants you to start regulating yourself, not through rational deliberation, but by paying attention to the factors that influence your working behavior beyond the scope of your transparent rational self-inquiry and by taking the necessary precautions to diminish their influence.

According to Moran, such self-regulation techniques count as external manipulations because they are not constitutive of deliberative processes in and of themselves. They involve exercising a kind of control over one's attitudes that is "not the expression of 'activity' relevant to autonomy or rational authority. In such cases of producing a desire [or belief] in oneself, the resulting attitude is still one I am essentially passive with respect to. It is inflicted on me, even if I am the one inflicting it." (2001, 117) An agent who adopts an instrumental perspective on his own mental states from a non-transparent, third-person perspective becomes 'alienated' from himself: insofar as he applies such techniques, he ceases to function as a rational being who 'declare[s] the authority of reason over his beliefs and his actions' (2001, 127).

We think that Moran creates a false dichotomy here. There appears to be a crucial difference between producing a mental state in oneself by, for example, taking drugs on the one hand and, say, cognitive rehearsal or seeking the support of friends, on the other. Whereas it could be argued that surrendering to one's drug addiction when feeling depressed is a case of not taking responsibility for one's mental life, this conclusion seems implausible when one decides to go jogging to clear one's head, or call a friend to make one feel better. The difference seems to be that in the latter case one *decides* to perform these self-management techniques in order to feel

better, on the basis of good reasons, whereas in the former one merely surrenders to one's craving. Such decision might take the form of deliberative avowal, but, crucially, such avowal would be in the service of instrumental self-regulation, which itself is not deliberative in nature.

Our rational capacities are fragile, and our inclinations to transparently avow from the stance of deliberative spontaneity are easily swayed by seemingly trivial and non-rational cues, as the confabulation studies demonstrate. Moreover, deliberative avowal is not the royal road to making one's dispositions line up with one's rational commitments. As our example illustrated, it can be deceivingly unreliable as means of shaping one's future dispositions. We conclude that Moran's construal of first-person mindshaping as deliberative avowal has little resources to mitigate the undermining effect of confabulation on the authority of mental state self-ascription. In the next section, we propose to broaden the scope of first-person mindshaping by introducing McGeer's (2008) account of self-regulation.

5 First-Person Mindshaping as Self-Regulation

McGeer (2008, see also McGeer 1996) starts from the observation that, in folk psychological practice, we are granted first-person authority insofar as we are able to live up to the expectations licensed by our self-ascriptions. (cf. Brandom 1994; Morton 2003; Zawidzki 2013) Deliberative avowal is indeed one important means by which we shape our attitudes into agreement with our commitments. But it is by no means the only way to authorize our self-ascriptions. According to McGeer, it is also a central assumption of folk psychology that we know our deliberative powers to be fragile, all too easily hijacked by our more subversive inclinations. Hence, she argues, social and moral development does not only aim at enhancing children's capacity for deliberative avowal, but also at teaching them how to regulate their inclinations (not) to avow. This may take the form of developing self-governing habits of mind, whereby we use our intentional self-ascriptions as reminders to reinforce inclinations that are compatible with these ascriptions and to disempower inclinations that are not. Or consider more explicit and effortful measures, whereby we try to steer ourselves into (or out of) a position in which deliberative avowal comes easiest, amplifying our deliberative powers in moments of boredom or distraction, or rather side-tracking them when under great stress.

Self-regulative authoritative agency implies that one is continually ready to take a step back from one's transparent deliberative powers in order to reflect on them from an allocentric or opaque point of view on oneself. By adopting such an instrumental stance towards one's rational agency, one makes room for an assessment of the way one's moods, appetites, current (social) environment, etc., shape, and potentially bias, one's deliberative judgments.

As an example, take Sarah who recently took the Implicit Association Test (IAT).[2] 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 IAT might proof otherwise and show that she

[2] The IAT asks 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

implicitly believes that African American people are less intelligent then Caucasian people.[3] 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. In such a case, we deny Sarah full first-person authority with respect to her 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.

