Acta Anal (2014) 29:469–485
DOI 10.1007/s12136-014-0226-8

A Problem for Self-Knowledge: The Implications of Taking Confabulation Seriously

Robin Scaife

Received: 30 April 2013 / Accepted: 25 February 2014 / Published online: 15 August 2014

Abstract There is a widespread assumption that we have direct access to our own decision-making processes. Empirical demonstrations of confabulation, a phenomenon where individuals construct and themselves believe plausible but inaccurate accounts of why they acted, have been used to question this assumption. Those defending the assumption argue cases of confabulation are relatively rare and that in most cases, we still have direct insight into our own decision-making. This paper reviews this debate and introduces two novel points. Firstly, I will point out that a rich source of evidence of confabulation is often overlooked. Secondly, I will argue that our inability to distinguish confabulations from cases in which we gain accurate information about the reasons for our actions gives rise to an empirically motivated scepticism which gives us grounds to doubt the accuracy of all our introspective insights into our own decision-making processes.

Keywords Self-knowledge · Introspection · Confabulation · Self-interpretation · Non-conscious

1 Introduction

The dominant view amongst both folk and philosophers is that we have reliable introspective access to the reasons for our own actions. When Carruthers (2011) ran an informal analysis of philosophy papers about self-knowledge published since 1970, he found that 94 % of authors supported the view that we have transparent access to our own thought processes. If we restrict the transparency claim to the domain of decision-making (the domain which is the target of the sceptical problem I am developing in this paper), then there is a similar consensus amongst the folk. When investigating the claim that the folk assume they typically have introspective access to their own minds, Kozuch and Nichols (2011) found that people only assumed introspective transparency

R. Scaife (✉)
Department of Philosophy, University of Sheffield, 45 Victoria Street, Sheffield S3 7QB South Yorkshire, UK
e-mail: r.scaife@sheffield.ac.uk

for certain domains. Of the domains they investigated decision formulation was regarded as especially available to introspection. That this folk view underlies many of the norms governing our social interactions becomes immediately apparent when we consider how rarely people admit that they do not know why they made a certain decision.

However, this dominant assumption about introspective transparency appears to be incompatible with the empirical evidence. There are now numerous studies which indicate that people can be mistaken about their own motivations. These studies demonstrate a phenomenon known as *confabulation*, where participants construct plausible but inaccurate accounts of their own motivations rather than gaining genuine insight into the decision-making process which gave rise to their behaviour. In these experiments, the participants are completely unaware that they are fabricating accounts of their own motivations and confidently treat their confabulations as if they were genuine introspective insights into their decision-making process.

Most modern philosophers are united in rejecting unacceptable Cartesian views of the transparency of mind. However, they still believe that there is such a thing as introspective access to our own states of mind and that there are areas—such as how things seem to us, what we are feeling and why we acted as we did—in which it can give us reliable self-knowledge. In respect to decision-making, they do not let evidence of confabulation discourage them from being confident about self-knowledge because they do not interpret the findings as indicating anything about how people typically understand themselves. The standard claim is that these results only provide evidence of how our typically accurate process for gaining introspective insight into our decision-making can go wrong under the unusual circumstances which psychologists construct in order to conduct these studies. Since such circumstances rarely occur in our everyday lives, we should remain confident that people generally have direct and infallible knowledge of the reasons for their actions.

The key move in this line of response is to posit two distinct processes for gaining information about our own motivations. The claim is that we typically have reliable self-knowledge derived from a process which allows people direct insight into their own decision-making and that it is only when this process fails that the second process kicks in which leads to people confabulating reasons for their behaviour. Those who make this kind of response have become known as 'dual-method theorists' because they think there are two processes by which we can gain an understanding of our own decision-making. Dual-method theorists typically claim that we normally have reliable insight into our own decision-making and that cases of confabulation are 'relatively rare' (Goldman 2006, p. 232).

A small number of philosophers take these findings much more seriously. Most notably, Carruthers (2010, 2011) argues that these findings indicate that all our insights into our own motivations are the result of a process of self-interpretation rather than the result of any direct introspective access to our own decision-making. I call this an 'interpretation-only' account of self-knowledge, because the account claims that the only process for gaining information about our decision-making is one which operates through self-interpretation.

There is some debate over which kind of account offers the best explanation of how we gain insight into our own motivations. Goldman (2006, 2009) argues that a dual-method account best explains the detailed understanding we have of our own decision-

making as well as fitting with the phenomenology of how we think about our own motivations. In contrast, Carruthers (2010, 2011) advocates an interpretation-only account on the grounds that it is simpler than its dual-method counterpart and that dual-method accounts owe us a principled explanation of why we sometimes gain genuine insight and other times engage in self-interpretation.

Despite agreeing with Carruthers that we should favour the interpretation account on the grounds of simplicity, I do not think the debate is likely to be decisively settled in the near future. Furthermore, I think this debate overlooks important concerns which arise when we carefully consider the implications of the experimental demonstrations of confabulation. Irrespective of how the debate between dual-method and interpretation-only accounts is resolved, we should be concerned about the reliability of our self-knowledge. This is because cases of confabulation are indistinguishable from cases where we gain correct information about our own decision-making. This leaves open the sceptical possibility that, any time we consider our own motivations, we might not be getting accurate information.

