Topoi (2020) 39:151–161
https://doi.org/10.1007/s11245-019-09631-y

Does Confabulation Pose a Threat to First-Person Authority? Mindshaping, Self-Regulation and the Importance of Self-Know-How

Leon de Bruin[1] · **Derek Strijbos**[1,2]

Published online: 9 February 2019

Abstract
Empirical evidence suggests that people often confabulate when they are asked about their choices or reasons for action. The implications of these studies are the topic of intense debate in philosophy and the cognitive sciences. An important question in this debate is whether the confabulation studies pose a serious threat to the possibility of self-knowledge. In this paper we are not primarily interested in the consequences of confabulation for self-knowledge. Instead, we focus on a different issue: what confabulation implies for the special status of self-attributions, i.e. *first-person authority* (FPA). In the first part of the paper, we propose that FPA is based on a capacity for *self-regulation*. Accordingly, FPA depends on the extent to which we are able to bridge the gap between our sayings and doings by aligning our actions with our avowed self-ascriptions and vice versa. FPA is withheld when we (systematically) fail at such re-alignment. In the second part of the paper, we contrast our view with the accounts of Scaife (Acta Anal 29:469–485, 2014) and Bortolotti (Rev Philos Psychol 9(2):227–249, 2018). We claim (contra Scaife) that the apparent fact that we cannot reliably distinguish, from a first-person perspective, when we are confabulating and when we are not, does not necessarily undermine FPA. We argue (contra Bortolotti) that a systematic failure to align our actions with our self-ascriptions and vice-versa is a genuine threat to FPA. In the last part of the paper, we introduce the concept of *self-know-how*—the know-how embodied in the way one is disposed to relate to oneself in making sense of oneself with or in the face of others—and briefly explored the importance of diminished or absent self-know-how in clinical cases.

Keywords Confabulation · Self-knowledge · First-person authority · Self-regulation · Mindshaping · Self-know-how

1 Introduction

Empirical evidence suggests that people often confabulate when they are asked about their attitudes, choices or actions. Confabulation, as Coltheart and Turner (2009, p. 180) explain, happens "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."

The implications of these studies are the topic of intense debate in philosophy and the cognitive sciences. Some philosophers, such as Carruthers (2009, 2011) and more recently, Scaife (2014), have argued that confabulation studies pose a serious threat to the possibility of self-knowledge (SK). Since we cannot reliably distinguish, from a first-person perspective, when we are confabulating and when we are not, we can never be sure that SK is accurate. Philosophers such as Bortolotti (2018), on the other hand, claim that confabulation studies do not necessarily pose a threat to SK. Confabulation studies only show that people are blind to the processes that lead to their attitudes and choices, but not to the actual attitudes and choices themselves.

Despite their differences, the accounts of Scaife and Bortolotti are both committed to an epistemic view of SK and confabulation in terms of *mindreading*. According to this

✉ Leon de Bruin
l.debruin@ftr.ru.nl

[1] Department of Philosophy, Radboud University Nijmegen, Erasmusplein 1, Postbus 9103, 6500 HD Nijmegen, The Netherlands

[2] Dimence Mental Health Care, Grasdorpstraat 6, 8012 EN Zwolle, The Netherlands

view, SK depends on the ability to correctly represent one's state of mind. The confabulation studies are potentially problematic because they seem to show that subjects are unable to do this—they fail to correctly represent their state of mind at some relevant time prior to the confabulatory response.

In this paper we are not primarily interested in the consequences of confabulation for SK. Instead, we focus on a different issue: what confabulation implies for the *special status of self-attributions*. Self-attributions of mental states are generally considered to have 'first-person authority' (FPA)—a special authoritative status compared to the attribution of mental states to others.

One might assume that there is a direct link between SK and FPA, and that failures of self-knowledge are detrimental to the authoritative status of self-ascriptions. Given this assumption, one might conclude that the confabulation studies, *if* they do indeed show that subjects lack self-knowledge, also undermine our everyday notion of FPA.

Our main aim in this paper is to show that this conclusion is unfounded if we adopt a broader and richer notion of FPA. In the next section we will present such a notion. Whereas SK is frequently understood in epistemic terms,[1] we take FPA to be an essentially social, pragmatic phenomenon. FPA, on our view, depends on a capacity for *self-regulation*: it depends on the extent to which we are able to bridge the gap between our sayings and doings by aligning our actions with our avowed self-ascriptions and vice-versa. FPA is withheld when we (systematically) fail at such re-alignment.

In Sect. 3, we discuss our view of FPA in relation to Scaife's account of the confabulation studies. Although we agree with Scaife that the results of these studies (if they do indeed generalize) might be indicative of failures of SK, we will show that this does not automatically lead to problems with FPA. According to the self-regulation view, although failures of SK typically create a gap between our sayings and doings, they do not necessarily undermine the authoritative status of our self-ascriptions. This is because, as we will explain, self-ascriptions also have a 'forward looking' function insofar as they help us to align our future behavior with our self-ascribed mental states. Furthermore, we will show how an account of FPA in terms of self-regulation is able to address Scaife's skeptical challenge of how we can distinguish whether we are confabulating or not.

In Sect. 4 we focus on Bortolotti's analysis. Bortolotti acknowledges that the explanations offered by subjects in confabulation studies can be ill-grounded and the result of ignorance, but she still thinks they can be authentic insofar as they are sincerely reported and genuinely endorsed. However, we will argue that even thought this argument might support and safeguard a specific understanding of SK, it is insufficient to mitigate the impact of confabulation on FPA.

In Sect. 5, we discuss the folk-psychological norms for self-regulation and propose a distinction between the 'know that' of explicit self-ascriptions and the skill set or 'know-how' of adopting certain self-directed attitudes in the process of self-regulation. We suggest that the 'know-how' of self-regulation plays an important role in achieving, maintaining or restoring the status of authoritative self-interpreters in folk-psychological practice.