In order to be granted first-person authority regarding her explicitly avowed belief that all races are of equal intelligence, we would demand that Sarah start systematically evaluating her own behavior and make attempts to change her implicit dispositional beliefs to the contrary, e.g., via changing her environment or job application procedures. We would demand such *non*-deliberative self-regulative techniques, precisely because we take her implicit bias to be fed by all kinds of subliminal factors that *cannot* be controlled by her deliberative powers. By exercising this kind of instrumental self-directed agency, we would therefore not be led to think that she is somehow *alienated* toward herself, acting in 'bad faith' or avoiding responsibility (cf. Moran). Quite to the contrary, we would consider such self-regulative practices to enhance her rational autonomy, and perhaps even the only means of restoring her status as authoritative self-ascriber of her explicitly avowed belief. (see also De Bruin et al. 2014)

Both McGeer and Moran would agree that the confabulation data need not automatically imply a diminishing of our status of authoritative self-ascribers of mental states. Such authority is not only concerned with seeking out the processes that presumably lead to the states ascribed, but also, and more importantly, with taking responsibility so as to ensure that their ascription will be vindicated in the future. Yet while Moran's account has only limited options to help us protect ourselves against subliminal influences on our reasoning, McGeer's model suggests more fruitful ways of responding to the challenges posed by the confabulation studies. The data point toward blind spots in our transparent deliberative point of view. Knowing this, we can devise self-regulative practices aimed at minimizing their influence. Ideally, such self-regulative techniques would be specifically directed at the subliminal factors identified in the confabulation studies. For example, armed with the knowledge that dirty desks tend to lead to more severe judgments of moral transgressions in vignettes, legal practice could implement a rule that proscribes the state of judge's desks (rooms, etc.) when deciding on legal verdicts. More in general, however, the knowledge that our deliberative powers are unconsciously influenced by a myriad of non-rational factors, could motivate us to adopt a lifestyle of self-regulation that increases our capacity to become aware of them and thereby to enhance rather then reduce our status as authoritative self-interpreters.

Our question was how to understand the implications of confabulation on the authoritative status of mental state self-attribution. We have argued that McGeer's account of first-person mindshaping as self-regulation suggests that, to a certain degree, we can protect ourselves against subliminal cues that cause us to confabulate. Indirectly, implementing self-regulative procedures can thus help maintaining our status as authoritative mental states self-ascribers. However, can first-person mindshaping also affect the authority of confabulations *directly*? What does a mindshaping account imply for the first-person authority regarding *the confabulated state itself*?

As a thought experiment, consider the following scenario based on Nisbett and Wilson's (1977) experiment mentioned in the introduction. Suppose, for the sake of the argument, that the subjects' choice was indeed determined by the relative position of the preferred pair (even under experimental conditions in which the pairs of pantyhose were *not* identical, but actually

[3] 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.

differed in consistency, knitting pattern, etc.). Now imagine one particular subject, who indeed erroneously took her preference for the far right pair to be the result of its superior knitting pattern, but who *starts living up to* her confabulation by buying only panty-hoses with this knitting pattern, recommending it to friends, etc. When confronted with this sequence of events, i.e., her misinterpretation followed by her change in preference, our subject responds as follows. She had never thought about her pantyhose preference before participating in the study. It was only then that she noticed the beautiful knitting pattern of the far left pair. It stayed with her, and every time she had to buy a new pair, the memory of the experiment and the knitting pattern came back to her, making it all too easy for her to choose between the pairs displayed in the store where she bought them.

Now would we judge our subject to be self-deceived? That is: *at the time of the experiment*. Would we still withhold from her first-person authority regarding her judgment that she prefers pantyhose with that particular knitting pattern? Or would we perhaps come to the conclusion that she did in fact, at the time of the experiment, make that judgment, but that her self-attribution at that time, though correct, was based on *the wrong evidence*. From this perspective, the mistake was to think that her judgment emanated from and was justified by mental processes *leading up to it*. In this sense she was deceived, for her judgment was in fact caused by the position of the preferred pair, not by its superior knit. Still, her self-ascription of the judgment need not be put into question. For she has made it true ever since, and she has therefore earned first-person authority with respect to it from that time onward.