You might not be inclined to take this scepticism very seriously. After all, we cannot prove that we are not brains in vats being fed sensory information by an evil demon, but most of us do not take this sceptical possibility very seriously. I will argue that we should not trivialise scepticism about self-knowledge in the same way that we do brain-in-a-vat scepticism because it is a much more serious type of sceptical concern based on generalising well-documented scientific findings. Unlike the fanciful hypothesis of brains in vats, we know that the sceptical scenario can and does occur. Confabulation is a scientifically documented phenomenon. We have experimental methods for detecting it, and we know some of the factors which give rise to it.

If you accept that from a first-person perspective, cases of confabulation are indistinguishable from cases of genuine insight into our motivations, then the only reason why you might not take this scepticism seriously would be if cases of confabulation were very rare indeed. However, I will argue that there are a number of reasons why we should not think that this is the case. There is actually a great deal more empirical evidence of confabulation than most philosophers are aware of. Furthermore, there is good reason to think that confabulation evidence will generalise from the psychology lab into the real world. This is because many people encounter the kinds of factors which lead to confabulations every day of their lives. I will argue that all these factors lead to the conclusion that we should take scepticism about our self-knowledge more seriously.

I will begin by outlining an array of experiments which provide evidence of confabulation. Then, in Section 3, I will discuss what I consider to be an often overlooked source of additional evidence of confabulation. In Section 4, I will outline the two competing types of account of self-knowledge: interpretation-only accounts and dual-method accounts. I will suggest that there are reasons to favour the interpretation-only view but acknowledge that despite this, the debate between these two accounts is likely to stagnate. In Section 5, I will argue that the debate between different accounts of self-knowledge overlooks an important sceptical concern which arises from the evidence of confabulation. I will argue that if cases of confabulation occur and individuals cannot tell when they themselves are confabulating, then this generates the concern that we can never be certain that any particular attempt to gain self-knowledge has succeeded. I will then argue that we must take this scepticism

seriously by replying to what I take to be the two main objections to scepticism about self-knowledge. The first objection is that self-knowledge comes with markers of certainty. I will argue that this is not reliably the case. The second is that sceptical problems of this type are not to be taken seriously. I will argue that scepticism about self-knowledge is precisely the type of scepticism which needs to be taken seriously. I will draw a distinction between traditional forms of purely philosophical scepticism and empirically motivated scepticism and argue that scepticism about self-knowledge needs to be taken seriously because it belongs in the latter category.

2 Experimental Evidence Demonstrating Confabulation

2.1 Split-Brain Cases

Gazzaniga (1995, 2000) conducted a number of experiments with split-brain patients. These participants had, for various reasons, undergone surgery to sever their corpus callosum. This procedure prevents any information being transferred between the two hemispheres of the brain. By a process of showing information to only the right eye, it can be ensured that only the left brain is able to process a certain piece of information and vice versa. The reason the eye location and brain hemisphere are opposite is because the optic nerve crosses over. The left hemisphere houses the speech centres of the brain, and therefore, only information presented to the right eye can be verbally expressed. This allowed Gazzaniga to investigate confabulation in ways which would not be possible with non-commissurotomised participants.

In one experiment, Gazzaniga flashed the instruction 'walk' in the left eye of these split-brain patients. Once this had been done, they would get up and wander off. Participants were stopped and asked why they had walked away. Because participants are required to answer verbally, they have to use their left brain. However, only the right brain was aware of the instruction to walk and has no way for this information to be transmitted to the left brain. One might expect that in such circumstances, participants would report that they had no idea why they walked away. Yet, the participants always came up with reasons, such as 'I wanted to get a coke'. It is possible that such an answer is not entirely confabulated. Maybe the left brain did generate a desire for a can of coke, although it does seem an unlikely coincidence that they decided to act on it just after the instruction to walk had been flashed up. It could have been the case that once the participant began to walk, their left brain considered the possible advantages of wandering off, such as getting a coke. If this is the case, the coke answer is no longer pure confabulation, but it still seems false to call it the reason why they walked off. Furthermore, if the participants had been allowed to wander off, they may well have got a can of coke. However, all this would show is that once someone has engaged in confabulation, the confabulated motive often *becomes* a genuine motivation.

In another experiment, a picture of a chicken's claw was shown to the right eye and a picture of snow was shown to the left eye. The participants were then asked to pick two associated pictures from a large selection. They always selected a chicken and a shovel. When asked why they made these selections, they would reply that the chicken went with the chicken's claw and the shovel was to clean out the coop. Both selections were explained by reference to the chicken's claw because the area of the brain generating

the report had no awareness of the picture of snow. This kind of finding was replicated across a wide variety of matching tasks with both selections being explained by participants with reference to the right eye cue.

In all the split-brain studies, participants were reminded about their operations and that they might not be aware of influences from their left visual field, but even after this prompt, they still insisted that they acted on the basis of the reasons they had cited. To everyone who has watched the videos of some of these trials, it is evident that participants are confident that the reasons they confabulated did cause their actions. They are as convinced by the confabulated reasons, as they are by genuine reasons (such as 'chicken's claw goes with chicken'), even when the confabulated reasons seem farfetched (such as 'shovel goes with chicken because you have to use one to clean out the chicken coop').

While these results do provide an impressive demonstration of confabulation, generalising these results is questionable because all the participants had major brain surgery. Perhaps, something about having one's corpus callosum severed causes confabulation. Furthermore, the participants all had cognitive problems to start with; otherwise, they would not have undergone the surgery in the first place. However, the case for confabulation does not rest only on split-brain cases. There are many other studies which suggest that confabulation is, by no means, a phenomenon which is restricted to split-brain patients.

