Interpretivism, first-person authority and confabulation

Eivind Balsvik[1]

Abstract:

Psychological experiments allegedly show that people have a tendency to confabulate explanations of their behavior, because their conscious selves do not know why they do what they do, and therefore create the explanations that make most sense. This article explains why confabulation is neither a threat to interpretivist social science, nor a threat to the presumption of first-person authority in Davidson's interpretation theory. The reason is that the interpretative endeavor, which is necessary in order to identify and provide evidence for confabulation is governed by a presumption of first-person authority. Explanations of confabulation thus depend on prior interpretations.

Key words:

Confabulation, interpretation theory, interpretivism, first-person authority

1. Introduction

A basic tenet of interpretivist social science is that in order to describe social phenomena adequately, a researcher must grasp the participating agents' own understanding of their actions, the situations in which they find themselves, and the rules that are constitutive of the institutions in which they take part. Social actors are "self-interpreting animals" (Taylor

[1] Faculty of Social Sciences, University of Oslo, Norway.

Corresponding author: Eivind Balsvik, University of Oslo, PB. 1084 Blindern, 0317 OSLO, NORWAY.
E-mail: eivind.balsvik@sv.uio.no

1985, 45), whose beliefs and desires, values and preferences "enter constitutively into what they do" (Giddens 1976, 13). Adequate descriptions of social phenomena therefore require that social scientists engage in a "double hermeneutic" where the object is to interpret how "knowledgeable agents" conceive of their own actions and the institutions they take part in (ibid.). Davidson's interpretation theory can be used to motivate interpretivist social science, because the presumption of first-person authority, which Davidson shows is a necessary presupposition in interpretation, also underlines that adequate interpretations must capture the agent's self-understanding. According to the presumption of first-person authority, people's self-attributions of psychological predicates should in the first instance be interpreted as being true, without need of supplementary evidence, even if error and correction is possible.

However, more and more psychological studies seem to challenge the presuppositions of privileged and authoritative self-knowledge. I shall concentrate on studies, which allegedly show that people have a tendency to confabulate explanations of their behavior, because their conscious selves do not know why they do what they do, and therefore create the explanations that make most sense (Nisbett and Wilson 1977; Gazzaniga 2011; Hall *et al.* 2010; Johansson *et al.* 2005, 2006, 2012). In the light of such studies, Davidson's interpretation theory and interpretivist social science might seem to rest on a flawed methodology.

The structure of my article is as follows: Section 2 offers a brief summary of Davidson's interpretation theory. Section 3 presents the confabulation-data and theories motivated by the confabulation-data that are skeptical of privileged and authoritative self-knowledge. Such theories seem to challenge the tenability of interpretivist social science and

the presumption of first-person authority. In section 4, I discuss the epistemic status of the presumption of first-person authority. In order to accommodate the confabulation-data, I introduce Henderson's (1993) distinction between first-approximation schemes and refined interpretation. With this distinction in hand, I argue that the epistemic status of the presumption of first-person authority is a priori when constructing a first-approximation scheme, but that it should be regarded as an empirical hypothesis in refined interpretation. In Section 5, I suggest that a combination of dual-method theory of self-knowledge and dual-systems theory of mental processing can be used in order to explain when first-person authority can be relied upon in refined interpretation, and when it cannot. In section 6, I conclude that it is possible to incorporate the psychological findings on confabulation, without abandoning the main tenants of interpretivism and Davidsonian interpretation theory. The presumption of first-person authority is necessary for the ascription of content, and for describing the point of view of those who are being studied. However, explanations of social behavior and social phenomena might deviate from how the agents themselves understand their own actions.

2. Davidson's theory of interpretation

Inquiring into the beliefs and desires of social actors is a peculiar kind of enterprise. Since it is only by saying things and doing things that social actors can express or reveal their mental contents, social scientists only have behavioral evidence to go on when interpreting their beliefs and desires. According to Davidson (1975, 160) "[t]his creates a problem",

> for it means that behavior, which is the main evidential basis for attributions of belief and desire, is reckoned the result of two forces less open to public observation. Thus where one constellation of beliefs and desires will rationalize an action, it is always

> possible to find a quite different constellation that will do as well. Even a generous sample of actions threatens to leave open an unacceptably large number of alternative explanations.