2 Confabulation as a Problem of Self-Regulation

Confabulation is typically understood as a 'backward looking' phenomenon: when giving an answer to some question regarding their behavior, confabulating subjects unintentionally misrepresent (the causes of) the mental states and attitudes leading to the behavior in question.

In previous work (e.g., De Bruin et al. 2014; Strijbos and De Bruin 2015; De Bruin 2016) we explored an understanding of confabulation as also having a 'forward looking' dimension. Accordingly, the significance of confabulation is not only determined by its inaccurate rendering of the causes of the behavior in question, but also by its failure to play a role in regulating the subject's *future* behavior.

This idea was inspired by clinical cases of confabulation, where patients often suffer from an accompanying lack of insight into their condition. Confabulating patients, such as split-brain patients, or patients suffering from Korsakoff's syndrome, often display indifference towards their apparent problems in self-interpretation. They usually do not attempt to compensate for their disability, e.g., by building in (socially extended) control mechanisms, such as using mnemonic devices, checking with others or asking others to correct them if necessary. We developed an account that sought to conceptualize this lack of self-regulation as a feature of confabulation itself, rather than as an accompanying symptom.

The accepted view of confabulation as an exclusively backward-looking phenomenon has its roots in the debate on folk psychology, where the consensus still holds that getting along in social practice is primarily about *mindreading*, i.e. explaining and predicting each other's behavior by inferring the mental states that caused it. Successfully getting along then implies accurately representing the causes of other another person's behavior.[2] Over the last two decade

[1] Although this is not entirely uncontroversial, see for example the agential accounts of self-knowledge proposed by Moran (2001) and Bilgrami (2006).

[2] This mindreading view of folk psychology is at the core of both the Theory Theory and the Simulation Theory.

or so, however, the mindreading approach to folk psychology has been under attack. Various arguments have been put forward in defense of the view that mental state attribution in everyday social practice is about *mindshaping*. According to this view, we are not primarily in the business of passively reading the mental states of others in order to predict or explain their behavior. Rather, we are being socialized in a community held together by social pressures to *make* behavior understandable. That is, we modify our own minds and those of others in accordance with the norms and normative expectations embodied in our community. This is effectuated by means of 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, the mindshaping view argues that folk psychological explications primarily serve to teach children what to think and how to act under certain circumstances, thereby regulating their thoughts and behavior so as to match socio-cultural norms. Correspondingly, in everyday practice folk psychology has the important function of normalizing, correcting and justifying behavior that is seemingly out of line, and by doing so re-enforcing the norms by which we judge each other's behavior. The mindshaping view does not deny that mindreading occurs in human social practices. Rather, it argues that our folk psychological competence to explain and predict behavior by mental state attribution should be explained within the (phylogenetic, ontogenetic and everyday practical) context of mindshaping.

On a mindreading view, the adequacy of the attribution of a mental state M in order to explain some past behavior B, depends on whether M was in fact instantiated some time prior to B and, furthermore, whether M played a causal role in producing B. On a mindshaping view, by contrast, the adequacy of ascribing M is measured by the extent to which the ascription of M helps to keep or bring (back) the interpretative target's behavior within the realm of expectable and admissible responses. Thus, on a mindshaping view, there is a 'backward looking' and a 'forward looking' element to the success of mental state attribution. By ascribing M, the past behavior B can be re-interpreted from within the culturally sanctioned repertoire of behavioral responses dictated by our folk psychology. Thus, 'looking back' the attribution of M might have a mitigating effect on the interpretative target's behavior, so as to make it more understandable and compatible with socio-cultural norms, and the target's status as a participant in folk psychological practice more reliable. At the same time, the ascription of M has a forward-looking function, in that it re-establishes these norms how to respond in certain circumstances and in this way plays a causal role in shaping the interpretative target's *future* behavior in alignment with the interpretation in terms of M.

Applying this regulating function of folk psychology to *self*-ascription of mental states, McGeer (2008) starts from the observation that, in folk psychological practice, FPA is determined to a significant extent by our ability to *live up to* the expectations licensed by our self-ascriptions. (cf. Brandom 1994; Morton 2003; Zawidzki 2013). In other words, we are able to *make* our self-ascriptions true by aligning our sayings and doings with the commitments we undertake by uttering them. This is what McGeer terms 'self-regulation'. The point is not that there is no backward-looking dimension to self-attribution in folk psychological practice. Of course there is: we do have an interest in making our past behavior understandable or permissible by self-ascribing reasons. The point is rather that this enterprise usually takes place with an orientation towards the future, and that FPA regarding these self-ascribed reasons is also dependent on our capacity to regulate our future behavior in light of these reasons.

Building on these insights, we proposed an account of FPA in terms of self-regulation: FPA depends on the extent to which we are able to bridge the gap between our sayings and doings by aligning our actions with our self-ascriptions and *vice-versa*. By contrast, FPA is withheld when we (systematically) fail at such re-alignment. We suggested that in everyday folk psychological practice, people hold each other accountable for bridging the gap by means of both 'upward' and 'downward' self-regulation. Upward self-regulation is the process of bridging the gap between saying and doing by adjusting one's actions to one's self-ascriptions; downward self-regulation concerns the alignment of one's self-ascriptions to one's actions, i.e. revising our beliefs about the causes and reasons for our actions.

Analyzing FPA from a mindshaping perspective has important implications. First of all, a mindshaping perspective suggests that FPA should be seen as the exercise of a capacity or a 'skillful activity', rather than the property of a stand-alone self-ascription of a specific mental state (more on this in Sect. 5). Furthermore, since FPA is taken to depend on the extent to which we succeed in aligning our sayings with our doings, it suggests that the adequacy of our mental state self-ascriptions is always assessed in relation to our past and/or future actions.