2.2 Direct Brain Manipulation

Strong evidence of confabulation comes from a study which used direct brain manipulation. One major advantage of this technique is that there can be absolutely no doubt as to what caused the participant to act. In one such study, Brasil-Neto et al. (1992) asked participants to lift either their right or their left index finger when they heard a clicking sound. The sound was in fact an electromagnet being turned on. This magnet was directed at the participants' motor cortex and was determining which finger participants lifted. In all trials, participants reported having chosen which finger they lifted despite this decision not causing the finger movement. It is statistically very unlikely that their choice of finger to move randomly matched the magnetic stimulation in all the trials. Furthermore, there is no evidence that decisions to act are made in the motor cortex. This precludes the explanation that the magnetic stimulation caused an intention to act which then became available to participants through introspection so that they subsequently treated it as their own decision.

2.3 Cognitive Dissonance

Some of the earliest evidence of confabulation comes from work on cognitive dissonance. In 1959, Festinger and Carlsmith found that students who were only paid $1 to tell another student that a boring repetitive task was interesting later reported that they genuinely did enjoy the task. However, students who were paid $20 to instil the expectation that the task would be interesting stood by their initial judgement that the task was dull. Festinger and Carlsmith attributed this difference to a phenomenon they called cognitive dissonance. They argued that the students who were paid $1 suffered from cognitive dissonance because they did not think that merely $1 justified lying to

another student. As a result, they would convince themselves that the task was not really so boring. The students who were paid more did not suffer from dissonance because the $20 (a lot for a student in 1959) justified telling a white lie to a fellow student.

Similarly, Cohen (1962) found that participants would show more sympathy for a position they initially disagreed with the less they were paid to write an essay justifying that position. Cohen concluded that participants were confabulating as follows: 'I must have some sympathy for the position if I agreed to defend it for so little money'. Whereas, those paid more money would simply conclude that the only reason they were writing the essay was for the money. Both these studies seem to show participants confabulating as they strive to keep a narrative which is consistent with both their actions and how they view themselves. Interpreting these studies in this way does require us to make a number of assumptions about the participants' reasoning. In this kind of study, it will always be difficult to rule out the possibility that other psychological processes could be driving the change in attitude. However, the burden of proof rests with those wishing to deny that confabulation occurs because it is not obvious what else could be generating these results.

2.4 Choice Blindness

Johansson et al. (2005) presented participants with 15 pairs of pictures of women's faces and asked them to choose which ones they found more attractive. After each selection, the experimenter would pass the picture to the participant and ask them to explain their decision. Three out of the 15 trials were manipulated in such a way that the picture passed to the participant was, in fact, the face of the woman they had rated as less attractive. The majority (74 %) of these manipulations went unnoticed. Furthermore, many participants confabulated reasons for their choice such as 'I prefer blondes' even though they had originally chosen the brunette!

Readers may think that participants did notice the manipulation but did not report it to the experimenter due to social conformity pressures. In order to rule this out, Johansson et al. (2005) included a question in the debrief about a hypothetical variation to the study which would use the manipulation which was, in fact, used in the study. Of the participants who failed to notice the manipulation, 84 % claimed that they would be able to spot this kind of manipulation. Later in the debrief, when the true nature of the experiment was revealed, many participants were clearly surprised, and some even refused to believe that the manipulation had taken place.

Hall et al. (2010) obtained similar results using the same type of methodology when asking participants to make choices between types of jam or tea. In this study, they revealed that the manipulations had taken place before explicitly asking participants if they had noticed the switches. The reason for this was to get demand characteristics to pull in the opposite direction to the effect, by creating a pressure which would result in an over-reporting of the manipulation having been detected. Despite this, only 33 % of manipulations were detected (of which just below 7 % were retrospectively detected). This all suggests that the vast majority of participants providing confabulated justifications did themselves believe these to be the real reasons behind their choices.

We might not think this is particularly worrying when we choose between types of jam or choose which photo depicts someone more attractive to us. Many of us will find

it acceptable that people are careless when making such choices. However, it is more worrying that there is evidence of the same pattern in our moral judgments. Hall et al. (2012) found evidence of the same choice blindness and confabulation when people were given false feedback about how they had filled in a questionnaire asking moral questions. They gave participants either 12 statements of moral principles or 12 concrete applications of those principles and then asked them to indicate if they agree or disagree on a nine-point scale. The principles included statements such as 'To be moral is to follow the rules and regulations of the society, rather than weighing the positive and negative consequences of one's actions', and the concrete applications were statements such as 'It is morally deplorable to harbour immigrants when they have been declared illegal and scheduled to return to their home country by the Swedish government'. The experimenters used a manipulation which changed two out of the 12 statements. For example, 'It is morally deplorable to harbour immigrants…' was changed to 'It is morally commendable to harbour immigrants…'. This manipulation changed the questions completely but left the participants' answers on the same nine-point scale. For example, a judgement of 'strongly agree' that followed the statement 'harbouring immigrants is deplorable' now followed the opposing statement that 'harbouring immigrants is commendable'. The change resulted in participants' answers to two of the 12 statements expressing exactly the opposite view to the preference they had selected earlier. Participants were then asked to go through some of their answers and explain why they had expressed the preferences they had. Across the two conditions, 69 % of participants accepted at least one of the two altered statements and had no problem in providing reasons in support of the opposite preferences to the ones they had expressed earlier. Again, it seems clear that participants are simply generating plausible reasons in favour of an opinion, rather than gaining genuine access to their decision-making process. If they were gaining access to their decision-making process, this would surely lead them to realise that they did, in fact, decide in favour of the opposite preference to the one being attributed to them.