The main challenge for interpretation theory is therefore to offer an account of how an interpreter can manage to narrow down the large number of possible, alternative interpretations. According to Davidson, there must be certain a priori requirements or principles that constrain adequate interpretations, for interpretation to be possible at all (Davidson 1975, 168-69). Firstly, a requirement of holism excludes a large number of interpretations. The requirement that adequate interpretation must fit a holistic pattern entails that the interpretations of what an agent says or does must cohere with the interpretation of a set of related interpretations for other utterances and actions, and that these interpretations are conditioned in the same way (Davidson 1970, 221). Secondly, when interpreting a range of behavior, the holistic pattern must be a fundamentally charitable one. The principle of charity requires that we as interpreters see others as "more or less rational creatures mentally inhabiting a world much like our own" (Davidson 1984b, 36). Although rationality is a matter of degree, "there must be enough rationality in the complete pattern for us to judge particular beliefs as foolish or false, or particular acts as confused or misguided" (ibid.). The underlying reason is that it is "only in a largely coherent scheme" that propositional contents can "find a lodging" (ibid.).

And thirdly, "[w]hen a speaker avers that he has a belief, hope, desire or intention, there is a presumption that he is not mistaken, a presumption that does not attach to his ascriptions of similar mental states to others" (Davidson 1984a, 3). By listening to what an individual has to say, it is possible to make very fine-grained descriptions of how he conceives of his own actions. Davidson holds that interpretation would not be possible

without a presumption of first-person authority: “People are in general right about the mental causes of their emotions, intentions, and actions because as interpreters we interpret them so as to make them so. We must, if we are to interpret at all” (Davidson 1976, 290).

3. The confabulation-data and Theory-theories of self-knowledge

Psychological studies, which supposedly demonstrate that people have a tendency to confabulate reasons for their choices and behavior, challenge the practice of relying extensively on verbal reports, as is common in interpretivist social science. This section offers a brief presentation of the research on confabulation, and introduces so-called Theory-theories of self-knowledge, which are launched in order to explain the confabulation-data. According to such theories, although self-knowledge often seems to be immediate and direct, it is actually the result of applying folk psychological theories in order to interpret or infer mental states to which one has no direct, introspective access. I shall focus mostly on Carruthers Interpretive Sensory-Access theory, since this is the most sophisticated Theory-theory on the market.

Psychologists have discovered that patients who suffer from organic amnesia, patients with spilt-brains, and participants in experiments who act on posthypnotic suggestion, “easily generate stories to explain their behavior, with no realization that their explanations are works of fiction” (Wilson 2002, 96-97). The phenomenon is called confabulation, and it is defined as “giving a fictitious account of a past event, believing it to be true” (Gazzaninga 2011, 77). Gazzaniga and Ledoux (1978) hypothesize that the actual causes of the tendency to confabulate might not be organic amnesia, split-brains, or

hypnosis *per se*, but suggest that these conditions make it easier to observe what they take to be "a common human tendency to confabulate" (Wilson 2002, 97).

Gazzaniga's experiments on split-brain patients are particularly telling. These patients have for various reasons undergone surgery to sever two fiber pathways that connect the brain's hemispheres: the *corpus callosum* and the *anterior commissure*. Consequently, information cannot be transferred between the left and the right brain hemispheres. This offers researchers the opportunity to provide different stimuli to each of the two brain hemispheres. Since it is the left hemisphere that controls speech, only stimuli presented to the left hemisphere can be verbally processed.[2]

In one of Gazzaniga's experiments, the instruction "Walk!" was flashed to the right brain-hemisphere of a split-brain patient, known as Joe. Joe immediately got up from his chair and walked out of the van in which the experiment took place. When asked why he left the van, he sincerely replied that he wanted to go to the house to get a Coke. In this experiment, Gazzaniga thinks that Joe's self-attributed intention is a confabulation, and that his behavior was in fact triggered by the instruction to walk that was flashed to his right brain-hemisphere. Nevertheless, Joe did not seem to be aware that his explanation was confabulated. Since Joe's left-brain did not have information about the instruction to walk, his left-brain interprets the limited information that is available to it in a way that makes sense of him leaving the van.[3]

Nisbett and Wilson's "Telling more than we can know: Verbal reports on mental processes" (1977) is supposed to provide evidence of widespread confabulation among normal subjects. The article assembles a number of experiments, which demonstrate that