Another implication of analyzing FPA from a mindshaping perspective is that there seems to be no strict dividing line between FPA as relating to particular self-ascriptions and FPA as relating to the person—one's status as a reliable, trustworthy self-ascriber. According to the mindshaping account of folk psychology, the adequacy of self-ascription is measured in social practice by the extent to which it helps to keep or bring (back) the interpretative target's mental states and behavior within the realm of expectable and admissible responses. In other words, the perceived

success of particular self-ascriptions is determined against the background of the perceived success of the overreaching activity of making oneself understood in the social situation one faces, i.e. against the background of one's status as a reliable self-ascriber under those circumstances.[3]

Thus, FPA of particular self-ascriptions and FPA-status appear to be interdependent. One's status as an authoritative self-ascriber depends, inter alia, on the acceptability of particular self-ascriptions given in the past and at present, and vice versa. A history of unreliable self-ascriptions tends to have a negative effect on one's overall FPA-status. Conversely, having high FPA-status will likely have a positive effect on the authority of particular self-ascriptions. Moreover, this not an all-or-nothing affair. First, interpretation of FPA has a hermeneutical dynamic to it, in the sense that the *prima facie* authority regarding particular self-ascriptions may change through time under the influence of one's interpretation of the overall FPA-status of the self-ascriber. Conversely, one's assessment of FPA-status may likewise change depending on how one judges (a series of) self-ascriptions. We will come back to this point in Sect. 4. Second, people may be viewed by others (who know them well) as generally reliable in their self-attributions, but particularly *unreliable* under certain specific circumstances (e.g., when under stress, drunk, or trying to impress), or relative to a certain range of self-ascriptions, relating to certain domains of mental states or aspects of the self. The latter is particularly striking in cases of clinical confabulation (e.g. cases of anosognosia).

What this shows is that FPA is a complex phenomenon: the adequacy of a particular self-ascription cannot be determined in isolation, but (i) depends on one's past and (expected) future behavior, (ii) should be understood as the exercise of a skill rather than the property of a single mental state self-attribution, (iii) is assessed in the light of one's overall FPA-status as a person with a history of more or less reliably self-ascriptions, (iv) is sensitive to the specific context of self-ascriptions, and (v) is relative to a certain range of self-ascriptions.

With the basics of the self-regulation view of FPA in place, we will now discuss its implications in relation to the accounts proposed by Scaife and Bortolotti.

[3] This rendering of FPA does not follow from a mindreading view of self-attribution, according to which the success of a particular self-ascription is determined by whether or not the attributed state was in fact instantiated at the implied time. From a mindreading perspective, integration of normative dimensions pertaining to justification, normalization, etc. in social settings that affect the success of self-attributions, is ad hoc (and on most mindreading accounts is considered to occur post hoc).

3 Addressing the Skeptical Challenge

Scaife (2014) has argued that the empirical evidence on confabulation leads to skepticism about SK. His argument goes as follows:

1. There are cases where people generate inaccurate information about their own decision-making (e.g. confabulation cases).
2. People cannot reliably differentiate between accurate and inaccurate information about their own decision-making from the first-person perspective.
3. Therefore, if we only have evidence from the first-person perspective, then we can never be certain that our self-knowledge is accurate.

Scaife considers two objections to this line of argument. First, he examines the 'markers of certainty' objection, which claims that SK comes with some dependable markers of certainty and that we should therefore be able to determine whether someone is confabulating or not. This objection has been advanced by Fiala and Nichols (2009), who state that there are systematic differences in the confidence levels between confabulation and veridical self-attribution.

However, Scaife argues that research has shown that subjects themselves cannot tell when they are confabulating. In a series of choice-blindness experiments, Johansson et al. (2005, 2006) checked for various markers of certainty (response time, length of statement, word frequency checks for markers of certainty, unfilled pauses, laughter, the use of the past vs. present tense, the use of first vs. third person, emotional content, word length) and found no evidence that participants were less confident about their confabulations. On the basis of this research, Scaife concludes that, from the first-person perspective, instances of confabulation are indistinguishable from accurate self-ascription.

The second objection is basically to acknowledge this, but to argue that we do not need to consider this kind of skepticism as a serious threat to SK. Take the skeptical hypothesis that I am a brain in a vat. Even though I cannot prove that I am not a brain-a-vat, neither is there any evidence which counts in favor of this conclusion. According to Scaife, however, the difference is that skepticism about SK is *empirically motivated*—there is (allegedly) a substantial body of empirical evidence demonstrating that confabulation does occur.

What are the implications of Scaife's skeptical argument for FPA? As we explained in the previous section, on the self-regulation view, FPA hinges on the ability to align our self-ascriptions with our behavior, by upward and/or downward self-regulation. What matters *primarily* is not the truthfulness of the story about the causes of one's behavior, but

the alignment itself. In other words, what matters primarily is not whether one's story about the causes of one's decision-making processes 'is true', but also whether one is able to 'make it true'.

Roughly, this means that one should aim at rendering one's self-ascriptions (past, present, and future) more-or-less compatible with the commitments undertaken by one's behavior (past, present and future). So the question to ask on the self-regulation view is the following: do the empirical findings from the confabulation studies give us reason to accept the conclusion that, whenever we consider our motivations in everyday life situations, we should be skeptical about our ability to align our self-ascribed reasons with our behavior?

Consider Nisbett and Wilson's (1977) experiment, which is generally taken to show that subjects confabulate by self-ascribing a preference as the cause of their choice for a pair of pantyhose, while this choice is in fact determined by the relative position of the preferred pair. We believe that the findings from this experiment first and foremost demonstrate that subjects in the experiment *are ignorant of the gap between* their actions and their self-ascriptions. The experiment was not designed to address the question whether and to what extent the subjects are actually able and willing to bridge this gap. Suppose the subjects would accept the alternative explanation offered by the experimenters as the real cause of their choice. What would happen if the same subjects would again participate in the same experiment? We think it is likely that they will attribute the cause of their choice to the position effect. This would effectively show that they are capable of downward self-regulation, i.e. that they can bridge the gap by aligning their self-ascriptions with their actions. The point is that this would restore their FPA in normal social settings. Understood as a (socially and temporally) extended process of self-regulation, this behavior would differ remarkably from the behavior displayed in clinical cases of confabulation. The significant feature of clinical confabulation is that patients lack this capacity for self-regulation. They lack FPA because they are unable to restore the fit between self-ascriptions and behavior.