What is of particular interest in these studies is that the justifications which the participants used for the choices they did not make were, in many respects, indistinguishable from the justifications they used for the choices they did make. In both Johansson et al.'s (2005) original study and a reanalysis conducted a year later (2006), they were unable to find any significant differences in the justifications provided in the manipulated and the non-manipulated trials. They checked a large number of factors which might differentiate confabulated justifications in terms of manner of delivery or markers of attitude. Amongst other things, they checked for 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 and word length. None of these factors provided any indication of whether or not the participant's justification was confabulated.

3 Non-conscious Influences as Additional Evidence of Confabulation

Perhaps the most well-known evidence of confabulation comes from a classic study by Nisbett and Wilson (1977) in which they found that participants would confabulate in order to rationalise decisions made as a result of right-hand bias. They found that right-

handed participants would favour the objects closest to their right hand when choosing between identical objects (such as identical pairs of socks). However, when interviewed, they would justify their selection by appeal to any number of factors. For example, claiming the object was made of better material or was shinier.

One explanation is that the bias is caused by such visual illusions which make people favour objects near their right hand. On this reading, the participants are not, in fact, confabulating because they are genuinely motivated by these perceived differences. However, this interpretation of the results seems unlikely when we consider that participants offer a vast array of different explanations for exactly the same bias. To deny that this is caused by confabulation, one must believe that the same manipulation induces different reasons in different people, yet this array of different reasons just happens to consistently produce the same behaviour. When we consider this, it becomes clear that it is simpler and therefore more plausible to interpret these results as showing that participants are confabulating reasons for their choice rather than providing a genuine insight into what caused it.

Apart from this Nisbett and Wilson (1977) study, cases in which participants' behaviour is driven by non-conscious biases are not typically cited as a source of evidence for confabulation. This is because these studies do not set out to investigate confabulation, and consequently, they do not typically mention confabulation in the published reports of their findings. However, I want to call attention to the fact that research on non-conscious influences on human behaviour is a rich source of additional experimental evidence of confabulation. For example, research on priming effects reliably produces evidence of confabulation. A priming effect occurs when non-conscious perception of certain stimuli directly influences subsequent behaviour. Psychologists typically generate priming effects by trying to activate a particular concept outside of the participant's awareness and measuring to see if this activation influences subsequent behaviour.

One example of this is a study by Bargh et al. (2001, experiment 2) in which they managed to use priming to induce a goal-directed behaviour. The prime was administered through the use of a scrambled sentence task (method of Srull and Wyer 1979) where participants had to unscramble a list of words to form a sentence. For the experimental group, the task contained words such as *share*, *cooperative* and *helpful* to try to induce the goal to cooperate. For the control group, the scrambled sentence task contained neutral words such as *salad*, *wet* and *zebra*. Following the scrambled sentence task, participants were then asked to do a supposedly unrelated resource management game. In this game, participants either keep resources (in this case, fish) for personal profit or return them to the lake in order to preserve fish stocks. Bargh et al. (2001) found that activating the goal to cooperate outside of participants' awareness using words which primed for cooperation caused participants to cooperate significantly more than participants from the control group who were only exposed to the neutral words. In fact, they found that those exposed to the cooperation priming words display the same level of cooperation as participants who did not complete a scrambled sentence task but were instead given the explicit conscious goal to cooperate. The only marked difference between these two groups was that those who had been primed did not report any intention to cooperate. This suggests that the prime had a direct non-conscious influence on behaviour rather than influencing behaviour through inducing the conscious goal to cooperate. The participants in the cooperation priming condition

did not report the influence of the priming words. Instead, they reported a variety of other reasons for adopting the strategy that they did.

This study and these kinds of studies in general, I would argue, provide evidence of confabulation—but not in a straightforward way. It would be wrong to say that all the participants displayed evidence of confabulation. It would even be incorrect to claim that we can say with any certainty that we know that any particular individual confabulated the reasons that they gave for their cooperation. However, we can say with some certainty that some of the participants in this study were confabulating about why they adopted the strategy of cooperating. The fact that those exposed to the priming words were significantly more likely to cooperate than those who were not indicates that exposure to the priming words was the reason why some of the participants in the experimental condition cooperated.

When I say 'reason' in this context, I do not mean that they consciously held this reason, but rather that it is an accurate causal explanation of why they cooperated. It is an accurate causal explanation of why many participants cooperated because we know from comparison with the control group that had they not been exposed to the cooperation priming words, then many of them would not have cooperated. That is to say, the statistically significant difference between the two groups indicates that for many participants, the priming words are the difference maker. If when we want to know the reasons for our decision we are interested in why we made that decision as opposed to choosing another option, then we are interested in the difference maker. If someone reports reasons for their decisions that are not the difference maker, then it seems correct to say that they are confabulating.

We know from the experimental debrief that none of the participants in the cooperation priming condition cited the priming words or even an intention to cooperate when explaining their cooperation behaviour. Instead, they gave other reasons. This enables us to conclude with some certainty that those participants whose behaviour was influenced by the priming effect also confabulated about their reasons for behaving as they did. Because these psychology studies only provide an indication of differences across large samples, we are not able to identify specific cases of confabulation in the same way that we are not able to identify specific cases of priming, but just as we conclude that this study provides evidence of priming, we should also conclude that it provides evidence of confabulation.