[2] See Gazzaniga (2011, 53-60) for more about split-brains.
[3] For discussion, see Carruthers (2011, 39-40); Peters (2014, 584-85); Scaife (2014, 472).

people are not at all good at detecting influences on their preferences, choices, or behavior. Nonetheless, people easily make up justifications for their choices. In the much discussed Panty-hose study, which was conducted at a shopping mall, Nisbett and Wilson placed four identical pairs of stockings on a table, and asked people passing by which pair they preferred and what their reasons were. From a previous version of the experiment, they knew that people tend to prefer items on the right side of the display. The aggregated preferences in the Panty-hose study confirm that there is a position effect. Whereas the two pairs on the left side of the display were preferred by only 12 and 17 percent of the participants, the pairs to the right were preferred by 31 and 40 percent (Wilson 2002, 102-3). However, the participants did not realize that the position of the stockings had an effect on the choice they made. When asked about the reasons for their choices, "no subject ever mentioned spontaneously the position of the article in the array" (Nisbett and Wilson 1977, 243–4). Rather, they typically pointed to an attribute of their preferred pair, such as its superior knit, sheerness, or elasticity. When Nisbett and Wilson asked people directly whether they thought the position of the panty-hose had influenced their choice, "virtually all subjects denied it, usually with a worried glance at the interviewer suggesting that they felt either that they had misunderstood the question or were dealing with a madman" (Nisbett and Wilson 1977, 233). The participants seemed to know which pair of stockings they preferred, but not why.

Whereas Nisbett and Wilson's Panty-house experiment demonstrates confabulation at the group-level, Johansson and his colleagues (2005a; 2006; 2012; Hall *et al.* 2010) have created a novel experimental design which makes it possible to demonstrate for each of the individuals participating in their experiments, whether the reasons they provide for their choices are confabulatory or not. Since their research tool is a modification of the change

blindness paradigm, they call it the choice blindness paradigm (CBP). In their first series of experiments, participants were shown pairs of photographs of female faces, and were asked to select the most attractive one. When they had made their choice, both photographs were laid facedown on the table, before the experimenter re-presented "the photograph", and asked them why they had chosen it. Without the participants realizing it, on some trials, a double-card ploy was used, and the participants were in fact shown the rejected photograph from the pair. So on these trials, the outcome of the choice became the opposite of what the participants actually intended. Johansson *et al* (2005a, 117) compares this with a situation where you ask "for a bottle of beer … [but] end up with a glass of milk in your hands". However, would the participants notice that the photographs were sometimes switched? If not, how would they explain their choice, and would there be any differences between confabulatory and "real" reasons for choices?

The results of the experiment are quite astonishing. Firstly, the number of manipulated trials detected by the participants was low. Even when participants were offered unlimited time to make a choice, and the two photographs were quite dissimilar, only 27 % of the manipulated trials were detected. Secondly, the kinds of reasons that the participants offered for manipulated and non-manipulated choices did not differ significantly in terms of emotionality, specificity, certainty, deceit and complexity (Johansson *et al* 2006a, 678-684). This similarity could imply that justifications for "real" choices are based on the same processes that generate justifications for confabulatory choices. Thirdly, sometimes the reasons given by the participants were specific to the rejected photograph, such as "I prefer blondes," in a trial where the selected photograph was of a brunette. These reports are obviously not based on remembering the considerations that led to the original choice. Johansson and his team are cautious in their conclusion, but state that their research

"suggest that introspections about (some forms of) decisions may (sometimes) be confabulatory" (Hall *et al.* 2006, 697). They have continued employing the CBP to different kinds of choices, also in natural settings, with similar results as in the original study. For instance, Hall *et al.* (2010) studied justifications for consumer choices by setting up a tasting venue for teas and jams at a supermarket; and Johansson *et al.* (2012) studied justifications for risky choices and political opinion by modifying the CBP. In the latter experiment, the participants were sometimes asked to justify the opposite opinions of what they had just previously answered on a political survey. Remarkably, the participants noticed only 20 % of the manipulated answers.

Some psychologists and philosophers believe the confabulation-data have radical implications for theories of self-knowledge. For instance, Nisbett and Wilson (1977, 248) explain their findings by suggesting that

> [w]hen people are asked to report how a particular stimulus influenced a particular response, they do not do so by consulting a memory of the mediating process, but by applying or generating causal theories about the effects of that type of stimulus on that type of response.