On a mindshaping account, being mistaken about the psychological processes that lead up to one's judgments might not be as detrimental to first-person authority as it might appear on accounts that model self-interpretation exclusively in terms of mindreading. As long as one is right about the consequences of one's self-ascription of a mental state, such 'maker's knowledge' might save part of one's status as an authoritative self-ascriber of that same state.

6 Confabulation Reconsidered

Let it be clear that we do not want to suggest a mere 'future-directed' understanding of confabulation. The adequacy of self-interpretation, and hence the *in*adequacy of confabulation, depends on one's past, present and future experiences and (intentional) behavior. The future-directed aspect of self-interpretation, highlighted by the mindshaping view, has so far received little attention in confabulation research, however.

In the philosophical and psychological literature, confabulation is characterized as a 'backward-looking' phenomenon: we are asked to represent the cause of a past action, and we are said to confabulate when we (mistakenly) self-attribute a previous judgment (or another mental state) as the cause of this action. Consider our thought experiment once more. From a mindreading perspective, the exact time of the subject's judgment is of primary importance in order to judge whether or not our subject fell victim to confabulation. If it is true that her judgment about her preference in knitting patterns did not occur prior, if only milliseconds, to her choosing the pair on the far left, then she must have confabulated her reasons for choosing that pair.

Starting from a mindshaping view, however, confabulation might also be cast as in part being a 'forward-looking' phenomenon: we can then be said to confabulate when, for example, we attribute a judgment that p to ourselves, but are unable to self-regulate ourselves such that our future behavior aligns with this judgment, i.e., that we acquire the dispositional state of believing that p. If confabulation, in the most general sense of the term, is about concocting, without intent, mental states that we are not entitled to ascribe to ourselves, then, from a

mindshaping perspective on first-person authority, the issue of entitlement should be decided in light of one's past *and* one's future behavior and circumstances.[4]

This comes out most clearly in clinical cases of confabulation. Consider patients with Korsakoff's syndrome. Due to severe amnesia, these patients make up their reasons as they go along, reasons which are clearly at odds with their actual past, and which, moreover, they are incapable of living up to in the future. Part of the problem, however, is that these patients not only suffer from amnesia, but also from an almost complete *lack of insight* into their illness. As a result, these patients do not take the necessary precautionary measures to protect themselves against their amnesia. They do not keep journals, neither do they recruit others to help them compensate for their memory loss. To a significant extent, their lack of insight consists in their failure to recognize that, due to their amnesia, they are in desperate need of drastic forms of self-regulation. When understood in this way, what causes their confabulated answers to have such stultifying effect on us, is that these answers are generated in complete *ignorance* of or even a *lack of interest* in their severely disabled capacity for self-directed agency. An even more radical example can be taken from the literature on *anosognosia for hemiplegia*, a condition in which a patient suffers from a certain disability but denies that this is so. Ramachandran (1995) describes the case of a 76-year-old lady who had a stroke that left her completely paralyzed on the left side. 'Mrs. M', as Ramachandran calls her, persistently denied her paralysis even when pressed after clearly failing to lift her left arm. Strikingly, her answers were without any hesitation or lack in conviction.

Clinical cases suggest that confabulation is a more complex phenomenon than as usually presented in philosophical and psychological literature. Authoritative self-ascription of mental states not only depends on mindreading capacities directed at the past. It also ties into our ability to shape our mental states with an eye towards the future and our concern to nourish and preserve this ability. Part of what makes clinical cases of confabulation so interesting, and perhaps also particularly pathological, is the disabilities they show in precisely this respect.

7 Conclusion

In this paper we have explored the implications of a view of self-interpretation as first-person mindshaping. On this view, the adequacy of self-interpretation is not merely determined by the extent to which one is able to represent one's past mental states, but also by one's future behaviour, i.e., to what extent one is able to live up to the interpretation one has given. We have argued that this mindshaping view, in particular McGeer's account of self-regulation, implies a different understanding of how confabulation affects first-person authority of mental state self-attribution. By broadening the notion of first-person authority to include future commitments entailed by self-attributed states, the mindshaping view implies that one can partly restore first-person authority of confabulated self-attributions. At the same time, however, it broadens our understanding of the phenomenon of confabulation itself. The mindshaping view of self-interpretation suggests that confabulation can also consist in a failure to shape one's behaviour in accordance with self-attributed mental states. These future-directed aspects of confabulation raise a number of interesting questions for further philosophical reflection and empirical research.