It is important to note that this is not just true for this study, but for all empirical demonstrations of priming. For a prime to be successful, the participants must be unaware of the fact that they are being primed. It is a standard procedure that the experimental debrief in priming experiments checks for this. Typically, the check is carried out by asking participants why they acted the way that they did. Participants rarely say they do not know why they acted the way they did nor do they identify the prime as the cause of their behaviour. Instead, they come up with seemingly plausible reasons for their actions. Furthermore, they almost universally stick by these confabulated motivations when given increasingly leading questions asking whether the task they had previously engaged in (the priming task) could have influenced their behaviour. This indicates that such studies are a rich source of additional evidence of confabulation. This is an important point because if I am correct to interpret priming studies in this way, then there are at least a hundred more studies demonstrating confabulation than philosophers typically think there are. I would urge any philosopher

who thinks that confabulation lacks an empirical support to take a detailed look at the psychology literature on priming (for a summary, see Ferguson and Bargh 2004 or Bargh and Chartrand 2000). Furthermore, if I am right about this, then it is probably the case that many other studies demonstrating other types of non-conscious influences on our behaviour such as the bystander effect (Latané and Darley 1970) also constitute evidence of confabulation, but I do not have space to discuss these cases here.

This interpretation of what counts as confabulation is not uncontroversial. However, I will not attempt to make a detailed defence of this interpretation here because it is an issue of such complexity that it would require a defence long enough to be a paper in its own right. For those of you unconvinced by my arguments, it is worth noting that I do not rely on this interpretation to generate the sceptical concerns about self-knowledge that I am arguing for. This interpretation only allows me to establish that there is more empirical evidence of confabulation than philosophers typically think there is. This would strengthen what is already a strong case in favour of taking scepticism about our self-knowledge seriously, but it is not necessary to get the scepticism going.

4 Accounts of Introspection

4.1 Carruthers' Interpretation-Only Hypothesis

Recently, Carruthers (2010, 2011) has questioned the dominant assumption about the introspective transparency of our thought processes by arguing that 'neither judgements nor decisions are introspectable, but are known only via a process of self-interpretation' (2010, p. 79). One of Carruthers' main arguments for his all self-knowledge arises from self-interpretation account of introspection is that it offers the best explanation of the empirical evidence demonstrating confabulation.

Carruthers argues that the human brain is not organized in a way which grants introspective access to the decision-making systems. As a result, reflecting about our own motivations is carried out in much the same way we attribute motivations to other people—a process of interpretation involving abductive inference based on the available perceptual input. Cases of confabulation are simply cases where we get the self-interpretation wrong and come to false conclusions about our own motivations.

Carruthers does think there are some reasons to believe that our self-interpretation process is reliable (at least in comparison with our ability to interpret the motivations of others). He points out that although the inward-directed and outward-directed attribution processes are identical, they will drastically differ in regard to the amount of data which is available to them. Even in the best outward-directed case, we will have only witnessed a small fraction of the individual's life leading up to the decision in question, whereas in the inward-directed case, we will have witnessed every waking moment of our own lives. Furthermore, Carruthers claims that in the inward-directed case, the interpretation process can draw on additional types of data such as mental imagery and inner speech.

4.2 Dual-Method Theories

The large body of empirical demonstrations of confabulation make it very difficult to argue that we always have genuine introspective insight into our own decision-making.

However, it is possible to maintain that we sometimes have introspective access, and it is only when this fails that we engage in interpretation in order to understand our own decision-making.

This is the kind of approach that is favoured by Alvin Goldman and has become known as a dual-method theory of self-knowledge. The central claim in dual-method accounts is that cases of confabulation are unusual cases of first-person mental state attribution, and typically, we do come to know about our own motivations through genuine introspection. Goldman (2006) outlines his position as follows: '…cases of confabulation may be relatively rare. Standard cases of first-person attribution may involve introspection or self-monitoring. This is the theory I wish to embrace, and the resulting overall theory might be called a *dual-method* theory. To sustain the privileged access theory, it suffices that a substantial range of mental attributions are executed by a special, distinctive method that works (fully) only in the first person. That there is a second, backup method, which can be used equally for first- and third-person attribution, doesn't detract from privileged access' (Chap 9, p. 232). From this, it is clear that Goldman's defence of privileged introspective insight works not by arguing against the empirical evidence supporting confabulation, but by trivialising it.

4.3 Assessing the State of the Debate

One clear advantage of Carruthers' interpretation-only hypothesis is that it is based on generalising a phenomenon for which we have empirical evidence. The studies which demonstrate confabulation clearly show that there are cases in which we use self-interpretation when considering our own motivation. This is in stark contrast to dual-method theories which posit direct introspective transparency despite there not being a single case where the empirical evidence clearly demonstrates that someone is introspecting without relying on a process of self-interpretation. This should at least shift the burden of proof onto those claiming we have direct introspective access.

The interpretation-only hypothesis seems to be the stronger theory because it offers a unified explanation of all cases in which we consider our own motivations. By contrast, dual-method theorists such as Goldman owe us a principled account of when and why the backup method will be used instead of gaining genuine introspective insight. They not only need to provide an account of why introspection sometimes fails but also an account of why it is that we confabulate when it does. Furthermore, the dual-method theorist needs to explain why we regard confabulations with the same degree of certainty that we do our genuine introspective insights. The findings of Johansson et al. (2005, 2006) which I discussed at the end of Section 2.4 indicate that if there are two methods, they deliver absolutely similar subsequent cognitive processes in all respects. This appears to be a serious anomaly for the dual-method theorist and another area where their account currently lacks a plausible explanation.