Unconsciously, people interpret their own behaviour by using such theories, because they are largely ignorant of the "cognitive processes underlying [their] choices, evaluations, judgments, and behavior" (1977, 231). Cases of confabulation are therefore understood as failed attempts at self-interpretation. The upshot is that people have "no more certain knowledge of [their] own mind[s] than would an outsider with intimate knowledge of [their] history and of the stimuli present at the time the cognitive process occurred" (1977, 257).

Nisbett and Wilson therefore thought that they had provided an empirical refutation of the presumption of authoritative and privileged self-knowledge of mental states and processes.

Carruthers (2011) uses the confabulation-data to motivate his Interpretive Sensory-Access (ISA) theory of self-knowledge. The ISA-theory is the most elaborate theory launched explicitly in order to account for the confabulation-data. According to the ISA-theory, a single "mindreading" mental faculty is responsible for the attribution of mental states both to oneself and to others. This faculty receives only sensory input. The theory therefore allows that we have transparent access[4] to our own sensory states, but it holds that we lack transparent access to our non-sensory states, such as beliefs, desires, intentions and decisions. In order to self-attribute thoughts, the mindreading faculty interprets the available sensory evidence. This can concern one's physical circumstances and overt behavior, or it can involve one's own visual imagery, affective feelings, and inner speech (Carruthers 2011, 2). The upshot is that just as interpretation is needed in order to figure out what someone else is thinking, so too is interpretation required to figure out what oneself is thinking (2011, 1). Moreover, since people are often misled when interpreting others, they should often be misled when they interpret themselves. In fact, "the ISA theory predicts that confabulation should occur whenever there is sensory evidence of a sort that might mislead a third party (2011, 6)."

Since the process of interpretation is typically "unconscious in character," people will experience themselves as reaching decisions or making judgments, without awareness that their attributions of attitudes to themselves are interpretive in nature (2011, 3). According to the ISA-theory, it should therefore be no mystery that "confabulating agents should

[4] Carruthers uses the term 'transparent access,' to refer to what philosophers and psychologists call privileged access.

generally be under the impression that they are merely introspecting" (2011, 6). Carruthers argues that we cannot use our own strong impression of having transparent access to our own mental states as evidence that interpretation is not occurring, because the split-brain patients who are known to confabulate explanations of their own behavior, and therefore must be interpreting themselves, also have a strong impression of having transparent access to their own mental states. In order to explain the universal intuition that we have transparent access to our own thoughts in a way that does not entail that the intuition is true, Carruthers suggests that "some form of belief in transparency of mind might be innate" (2011, 34), or that "a tacit assumption of self-transparency [is] built into the structure of the mindreading system" (2011, 46).

4. The epistemic status of the presumption of first-person authority

Nisbett and Wilson (1977) and Carruthers (2011) are not alone in thinking that the research on confabulation undermines the authority and privilege of self-attributions of mental states and processes. Since the subjects of the experiments that demonstrate confabulation are not able to tell whether they are confabulating, Scaife (2014, 471) argues that the confabulation-data "leaves open the skeptical possibility that any time [people] consider [their] own motivations, [they] might not be getting accurate information." Moreover, Gazzaninga (2011, 77) warns us that when people attempt to explain their actions, their explanations "are all *post hoc* using *post hoc* observations with no access to nonconscious processing." Therefore, "listening to people's explanations of their actions is … often a waste of time (Gazzaninga 2011, 78)." Similarly, Nisbett (2015, 203) sums up his chapter, "Don't ask, can't tell" by stating, "*[r]eports about the causes of our behavior … are susceptible to a*

host of errors and incidental influences. They're frequently best regarded as readouts of theory, innocent of any "facts" uncovered by introspection."

The confabulation-data therefore challenge the tenability of the presumption of first-person authority in Davidson's interpretation theory. Since I agree with Davidson that the presumption of first-person authority is necessary for interpretation to be possible, and since I grant that confabulation does occur, I need to provide an account of the epistemic status of the presumption of first-person authority, which is compatible with the confabulation-research. In order to accommodate the confabulation-data, I introduce Henderson's (1993) distinction between first-approximation schemes and refined interpretation. Whereas the main purpose of first-approximation schemes is to attribute meaning, the main goal of refined interpretation is to provide explanations. With this distinction in hand, I argue that it would be impossible to identify and provide evidence for confabulation without a presumption of first-person authority. Whereas the epistemic status of the presumption of first-person authority is *a priori* when constructing a first-approximation scheme, I argue that it should be regarded as an empirical hypothesis in refined interpretation.