On the self-regulation view, then, we can meet Scaife's skeptical challenge in the following way. First, we can point out that, although there is a substantial body of empirical evidence demonstrating that confabulation does occur, this evidence is based on a narrow concept of confabulation as inaccurate mindreading. As we mentioned above, the experiments on confabulation are not designed to address a capacity for self-regulation, and as such do not give us reason to be skeptical about our ability to align our self-ascribed reasons with our behavior. Second, on the self-regulation view, the apparent fact that we cannot reliably distinguish, from a first-person perspective, when we are confabulating and when we are not, does not necessarily undermine FPA. This is because self-regulation in folk psychological practice is essentially a *second-person activity*. The (re-)alignment of our self-ascriptions with their actions and vice-versa need not be and is not a solipsistic enterprise. In the game of giving and asking for reasons, the second-person perspective may come to the rescue where the first-person perspective stops being a reliable guide. Others may confront us with inconsistencies between our self-ascriptions and our behavior and point us towards subliminal factors influencing our decision-making.

But do people actually display this kind of self-regulative behavior? Some confabulation studies show that subjects are in fact resistant to alternative explanations of the causes of their choices. Participants often express surprise or even disbelief when being informed about the nature of the experiments. Thus, Johansson et al. (2008) report that the debriefing after a choice-blindness experiment involved[4]:

> "asking the participants a series of questions, the last one being if they thought they would have noticed if a switch had been made during a 'similar' experiment. Of the participants that did not notice any manipulations during the experiment, 85% believed that they would have detected such a switch if it had been performed. When the actual purpose of the experiment was finally revealed, the participants showed considerable surprise, and sometimes even questioned our claim that we had switched the pictures." (p. 151)

On our view, these findings suggest that people are often ignorant of the influence of subliminal factors on their judgments and decisions, and that studies teasing out these factors are genuinely revealing from a folk-psychological point of view. People tend to be epistemically conservative when being offered explanations that do not square with folk-psychological consensus. This would explain the surprise and disbelief of the participants in this study. We would not be surprised, however, if the subjects would revise their beliefs when being offered a video or a live demonstration of how the experiment was actually carried out. In the face of this evidence, folk psychological norms would dictate such downward self-regulation (i.e., revision of their beliefs).

When we consider the implications of the confabulation studies on FPA from a self-regulation perspective, Scaife's skeptical question about our epistemic abilities has to be reformulated as a question regarding our self-regulation abilities. This reformulation does not fully dissolve the skeptical worry, but reframes it in terms of different empirical

[4] In this experiment participants had to choose which one of two abstract patterns they found most appealing. Subsequently, the patterns were switched and the participants were asked to indicate which one of the two patterns they had previously found most appealing.

questions, such as: is our folk psychology in fact critical of the stories we weave around the subliminal causes of our behavior? Two considerations seem to be pointing in different directions here. One the one hand, systematic and profound inconsistency between one's self-ascriptions and one's (future) behavior will evoke critical responses by others. And although in theory, it may be possible to be systematically wrong about the causes of one's decisions and right about its consequences, we think this is empirically highly implausible. That is, living up to one's self-ascriptions requires deep knowledge about one's dispositions—what causes one to behave in certain ways in certain situations, and what adjustments are to be made to shape one's dispositions in the favored or normatively required direction. So if Scaife is right about the unreliability of our understanding of the causes of our own behavior, this epistemic shortcoming will probably have a negative effect on our ability to self-regulate. And this, in turn, will evoke extended social self-regulation mechanisms to come in to play when we become too unreliable to be counted on by the people around us.[5]

On the other hand, however, it should be noted that the accuracy of self-ascriptions (i.e., the degree to which self-ascriptions reliably track the causes of one's behavior and the consequences of these ascriptions) is not only judged on empirical grounds in folk psychological practice. The folk-psychological norm for self-regulation is not perfect alignment (see Sect. 5), and the success of our self-attributions is also measured by the extent to which they enable us to e.g., save face and to maintain or restore social status as trustworthy participants in social practice. These two criteria for success of self-ascription need not always converge.

It appears to be an open question to what extent folk psychological practice actually adopts a critical stance towards the interpretations we give of our motivations for action. Moreover, folk psychology is not static, but evolves, partly under the influence of scientific insights. We would not be surprised if the influence of subliminal factors on our behavior were to become part of our (scientifically informed) folk-psychology. Presumably, this would have effects on folk psychological norms and on our (socially extended) self-regulation practices.

4 What is Required for FPA?

Bortolotti (2018) argues that, even though the confabulation studies show that people are blind to the processes responsible for their choices, this does not mean that they are also blind to what choices they made. She distinguishes between two necessary features that allow us to identify cases of (non-clinical) confabulation[6]:

1. Ignorance: People ignore some of the key causal factors leading to the formation of their attitudes and choices.
2. Ill-groundedness: People produce ill-grounded claims about the causes of their attitudes and choices.

To tease out the difference between the first and the second feature, Bortolotti provides two interpretations of the Nisbett and Wilson (1977) study. The first interpretation focuses on Roberto, who chooses the rightmost pair of stockings *because of the position effect*. When asked to explain his choice, he answers that he chose that pair because it was the brightest. Roberto confabulates—he forms the belief that the rightmost pair is the brightest as a result of his ill-grounded explanation. His explanation is ill-grounded because he does not mention the role of position effects in his choice. The second interpretation focuses on Sylvia, who chooses the rightmost pair of stockings *because she believes that it is the brightest*. The explanation she offers for her choice is accurate, although her belief about the chosen pair of stockings being the brightest is false. Sylvia does not engage in confabulation, because she offers an explanation for her choice that is well-grounded, even though it is based on a false belief.[7]

Only Roberto is a case of confabulation, according to Bortolotti. Despite the fact that Sylvia's belief was caused by factors of which she was ignorant (the position effect), her explanation is not ill-grounded because her choice was caused by this belief. But even in Roberto's case, which qualifies as a case of confabulation (his explanation is both a result of ignorance and ill-grounded), we have no reason to doubt whether he has SK. According to Bortolotti, Roberto's choice is still authentic, in the sense that it is sincerely reported and genuinely endorsed. Therefore, she concludes: "If successful mental-state self-attributions require awareness of one's attitudes and choices, then they are not threatened by the form of confabulation reviewed here" (2018, p. 236).