Despite my view that the interpretation-only hypothesis is stronger, I think it is unlikely that we are going to be able to make any progress on the debate between the two accounts without significant breakthroughs in our understanding of neuroscience. This is because it is going to be difficult to generate more empirical evidence which would count decisively in favour of one of the two accounts. Further or better empirical demonstrations of confabulation are not going to significantly weaken the dual-method theorist's position because as long as they can claim that cases of genuine introspection

also occur, then the evidence is consistent with their position. All that more demonstrations of confabulation do is suggest that confabulation is less rare than most dual-method theorist think it is. This might undermine some of the intuitive appeal of dual-method accounts much of which I suspect rests on the claim that confabulation is relatively rare. However, even claims about how common confabulation is are going to be hard to establish empirically without complex longitudinal studies. It is clear that cases of inaccurate self-knowledge are not going to provide any further evidence which will help us decide between the two accounts.

If this is the case, then investigating cases in which we gain accurate self-knowledge might help us progress the issue. However, in order for any progress to be made on this front, we will require a method to detect if our correct insights are introspection or accurate self-interpretation. It is currently not clear how this could be achieved.

5 Scepticism About Self-Knowledge

Evidence of confabulation plays a central role in the debate about the mechanisms which give rise to our insights into our own decision-making. While this debate is one I follow with keen interest, it strikes me that it overlooks an important concern which the literature on confabulation gives rise to. The concern is this. If individuals themselves are unaware of when they are confabulating about their own decision-making, then we can never be sure that we are not confabulating. This may be grounds to adopt a degree of scepticism about all our insights into our own decision-making.

This has an important bearing on the consequences of the debate between interpretation-only accounts and dual-method accounts of self-knowledge. While it is independently interesting to discover the kind of mechanism which provides insight into our own decision-making, I suspect that much of the motivation behind arguing for a dual-method account is to preserve the dominant view that we have first-person authority over our own decision-making. The benefit of maintaining that we do have introspective access is supposed to be that this allows us to treat the information we generate about own decision-making with a high degree of certainty. However, it is not at all clear that dual-method accounts do lead to this conclusion. Because from a first-person perspective, cases of introspection are just like cases of confabulation, and this leads to the sceptical concern that we cannot be certain that any particular insight into our own decision-making is not confabulated.

The argument which gives rise to this scepticism is simple:

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.

I foresee two main types of objection to this sceptical concern. The first is that the scepticism is not warranted. Some may argue that the evidence does not justify any sceptical concern at all. Those wishing to peruse this line of objection are likely to deny

premise 2 either by claiming that individuals can tell when they are confabulating or that they can tell when they are not confabulating. This first line of objection can be summarised as the idea that self-knowledge comes with some dependable markers of certainty. I call this the markers of certainty objection. The second line of objection is to concede that there is a sceptical concern, but to argue that we do not need to take this concern serious. This second line of response is the idea that certain forms of scepticism should not be considered a serious threat to our knowledge, and the concerns that I have outlined are one such form of scepticism. I call this the incredulity about scepticism objection. I will respond to each of these objections in turn.

5.1 The Markers of Certainty Objection

Fiala and Nichols (2009, p. 145) responded to an early version of Carruthers' interpretation-only hypothesis (2009a) by claiming that arguments from confabulation are 'toothless' because '…there are systematic differences in confidence levels between confabulation and apparent introspection, which in turn suggests a difference in underlying mechanism'. While I am not trying to advance a claim about the underlying mechanism, if this objection is true, it clearly undermines the argument I need to get my scepticism up and running by showing that premise 2 is false. Fiala and Nichols support their claim by pointing out that in one of the studies most often cited as evidence of confabulation—Nisbett and Wilson's (1977) 'Telling more than we can know'—a typical case of confabulation is reported as beginning 'Gee, I don't really know…' (p. 237). This suggests that although participants in the study did confabulate, they were aware that they were self-interpreting and were not confident about the accuracy of their interpretations.

My line of response to this argument is a development of Carruthers' (2009b) response. Carruthers points out that the existence of some cases in which people can tell that they are confabulating does not show that people are always aware that they are confabulating. He claims that '…there are also a great many instances in which subjects express their metacognitive beliefs unhesitatingly and with high confidence' (p. 169). As long as there are some cases in which people cannot tell if they are confabulating, then my arguments for scepticism stands.

Fiala and Nichols (2009) themselves concede that further systematic study is needed to investigate how well feelings of certainty detect confabulations. I would presume that they were not aware of the work by Johansson et al. (2005, 2006) in which they conducted a comprehensive analysis and found no evidence that participants were less confident about their confabulations. Amongst other things, they checked for 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 and word length. In later studies conducted by the same research team, Hall et al. (2010, 2012) even allowed participants to report if they had been aware of their confabulations once they have been informed that they must have been confabulating. Despite giving participants the opportunity to retrospectively report uncertainty once they know that their explanations cannot be accurate, the vast majority of participants still indicated that they had been totally unaware that they were confabulating. These studies demonstrate that from the first-person perspective, some instances of confabulation are indistinguishable from accurate self-knowledge.