To begin with, notice that discarding the presumption of first-person authority altogether would make it impossible to identify and describe cases of confabulation and other sorts of irrationality. First-person authority over occurrent beliefs is presupposed in the very characterization of such phenomena. For instance, unless it is presumed that Joe knows what belief, he expresses when he says that he left the van in order to get a Coke, and knows that he is sincere; his response would not count as an instance of confabulation. In general, without a presumption that the participants in the confabulation studies know what belief they are falsely self-attributing, it would be impossible to identify and provide

evidence for confabulation. This is in line with the basic intuition, which motivates interpretivism in the social sciences. In order to describe social behavior adequately, it is necessary to interpret how social actors view their own behavior. Describing the phenomena that are used to cast doubt on privileged and authoritative self-knowledge therefore presupposes a presumption of first-person authority. Therefore, the problem of confabulation concerns privileged and authoritative access to *mental processes*, and primarily affects explanations of behavior.

So perhaps the presumption of first-person authority is a *simplifying, but untrue heuristic*, which researchers employ as a preliminary to getting on with the central business of the social sciences, namely explanation. This proposal is reminiscent of how Henderson (1993) thinks the principle of charity should be adjusted in the light of psychological studies that demonstrate systematic deviations from the normative requirements of rationality. Henderson argues that the principle of charity should be a rather strong constraint in the early, preparatory, stages of interpretation, in which an interpreter lays the basis for a less charitably constrained investigation, where some beliefs and actions might be explained as seriously mistaken, or irrational (Henderson 1993, 41). He therefore distinguishes between first-approximation schemes and refined interpretive schemes. Whereas the primary goal of first-approximation schemes is to attribute meaning to verbal and non-verbal actions, the goal of refined interpretation is to offer explanations by uncovering the mental processes that are the causes of actions. Henderson offers two reasons why the principle of charity must be a strong constraint on interpretation when developing a first-approximation scheme, which also applies to the presumption of first-person authority. Firstly, it is only by using a first-approximation scheme that an interpreter can identify and provide evidence for possible cases of irrationality in a sufficiently precise way to warrant further investigation.

Secondly, it is not possible to provide evidence for irrationality without using such a scheme (Henderson 1993, 52). Henderson concludes that first-approximation schemes "provides the context within which … anomalous belief or action can be examined" (Henderson 1993, 53). The role of the principle of charity and the presumption of first person authority can therefore be understood as preparatory for "the central business of the social sciences: explanation" (Henderson 1993, 55). In refined interpretive schemes, the main concern is for explicability, and interpreters should be "indifferent to whether what is explained is rational (or correct) belief and action" (Henderson 1993, 60).

Although Davidson would agree with Henderson that adhering to the principles that constrain interpretation is more important in the early stages of investigation, than in the later, he would not agree to characterize these principles as being untrue, but simplifying heuristics. Since Davidson (1975; 1984a) thinks these principles are necessary presuppositions for the attribution of mental content, perhaps he would agree to characterize them as being *a priori, yet empirically defeasible presuppositions of interpretation*. That the warrant is *a priori* means that its justificational force does not derive from experience; that it is *defeasible*[5] means that it is open to empirical correction.

Hopkins' (1999) puts a spin on Davidson's position, which highlights the differences between the Davidsonian view and Henderson's alternative. When interpreting an individual's actions, the interpreter must infer the beliefs and desires that motivated and guided the actions from observations of what the individual says and does. When interpreting what an informant says, the interpreter does not merely attribute meaning to the informant's utterances, but interprets them as expressions of desires, beliefs, or

[5] First-person authority in speech might be defeated by insincerity, slips of the tongue, and perhaps some instances of sincere attributions of beliefs that are not matched by appropriate actions (e.g. Balsvik 2003, 109-125; Schwitzgebel 2010).

intentions. Since beliefs and desires provide impetus and guidelines for action, an interpreter can expect a close connection to obtain between what a sincere individual says and what he does. As a number of philosophers emphasize, to self-attribute a belief is to express a commitment to its truth: to be willing to assert it, defend it, take it as a premise, and to act in accordance with it (Frankish 2016; Schwitzgebel 2010). The presumptions of first-person authority and rationality therefore entail a *presumption of 'normative accord'* between an individual's utterances and actions (Hopkins 1999, 276).