[5] We believe that a high degree unreliability in this sense and the lack of effect of (socially extended) self-regulative measures provides the watershed between clinical and non-clinical cases of confabulation.

[6] According to Bortolotti, there is also a third (optional) feature that is quite common in cases of confabulation: (3) Further ill-groundedness: as a result of producing the ill-grounded causal claim, people commit to further beliefs that, even if generally plausible, do not fit the specifics of the situation in which the attitude is formed or the choice is made.

[7] Bortolotti claims that "Sylvia gets the world wrong (the chosen pair of stocking is not the brightest), but she accurately identifies the reasons for her choice." (p. 231) Furthermore, she suggests that Sylvia's belief that the chosen pair of stockings is the brightest might be caused by the position effect: "It is possible that position effects generate a perceptual salience which manifests as brightness for some participants and as softness for other participants. This may give rise to the situation described in Sylvia's case." (ibid.)

There are two issues here that we would like to highlight. First, how can we distinguish between Roberto's case and Sylvia's case, i.e. between an ill-grounded explanation and a well-grounded explanation that is based on a false belief? According to Bortolotti, this is necessary in order to determine whether someone is confabulating or not. Second, it is one thing to argue that SK of one's attitudes and choices only requires an awareness of these attitudes and choices. But what about FPA? Does being ignorant of the processes responsible for these attitudes and choices not have any impact on one's FPA whatsoever?

On Bortolotti's account, the distinction between an ill-grounded explanation and a well-grounded explanation that is based on a false belief is crucial to establish whether someone is confabulating or not. If we consider her two interpretations of the Nisbett and Wilson experiment, the difference between Roberto and Sylvia hinges on whether or not the belief about the preference was formed before or after choosing the preferred pair. However, it is not clear how we can determine this. Of course, one could decide to get the answer straight from the horse's mouth, and ask subjects whether their belief about the preference was formed before or after choosing the preferred pair. But the reliability of these subjective reports is precisely what is at stake in the confabulation studies. Would we believe Roberto when he claims that his belief about his preference was actually formed before choosing the preferred pair? As we saw in Sect. 3, Scaife concludes on the basis of empirical research that, from the first-person perspective, instances of confabulation are indistinguishable from accurate self-ascription. Bortolotti would need to show that this conclusion is unsupported.

The same problem arises when we consider the assumption that SK requires only that people are aware of their attitudes and choices, and that being ignorant of the processes responsible for these attitudes and choices does not undermine their SK. How can we be sure that subjects in the Nisbett and Wilson's experiment such as Roberto and Sylvia are indeed aware of their attitudes and choices? How do we know for certain that these attitudes and choices are not confabulated as well? We already find this worry in Carruthers' (2009):

> Nisbett and Wilson themselves cast this result in terms of confabulation about the causes of action, and those who believe in the introspectability of judgments will often dismiss it on that ground […] But this is to miss the point that subjects are also confabulating and attributing to themselves a judgment

Our point is not that Carruthers is necessarily right. Our point is that it is not clear how Bortolotti can dismiss this explanation by appealing to 'sincerely reported and genuinely endorsed' subjective reports. For the authoritative status of these reports is precisely what is at issue here.

We do agree with Bortolotti that confabulation studies do not necessarily undermine FPA of our everyday self-ascriptions. However, we have a rather different interpretation of *why* this is so. Bortolotti argues that it is 'implausibly demanding' to suggest that successful mental-state self-attribution requires that people are aware of the mental processes responsible for their attitudes and choices. She also challenges the suggestion that SK requires that people's self-attribution reliably predicts and explains their subsequent behavior. This imposes "more stability and consistency on people's mental life than is reasonable to expect." (p. 236) This is why she thinks someone like Sylvia has SK even if (a) she is ignorant of what caused her belief and (b) her self-ascription has little/no effect on her future behavior.

However, the question is whether a failure to meet these two conditions (a and b) has no impact whatsoever on FPA. From a self-regulation perspective, both backward- and forward-looking aspects are crucial for FPA, to the extent that they matter for the alignment of self-ascriptions with behavior. This means that we are skeptical about whether Sylvia should be granted FPA. For even though her belief is well-grounded, it is not clear that she succeeds in aligning her future behavior with this belief. It is true that the criteria by which FPA is determined may differ between contexts and the contents involved. In this sense, Bortolotti is right: our preference for stockings is usually of no importance in everyday life, so the criteria will not be very stringent when we are considering FPA regarding the self-ascribed belief. But when it comes to important personal (e.g., the self-ascribed belief that you love your children equally) or moral issues (e.g., the self-ascribed belief that you think discrimination is morally wrong), your future actions *will* play a significant role in determing the authority of your self-ascriptions.

In other words, even if Bortolotti's line of argument supports a specific notion of SK, it seems to be insufficient for a robust (mindshaping) conception of FPA. This is also why we are hesitant about the claim that successful mental state self-attribution does not require that people are aware of these processes, because

> "the causal factors leading to the attitude or the choice are psychological processes that involve priming effects, socially conditioned emotional reactions, and implicit biases whose role cannot be directly experienced or easily observed, but needs to be inferred on the basis of the systematic, scientific study of human behavior." (p. 236).