[4] From a mindshaping perspective, it could also be argued that the subject's judgment in our thought experiment was not an example of confabulation, but rather of self-constitution. On a self-regulation account, the exact time of the judgment during the experiment need not be very relevant. Our subject might have stumbled upon her preference at a time she is not fully aware of in a way she doesn't fully understand, *yet it is still true* that she chose the left pair *because* she favored the particular knitting pattern. This is the reason for choosing pairs she *constituted* during the experiment.

Acknowledgments We would like to thank 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.

References

Andrews K (2012) Why do apes read minds? Toward a new folk psychology. MIT Press, Cambridge
Brandom R (1994) Making it explicit. Harvard University Press
Carruthers P (2009) How we know our own minds: the relationship between mindreading and metacognition. Behav Brain Sci 32:121–181
Carruthers P (2011) The opacity of mind. Oxford University Press
Coltheart M, Turner M (2009) Confabulation and delusion. In: Hirstein W (ed) Confabulation: views from neuroscience, psychiatry, psychology and philosophy. Oxford University Press, Oxford
De Bruin LC, Jongepier F, Strijbos DW (2014) Mental agency as self-regulation. Rev Philos Psychol. doi:10.1007/2Fs13164-014-0190-7
Dennett DC (1987) The intentional stance. MIT Press, Cambridge
Gallagher S (2012) In defense of phenomenological approaches to social cognition: interacting with the critics. Rev Philos Psychol 3(2):187–212
Gazzaniga M (1998) The mind's past. California University Press
Geschwind N (1965) Disconnection syndromes in animals and man. Brain 88(237–294):585–644
Haidt J (2001) The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychol Rev 108:814–834
Hieronymi P (2009) Two kinds of agency. In: O'Brien L, Soteriou M (eds) Mental actions. OUP, Oxford, pp 138–162
Hirstein W (2009) Confabulation: views from neuroscience, psychiatry, psychology and philosophy. Oxford University Press, Oxford
Hutto DD (2008) Folk psychological narratives: the sociocultural basis of understanding reasons. MIT Press, Cambridge
Hutto DD, Rattcliffe M (eds.) (2007) Folk-psychology re-assessed. Springer, New York
Kraepelin E (1913) Lectures on clinical psychiatry, 3rd edn. William Wood & Company, New York
McGeer V (1996) Is "Self-Knowledge" an empirical problem? Renegotiating the space of philosophical explanation. J Philos 93(10):483–515
McGeer V (2007) The regulative dimension of folk psychology. In: Hutto D, Ratcliffe M (eds) Folk-psychology re-assessed. Springer, New York
McGeer V (2008) The moral development of first-person authority. Eur J Philos 16(1):81–108
Moran R (2001) Authority and estrangement. Princeton University Press
Morton A (2003) The importance of being understood: folk psychology as ethics. Routledge
Nisbett R, Wilson T (1977) Telling more than we can know. Psychol Rev 84:231–295
Ramachandran VS (1995) Anosognosia in parietal lobe syndrome. Conscious Cogn 4:22–51
Schnall S, Haidt J, Clore G, Jordan A (2008) Disgust as embodied moral judgment. Personality Soc Pathol Bull 34:1096–1109
Wegner D (2002) The illusion of conscious will. MIT Press
Wheatley T (2009) Everyday Confabulation. In: Hirstein W (ed) Confabulation: views from neuroscience, psychiatry, psychology and philosophy. Oxford University Press, Oxford
Wilson T (2002) Strangers to ourselves. Harvard University Press
Wing JK, Cooper JE, Sartorius N (1974) Classification of psychiatric symptoms. Cambridge University Press, Cambridge
Zawidzki T (2013) Mindshaping: the linchpin of the human socio-cognitive syndrome. MIT Press