In order to test whether or not an individual's self-attributions of beliefs and desires are in normative accord with his conduct, an interpreter should correlate and compare this individual's self-attributions (verbal behavior) with his other actions (both verbal and non-verbal), in order to triangulate on the common set of beliefs, desires and intentions which caused them (Hopkins 1999, 276). Whenever the informant is sincere and rational, and the presumption of first-person authority is not defeated, the method should result in precise and accurate interpretations. The interpretation of his utterances permits a more accurate and precise understanding of his actions, and the observation of his actions confirms and strengthens the interpretation of the utterance. In general, it would neither be possible to interpret the beliefs and desires expressed by an individual's utterances, nor the intentions underlying an individual's actions, with any degree of accuracy and certainty, unless it is possible to correlate and compare what an individual says with what he does. Echoing Kant's first *Critique* (B 75), Hopkins states that "words with no relation to deeds would be unintelligible, and deeds with no relation to words would be inarticulate" (Hopkins 1999, 276).

The action-guiding role of beliefs and desires thus ensures a convergence between an individual's authoritative warrant for his self-attribution, and the justification an interpreter

has for attributing the self-attributed attitudes to him. The two kinds of justification normally converge because when a self-attribution of belief is sincere, the speaker knows he believes what he claims to believe, and hence intends to act in such a way that his subsequent behavior can be expected to continue to support the belief he has self-attributed. Unless there is a presumption of first-person authority, it would not make sense to correlate and compare an individual's self-attributions with his actions, and it would not be possible to make the precise and accurate interpretations that we actually succeed in making.

Even though an individual's self-attributions of propositional attitudes do not require justification from the individual making the attribution in order to count as authoritative, they nevertheless provide the interpreter with a default warrant for attributing these attitudes to the individual. Further justification is required only when specific interpretative difficulties arise. For instance, if there is a discrepancy between what an individual says and does, an interpreter might need to justify the attribution of sincerity, rationality or first-person authority. In such cases, Davidson could agree with Henderson that the interpreter should be informed by psychological theory.

Nevertheless, Davidson's position is problematical because it seems to put *a priori* limits on what empirical findings there could be regarding lack of self-knowledge and irrationality (Henderson 1993, 6). Take Joe, the split-brain patient as an example. If he were to insist on getting a drink after having been queried why he left the van, the method of testing for normative accord would take this as confirmation that he left the van in order to get a Coke. However, the fact that he was flashed the instruction to walk just before he left the van, suggests that this cannot be the complete explanation. A better explanation would perhaps be that he left the van because he was flashed the instruction to walk. When

queried why he left the van, he confabulated a motive, which then became a reason for action. In holding that interpreters are methodologically obligated to interpret people as far as possible as being rational, consistent, and as knowing their own minds, without regard to whether or not there is evidence that people tend to possess these qualities, the Davidsonian position is in tension with the principle that scientific results should be falsifiable in the light of experience. Therefore, I accept Henderson's claim that refined interpretation should "be informed by psychological theory regarding the relative likelihood of various types of error (Henderson 1993, 43)."

Do the confabulation-studies give us reason to reject first-person authority as a likely empirical hypothesis in refined interpretation? Are people generally not aware of the reasons for their choices, or the considerations that motivate their actions? I think there are reasons to resist jumping to this conclusion. Johansson and his colleagues (2005b, 4) are explicit that the CBP-experiments "simulate a choice situation in which no prior evidence indicates that a high level of monitoring is needed." It is therefore reasonable to suppose that the participants in these experiments do not attempt explicitly to memorize the alternatives nor the reasons for their choices. If the participants were motivated to monitor their choices carefully, the detection rate of the manipulated trials would surely have been much higher. Moreover and most importantly, the low detection rate could in large part be explained by the way in which the experiment is conducted (Petitmengin *et al.* 2013, 666). The experiment does not induce the participants to introspect, but rather to focus on the social interaction in which they are engaged. It is reasonable to suppose that the participants trust the experimenter, and that it is important to them to come across as being both competent and respectable. Therefore, when the participants are "re-presented" the

rejected photograph, and asked why they chose it, it is not surprising that they do not introspect, but simply look at the presented photograph and provide reasons for choosing it.