As we argued in response to Scaife, our folk psychology may become more scientifically informed, and compensating for our blind spots regarding the processes that give rise to our mental states could become the norm of socially extended

self-regulation. Arguably, this will also have an impact on our everyday understanding and attribution of FPA.

To flesh out the implications of our self-regulation account for the interpretation of FPA in the confabulation studies in a bit more detail, consider the case of Josephine. She is also a subject in de Nisbet and Wilson experiment, and she takes her preference for the far-right pair to be caused by (her belief about) its superior knitting pattern. However, Josephine *starts living up to* her self-ascription by buying only panty hoses with this knitting pattern, recommending it to friends, on blogs, etc. (cf. Strijbos and De Bruin 2015). When she is informed about the actual cause of her preference, Josephine 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 pair on the far right. 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 her choose pairs with that particular knitting pattern.

From the perspective of self-regulation, the question whether Josephine's belief about her preference was formed before or after her choice is not irrelevant, but it is also not *decisive* when determining FPA. This question only addresses the 'backward looking' dimension of confabulation. However, the self-regulation view suggests that the significance of confabulation is also determined by its 'forward looking' dimension, i.e. whether or not subjects succeed in regulating their future behavior. In this light, even if Josephine's belief was formed *after* her choice (which makes her explanation ill-grounded), she has aligned her behavior with it ever since. Of course, forward-looking considerations cannot post hoc miraculously make a false backward-looking claim true. In this sense, Josephine's claim to backward-looking SK at that time could be contested. However, the question we are interested in here is whether, despite this, we would withhold FPA, as we would in typical cases of clinical confabulation.

At the time of the experiment, Josephine instantiated a token of the self-ascription type "I like the knitting pattern of panty hoses of this kind" (or something like this) as her given reason for choosing the far-right pair. It seems implausible to grant her FPA regarding *this token* self-ascription as her reason, because it is, at that time, solely based a false backward-looking claim. However, Josephine could earn FPA regarding this self-ascription *type* by aligning her future behavior with it (by buying stockings with this knitting pattern, etc.). That is, future tokens of this type would be granted FPA in light of successful alignment. In this way, upward self-regulation (aligning her behavior with her self-ascription) might restore her FPA regarding claims of (future tokens of) the type "I like the knitting pattern of panty hoses of this kind".

But Josephine might also restore her FPA by means of downward self-regulation (aligning self-ascriptions with observed behavior), for example, by participating again in the same experiment (as we explained in Sect. 3). Of course, attributing the cause of her choice to the position effect would not restore her FPA for the particular token she instantiated during the first experiment. However, it would enhance her overall FPA status as someone who is aware of the subliminal factors that influence her choice for panty hoses in these specific (and rather artificial, given that there are four identical panty hoses) experimental conditions. And one could even argue that such an instance of downward self-regulation also indirectly enhances Josephine's FPA regarding the type "I like the knitting pattern of panty hoses of this kind"—she knows that her self-ascriptions of this type are unreliable in these conditions.

This illustrates the point we made in Sect. 2, namely, that FPA is a complex phenomenon whose interpretation unfolds within an interpretative or hermeneutical dynamic. The adequacy of self-ascriptions cannot be determined in isolation, and their assessment takes place in the diachronic context of past, present and future behavior, against the background of socio-pragmatic considerations pertaining to FPA status. Interestingly, although Bortolotti adopts a purely epistemic understanding of confabulation and SK, she does mention several important psychological and social mechanisms that are important for self-ascriptions in folk psychological practice: psychological adaptiveness, enhancing coherent self-image, self-perceived agency, etc. On our self-regulation account, these aspects are always in play when considering FPA. What matters in self-ascribing is that one can *make oneself understood*, and becomes a trustworthy, reliable, understandable, etc. participant in social practice. Self-ascriptions should help in the second-person coordination of social affairs. The potential costs and benefits of self-ascriptions, including confabulation, should be determined against this background.

As we already mentioned, the self-regulation view of confabulation was inspired by clinical cases. Especially in clinical cases, confabulated self-ascriptions seem to *undermine* the coordination of our affairs in folk psychological practice. In the next and final section, we will end with the further suggestion that it is not only the attitudes ascribed which determine the damaging effect to social interaction and coordination in such cases. It is also the way in which one relates to oneself in doing so.

5 The Importance of 'Self-Know-How'

Our proposal to understand the authority of self-ascriptions in terms of self-regulation highlights the fact that the success of our self-ascriptions in folk psychological practice is determined by the extent to which they facilitate interaction, coordination and cooperation in social practice.

In this final section, we want to direct attention to the fact that FPA also depends on the *manner in which one*

relates to oneself in providing one's reasons for actions and in responding to potential misalignment when faced with questions why. One of the features distinguishing successful self-ascriptions from unsuccessful ones (including confabulation) is the particular nature of the self-regulatory attitudes one adopts when providing them in response to the questions of others. In other words, it is not only the content (the 'what') of the ascription that is relevant, but also *how* one relates to oneself in self-ascription. This is what we will term *self-know-how*.

To tease out this distinction in relation to FPA, it is helpful to consider the fact that perfect alignment between our avowed self-ascriptions and our behavior is not the norm in folk psychological practice. We do not expect people to be fully consistent in their self-ascriptions, nor do we demand that their behavior and their self-ascriptions be completely compatible. One obvious reason is that, within certain limits, we grant people the opportunity to change their views, attitudes, and commitments and sometimes also expect people to change through the course of their lives, e.g. in response to certain events they experienced. In this sense, past self-ascriptions may come into conflict with present or future behavior (and *vice versa*), without necessarily undermining one's status as an authoritative self-ascriber. But even in more 'synchronic' cases, e.g., when giving reasons for occurrent behavior, full compatibility is not the norm. Within limits, we grant people some leeway in making themselves understood. We allow them to save face as long as the answers provided are not too far off. Depending on the subject matter and one's relation to the other, certain degrees of indeterminacy are acceptable. Apart from impression management, the simple fact, embodied in our social dealings with one another, is that a certain degree of misalignment is part of the human condition. We know ourselves and others not to be perfectly cognizant of our own minds let alone of the possible consequences of our self-ascriptions and behaviors for overall consistency. What we expect from people is not the superhuman ability to be fully consistent, but rather *the ability to adequately cope with certain degrees of misalignment*, holding in regard the demands of social practice. This ability does not merely or even primarily refer to the content of one's self-ascriptions, but also and perhaps more importantly to the know-how embodied in the self-regulatory skills one uses when confronted with inconsistencies between one's self-ascriptions and/or behaviors in social situations.[8]