5.2 The Incredulity About Scepticism Objection

You may well think that my sceptical argument succeeds but, nonetheless, not think that the consequences are to be taken seriously. After all, I cannot rule out with any certainty that I am not a brain in a vat. However, I argue that there is one important respect in which the scepticism I am endorsing here is entirely different to the traditional philosophical scepticism which is typically not considered a serious threat to our knowledge. Presumably, one of the reasons why we do not take brain-in-a-vat scepticism seriously is because, although, I have no evidence that I am not a brain in a vat neither do I have any evidence which counts in favour of concluding that I am a brain in a vat. However, the case for scepticism about self-knowledge is quite different. We have substantial empirical evidence demonstrating that confabulation does occur. The scepticism about self-knowledge I am endorsing is empirically motivated scepticism which cannot and should not be dismissed for the same reasons that we might reject traditionally philosophical scepticism.

My opponent may concede that the empirical evidence forces them to take scepticism about self-knowledge more seriously than brain-in-a-vat scepticism, but still not think that scepticism about self-knowledge needs to be taken seriously. They could make another comparison with hallucination scepticism. Hallucinations are often indistinguishable from genuine experience. Furthermore, we have evidence that cases of hallucination do occur. Yet, we do not take seriously the sceptical possibility that any of our experiences could be a hallucination. However, there are two key ways in which self-knowledge scepticism differs from hallucination scepticism.

The first respect in which the self-knowledge scepticism is not like the hallucination scepticism is that people are not good at spotting when they are likely to confabulate. People are most likely to confabulate when their behaviour was determined by factors they were not aware of, particularly if those factors are not the sort of factors that people typically regard as important to decision-making. Nearly all the empirical demonstrations of confabulation have these factors in common. In Gazzaniga's (1995, 2000) split-brain studies and all priming studies, confabulation is induced by stimulus that create associations without the participant being aware of them. Similarly, the decisions that participants are asked to explain in the choice blindness studies are caused by the experimenter's sleight of hand. There is a sharp contrast between this and cases of hallucination. Most people are aware of the kind of factors which might make them hallucinate, such as being seriously dehydrated, sleep deprived or having ingested LSD. If we have no reason to think that we are in circumstances which make us particularly prone to hallucination, then we do not doubt our visual experience. In contrast, we have no idea about when we are in circumstances which make us prone to confabulation and, as a result, have no idea about when we should doubt our self-knowledge.

The second way in which self-knowledge scepticism differs from hallucination scepticism is that we have good reason to think that confabulations are a great deal more common than hallucinations. Years can go by without someone hallucinating. While it is difficult to quantify how common cases of confabulation are, I think that we can be fairly certain that they are more common than this. My personal suspicion is that ordinary people confabulate on a daily basis. My main reason for this suspicion is that the factors which can give rise to confabulation are the sorts of things that are part of our everyday lives. We need only to look at the literature on priming (for a summary,

see Ferguson and Bargh 2004 or Bargh and Chartrand 2000) or situational effects (for a summary, see Doris 2002) to see that people's behaviour is often driven by factors which they are not aware of. Furthermore, these factors are the kind of things which people do not typically regard as important in decision-making. We have reason to believe that such influences are not uncommon and that such influences are likely to give rise to confabulation. However, I do not wish to get dragged into a debate about just how common cases of confabulation are. Whether a fraction, some, or even most of our self-knowledge is actually just confabulations is going to be difficult to establish. Even if it turns out just to be a fraction, then I still think that my scepticism about self-knowledge needs to be taken seriously.

This is because I think that the kind of self-knowledge scepticism which I am proposing is of the same type as the fungi scepticism which Williams (1978, pp. 54–55) discusses in his consideration of Descartes. Williams asks us to consider a man walking in a forest which contains various sorts of fungi. The man knows for certain that some of the fungi are poisonous and that he cannot tell which is which. Williams points out that it would be reasonable for the man to adopt the strategy of not eating any of the fungi because they are all possibly poisonous. I think we are in a similar position with our self-knowledge. We know that some of our self-knowledge is inaccurate, but we do not know which bits of it are.

5.3 Conclusions

The lesson we need to learn from taking scepticism about self-knowledge seriously is that reflecting on our own motivations is a fallible process even when we have a firmly held conviction that we 'know' why we did something. We must abandon the dominant view that we have a reliable and direct insight into our own decision-making. I want to be clear that I do not think we should assume we are always wrong when we consider why we acted. Nor do I think that we should forego all attempts to reflect about our own decision-making. After all, reflecting on our own decision-making is the main way in which we gain information about ourselves. To continue the fungi analogy, suppose that fungi were our main source of food. Then, we would have to take some risks. Williams does consider this scenario and rightly points out that it is no longer sensible to abstain from eating the fungi if there were no other source of food.

Fortunately, the stakes are not analogues to Williams' fungi case. So the risks are not as daunting. In reality, misinterpreting why we acted a certain way is unlikely to lead to death.

Of course, we generally want our reflections on our own motivations to be as error free as possible, and there can be bad consequences when we reach false conclusions about what caused us to act the way we did. These misinterpretations could influence what motivates us in the future or lead us to overlook cases in which we are motivated by reasons, which if we were aware of them, we would reject as good reasons for our actions. The literature on priming and situational effects which I cited in Section 5.2 indicates that such cases are not uncommon. Furthermore, such non-conscious biases can influence important choices. For example cases of implicit gender bias in job recruitment, see Steinpreis et al. (1999). This generates the concern that ignorance of our poor self-knowledge might be a factor reducing our self-control. However, I do not have space to flesh out this concern in detail in this paper.

One important issue to consider is what practices, if any, should change as a result of recognising that we do not have infallible access to our own decision-making processes? Ideally, if avoiding error was always our main concern, we should treat each piece of information yielded by our attempts at introspection with suspicion and try to come up with methods for avoiding and identifying all confabulations. Unfortunately, I do not think it is going to be practically possible to identify and avoid situations when we are particularly likely to confabulate. This is because they are ubiquitous and, by their very nature, difficult to identify from a first-person perspective. The advice to be on the lookout for the influences of factors which you are hiding from yourself is simply practically useless. This makes it tricky to systematically check for confabulations. One thing that we can do is to look for information about our own decision-making taken from other sources, which do not rely on our first-person perspective. Doing this for all our decisions would be such a time-consuming endeavour as to be completely impractical. However, this is something we can and should engage in when we can foresee that there would potentially be bad consequences of making incorrect judgements about our own decision-making. It is not the circumstances which should act as a guide to take precautions against confabulation, but the potential consequences of confabulating about one's motivations for that particular decision. The influence of unwanted motivations is certainly something that we want to avoid when making high-stake decisions which could have a major impact on our or someone else's life. One lesson we should learn from scepticism about self-knowledge is that when the stakes are high, we should try to be aware of the fallibility of our self-knowledge and seek information about our decision-making from external sources.

Acknowledgments Thanks are due to George Botterill who made helpful comments on several drafts of this paper and to the Leverhulme Trust whose research project grant awarded to the "bias and blame" project funds my current position.

References

Bargh, J. A., & Chartrand, T. L. (2000). The mind in the middle: A practical guide to priming and automaticity research. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology*. New York: Cambridge University Press.

Bargh, J. A., Gollwitzer, P. M., Lee-Chai, A., Barndollar, K., & Troetschel, R. (2001). The automated will: nonconscious activation and pursuit of behavioural goals. *Journal of Personality and Social Psychology, 81*, 1014–1027.

Brasil-Neto, J., Pascual-Leone, A., Valls-Solé, J., Cohen, L., & Hallett, M. (1992). Focal transcranial magnetic stimulation and response bias in a forced choice task. *Journal of Neurology, Neurosurgery, and Psychiatry, 55*, 964–966.

Carruthers, P. (2009a). How we know our own minds: the relationship between mindreading and metacognition. *Behavioral and Brain Sciences, 32*, 121–138.

Carruthers, P. (2009b). Mindreading underlies metacognition. *Behavioral and Brain Sciences, 32*, 164–176.

Carruthers, P. (2010). Introspection: divided and party eliminated. *Philosophy and Phenomenological Research, 80*, 76–111.

Carruthers, P. (2011). *The opacity of mind: An integrative theory of self-knowledge*. Oxford: Oxford University Press.

Cohen, A. (1962). An experiment on small rewards for discrepant compliance and attitude change. In J. Brehm & A. Cohen (Eds.), *Explorations in cognitive dissonance*. New York: Wiley.

Doris, J. M. (2002). *Lack of character*. New York: Cambridge University Press.

Ferguson, M. J., & Bargh, J. A. (2004). How social perception can automatically influence behaviour. *Trends in Cognitive Science, 8*(1), 33–39.
Festinger, L., & Carlsmith, J. (1959). Cognitive consequences of forced compliance. *Journal of Abnormal Psychology, 58*, 203–210.
Fiala, B., & Nichols, S. (2009). Confabulation, confidence, and introspection. (Commentary on Peter Carruthers). *Behavioral and Brain Sciences, 32*(2), 144–145.
Gazzaniga, M. (1995). Consciousness and the cerebral hemispheres. In M. Gazzaniga (Ed.), *The cognitive neurosciences*. Cambridge: MIT Press.
Gazzaniga, M. (2000). Cerebral specialization and inter-hemispheric communication: cerebral specialization and inter-hemispheric communication: does the corpus callosum enable the human condition? *Brain, 123*, 1293–1326.
Goldman, A. (2006). *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*. Oxford: Oxford University Press.
Goldman, A. (2009). Replies to commentators. *Philosophical Studies, 144*, 477–491.
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, 17*, 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. doi:10.1371/journal.pone.0045457.
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., Tarning, B., & Lind, A. (2006). How something can be said about telling more than we can know: on choice blindness and introspection. *Consciousness and Cognition, 15*, 673–692.
Kozuch, B., & Nichols. (2011). Awareness of unawareness: folk psychology and introspective transparency. *Journal of Consciousness Studies, 18*(11–12), 135–160. 26.
Latané, B., & Darley, J. M. (1970). *The unresponsive bystander: Why doesn't he help?* New York: Appelton-Century Crofts.
Nisbett, R. E., & Wilson, T. (1977). Telling more than we can know. *Psychological Review, 84*, 231–295.
Srull, T. K., & Wyer, R. S., Jr. (1979). The role of category accessibility in the interpretation of information about persons: some determinants and implications. *Journal of Personality and Social Psychology, 37*, 1660–1672.
Steinpreis, R., Anders, K., & Ritzke, D. (1999). The impact of gender on the review of the curricula vitae of job applicants and tenure candidates: a national empirical study". *Sex Roles, 41*(7/8), 509–528.
Williams, B. (1978). *Descartes: The project of pure enquiry* (pp. 54–55). Harmonsworth: Penguin.