We do not have the space here to give a detailed account of the know-how involved in adequately coping with possible misalignment regarding one's self-ascriptions in everyday social life. What we can do, however, is give an impression of some of the requirements involved by highlighting some shortcomings on this score in certain conditions. We already stated that perfect alignment is not the norm in human social practice, one reason being that it is psychologically and practically impossible for human beings to be fully consistent in their self-ascriptions. Here we would want to make the stronger claim that the impression of perfect alignment actually tends to raise suspicion. A person, who consistently, even convincingly, rationalizes away any inconsistencies between self-ascriptions and behavior, would not make a particularly good impression as a reliable and trustworthy partner in our social endeavors. Coping with possible misalignment should consist of more than one strategy—the inclination to rationalize (away) should be balanced out by others, for example the ability to admit errors and to negotiate about more viable self-ascriptions. Moreover, an obsessive striving for perfection in this sense might indicate vulnerabilities in psychological, neurocognitive or personality profile (cf. McGeer 2008). Such obsession with consistency can be associated with rigidity (e.g., in cases of autism), feelings of fear or nagging uncertainty (e.g., in the course of OCD), or intolerance of feelings of inferiority when being held to account by others (e.g., in narcissistic personalities).

In general, coping with inconsistencies between one's self-ascriptions and behavior in social settings in an authority preserving/promoting way seems to be facilitated by adopting certain kinds of self-relational attitudes and undermined by adopting others. Authority preserving/promoting attitudes are, e.g., self-directed attitudes of flexibility, openness to new perspectives, inquisitiveness to possible inconsistencies, attentiveness to one's own thoughts, feelings and emotions, feelings of certainty and self-confidence regarding the prima facie viability of one's answers to questions why, benevolent and respectful self-criticism, etc. By contrast, self-directed attitudes characterized by inflexibility, rigidity, lack of interest, groundless doubt, lack of self-confidence or devaluative self-criticism, tend to undermine one's authority as a reliable self-interpreter.[9]

[8] We think the term 'know-how' is appropriate here, because we are talking about *the skills involved* in adopting certain kinds of attitudes towards oneself and managing one's self-understanding, when facing others in explaining or justifying one's behavior or self-ascriptions. People learn these self-regulation skills through upbringing and socialization.

[9] The self-regulatory attitudes in play here seem to be partly epistemic (in the sense that they figure in the process of gaining knowledge about one's mental states in relation to one's behavior), partly agential (because they play a role in the regulation of one's mental states and behavior), partly moral (in being closely associated with or perhaps implying certain moral virtues such as benevolence, regard, and respect towards self and others). We are inclined to refer to these attitudes as 'self-relational virtues', given the association with e.g., intellectual and moral virtues often discussed in philosophical literature.

Why does adopting these self-directed attitudes promote or undermine FPA? An important reason is that human beings are prone to error when interpreting their dispositions and predicting how they will think, feel and act under certain circumstances. As social psychologists have tirelessly pointed out, we have limited access to many factors that determine our mental and behavioral dispositions, and are subject to all kinds of bias when trying to deal with possibly conflicting beliefs and desires. Flexibility in perspective taking, openness, inquisitiveness, and attentiveness to various aspects of oneself, protect us from shortsightedness and other epistemic biases when interpreting ourselves, while attitudes characterized by a fair degree of self-respect and self-confidence help to avoid endless doubt and unwarranted negative self-evaluation in the face of others. As we pointed out in our discussion of Scaife's account (Sect. 3), alignment of one's self-ascriptions and one's behavior requires deep knowledge of one's dispositions, i.e., what causes one to feel, think and act in certain ways under certain conditions, and how these dispositions can be shaped in favored directions. Structural neglect of factors that have high impact on one's behavior, and unwillingness or incapacity to address this (e.g., in the course of clinical conditions alluded to earlier), can cause *structural misalignment*. This, in turn, makes one's self-ascriptions unreliable and undermines one's FPA status significantly. Given our complex, layered, and sometimes conflicting nature, we need to relate to ourselves in certain ways so as to avoid structural misalignment. The know-how embodied in the self-regulatory skills of adopting attitudes of flexibility, attentiveness, self-esteem and the like, helps us to 'keep our car on the road', keeping misalignment to an acceptable minimum by negotiating and navigating between different perspectives, interests and concerns.[10]

We briefly mentioned certain clinical examples characterized by self-regulatory attitudes that typically undermine FPA-status. In cases of clinical confabulation, the neglect and apparent indifference towards incompatibilities between one's self-ascriptions and behavior, is perhaps the most eye-catching feature of the failure of self-regulation. From the perspective of folk psychology, these self-relational attitudes of indifference and neglect, whatever their precise neurobiological underpinnings, bring the process of making oneself understood towards others to a grinding halt, even up to the point that it becomes questionable whether the notion of FPA is still applicable. Ramachandran (1995) describes the case of a 76-year-old woman, Mrs. M., who had a stroke that left her completely paralyzed on the left side. Although quite lucid when discussing most other topics, she persistently denied her paralysis even when pressed, and her answers were not hesitant or lacking in conviction:

> "Mrs. M., when were you admitted to the hospital?"
> "I was admitted on April 16 because my daughter felt there was something wrong with me."
> "What day is it today and what time?"
> "It is sometime late in the afternoon on Tuesday."
> (*This was an accurate response*).
> "Mrs. M, can you use your arms?"
> "Yes."
> "Can you use both hands?"
> "Yes, of course."
> "Can you use your right hand?"
> "Yes."
> "Can you use your left hand?"
> "Yes."
> "Are both hands equally strong?"
> "Yes, they are equally strong."
> "Mrs. M, point to my student with your right hand."
> *(Patient points)*
> "Mrs. M, point to my student with your left hand."
> *(Patient remains silent)*
> "Mrs. M, why are you not pointing?"
> "Because I didn't want to…

There is a lot to say about this clinical example. Our point here is that her incapacity to *relate to* and *manage* this misalignment in appropriate ways—her structural failure in *self-know-how* (related to self-attributions involving her left arm)—is arguably more detrimental to her FPA-status than the particular misalignment itself. Recall from Sect. 2 that diminishment of FPA-status can be local, in the sense of pertaining only to certain domains of self-ascriptions or aspects of the self. This is exactly what is going on in the case of Mrs. M. When discussing other topics, unrelated to the content of her anosognosia, FPA did not seem to pose significant problems. It was only when being confronted with the lack of functioning of her left arm, that structural misalignment occurred, to the extent that the whole pragmatic point of attributing or withholding FPA seems to be lost.

From a mindshaping perspective on folk psychology, self-know-how is central to the practice of making oneself

[10] A direction to pursue further in future research is the parallel with self-regulation enhancing therapies in mental health care. A significant part of these therapies consists in teaching patients to trust themselves in adopting and maintaining a kind of tolerance, openness and inquisitiveness toward the gap between conflicting evaluations of self and others and towards the emotional tensions that go with it. The concepts of 'mentalizing' and 'epistemic trust', as put forward by e.g., Fonagy and Allison (2014), seem particularly promising in this respect: "Mentalizing in therapy is a generic way of establishing epistemic trust between the patient and the therapist with the aim of freeing the patient from rigidity, so that they can begin to learn from new experiences and achieve change in their understanding of their social relationships and their own behavior and actions […] Put simply, the experience of feeling thought about in therapy makes us feel safe enough to think about ourselves in relation to our world, and to learn something new about that world and how we operate in it." (p. 273).

understood and managing one's authoritative status as a reliable self-interpreter. As far as we know, the capacities involved in self-know-how have not yet been systematically investigated from the philosophical perspective of FPA (and SK). More work is needed to flesh out the relation between FPA and self-know-how, and to explain when a certain kind of responsiveness to misalignment becomes pathological. This will have to wait for another occasion. For now, let us end by briefly summarizing the main points of our paper.

We have suggested that FPA is based on a capacity for *self-regulation*, i.e. a capacity to bridge the gap between our sayings and doings by aligning our actions with our avowed self-ascriptions and vice-versa. On the self-regulation view, confabulation studies as of yet have little import on the subject of FPA because they do not specifically address the question whether and to what extent the subjects are actually able to bridge this gap. We claimed that the apparent fact that we cannot reliably distinguish, from a first-person perspective, when we are confabulating and when we are not, does not necessarily undermine FPA. This is because self-regulation in folk psychological practice is essentially a second-person activity. Furthermore, we argued that a systematic failure to align our actions with our self-ascriptions and vice versa is a genuine threat to FPA. To the extent that confabulation undermines this capacity, confabulation does affect FPA. In this last section, we introduced the concept of self-know-how—the skills involved in adopting certain attitudes towards oneself in making sense of oneself with or in the face of others—and briefly explored the importance of diminished or absent self-know-how in clinical cases, such as clinical confabulation.

Compliance with Ethical Standards

Conflict of interest Authors declare that there is no conflict of interest.

Ethical Approval This article does not contain any studies with human participants performed by any of the authors.

References

Andrews K (2012) Why do apes read minds? Toward a new folk psychology. MIT Press, Cambridge

Bilgrami A (2006) Self-knowledge and resentment. Harvard University Press, Cambridge

Bortolotti L (2018) Stranger than fiction: costs and benefits of everyday confabulation. Rev Philos Psychol 9(2):227–249 (2018)

Brandom R (1994) Making it explicit. Harvard University Press, Harvard

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, Oxford

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 (2016) First-person folk psychology: mindreading and mindshaping. Studia Philos Estonica 9:170–183

De Bruin LC, Jongepier F, Strijbos DW (2014) Mental agency as self-regulation. Rev Philos Psychol 6:815–825

Fiala B, Nichols S (2009) Confabulation, confidence, and introspection. Behav Brain Sci 32(2):144–145

Fonagy P, Allison E (2014) The role of mentalizing and epistemic trust in the therapeutic relationship. Psychotherapy 51(3):372–380

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, Berkeley

Hall L, Johansson P, Tärning B, Sikström S, Deutgen T (2010) Magic at the marketplace: choice blindness for the taste of jam and the smell of tea. Cognition 117:54–61

Hall L, Johansson P, Strandberg T (2012) Lifting the veil of morality: choice blindness and attitude reversals on a self-transforming survey. PLoS ONE 7(9):e45457

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

Johansson P, Hall L, Sikström S, Olsson A (2005) Failure to detect mismatches between intention and outcome in a simple decision task. Science 310:116–119

Johansson P, Hall L, Sikström S, Tärning B, Lind A (2006) How something can be said about telling more than we can know. Conscious Cognit 15:673–692

Johansson P, Hall L, Sikström S (2008) From change blindness to choice blindness. Psychologia 51:142–155

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, Princeton

Morton A (2003) The importance of being understood: folk psychology as ethics. Routledge, Abingdon

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

Scaife R (2014) A problem for self-knowledge: the implications of taking confabulation seriously. Acta Anal 29:469–485

Strijbos W, De Bruin LC (2015) Self-interpretation as first-person mindshaping. Theory Moral Pract 18(2):297–307

Zawidzki T (2013) Mindshaping: a new framework for understanding human social cognition. MIT Press, Cambridge