Petitmengin and her colleagues (2013, 655) give a twist to Johansson *et al.*'s (2005) CBP-experiment by introducing an "elicitation interview" for some choices. The elicitation interview lasts 17 to 45 minutes. While the photographs are facedown, the interviewer invites the participants to introspect, and helps them evoke the process that led them to make their choice. The interviewer never asks the participants directly why they chose their photograph, but rather lets the question "how" "guide the subject towards the description of more and more detailed elements of his evoked choice process" (2013, 658). At the end of the interview, the participants are shown the rejected photo, and asked if "*anything else concerning his choice process comes back to him when seeing this picture again"* (ibid.). The participants detected the manipulation in 80 % of these trials. In the trials where the participants did not undergo an elicitation interview, Petitmengin *et al.'s* results were similar to Johansson *et al*.

Siewert (1998, 47) responds to the psychological research by arguing that general skepticism on first-person authority does not have empirical warrant. When people self-attribute a belief, desire or intention, they typically act in ways that confirm that they have the self-attributed state. For instance, if customers at a restaurant have ordered one thing (e.g. a beer), but receive something else (e.g. a glass of milk), they would surely notice the substitution. Although there are not any studies that present evidence which confirm this, Siewert (1998, 49) suggests, "this is not because such evidence would be so scarce, but because of its bland ubiquity." Rey (2013, 273) adds that for a wide range of attitudes, feelings and sensations, there simply are "no serious doubts" about people's reliability, "which is perhaps why there have been no studies."

Moreover, Rey argues that the exceptional reliability of a significant set of self-attributions of attitudes, poses a serious problem for Theory-theories of self-knowledge. Such theories must be able to explain how a third-person inferential process can be as spectacularly reliable as a significant set of self-attributions seem to be. Carruthers wishes to explain the reliability of self-attributions of mental states and processes, by appealing to the fact that there is more evidence available for interpretation in the first-person case than in the third person case (Carruthers 2011, 94). Most importantly, when interpreting our own minds, we have access to our own inner speech. However, Carruthers (2011, 86-88) emphasizes that inner speech only provides indirect evidence of our thoughts and feelings and that this evidence too needs to be interpreted by the mindreading faculty. Since Carruthers (2011, 24) puts so much emphasis on the fact that more evidence does not necessarily lead to more reliable conclusions, the exceptional reliability of some kinds of attitudes do pose a serious problem for his overall theory.

5. First-person authority in refined interpretation

Even if it is probably true that "ordinary people confabulate on a daily basis" (Scaife 2014, 480), I do not think, as Scaife suggests, that the confabulation-data should lead to general skepticism about the authority of self-attributions of mental states and processes in refined interpretation. In this section, I argue that dual-systems theory within psychology can provide plausible criteria for when first-person authority is a likely empirical hypothesis in refined interpretation, and when it is not.

According to dual-systems theory, the mind is composed of two distinct processing systems, commonly referred to as System 1 and System 2. System 1 is supposed to "operate automatically and quickly, with little or no effort and no sense of voluntary control"

(Kahneman 2011, 20). System 1 processing can therefore cause behavior without requiring conscious reflection. System 2 processing is very different. It requires some expenditure of effort and it is typically "associated with the subjective experience of agency, choice, and concentration" (Kahneman 2011, 21). When agents consider different behavioral alternatives, by deliberating about their possible costs and benefits, System 2 is doing the work. Some behaviors are also a result of mixed processing. System 2 might for instance monitor the results of System 1 processing. The motivation for monitoring System 1 might come "from an enhanced desire for accuracy, a sense of accountability, a concern with social desirability, or … [a desire] to control prejudiced reactions" (Fazio and Olsen 2014, 156).

In order for interpreters to be able to rely upon first-person authority and privileged access, we need an account of the conditions under which it is reasonable to expect people to be capable of self-attributing mental states and processes in the privileged and authoritative first-person way, and the conditions under which it is likely that they will resort to unconscious self-interpretation. By combining dual-systems theory with a dual-method theory of self-knowledge, it is possible to provide such criteria (Cf. Wilson 2002, 105-6). On this approach, people will resort to self-interpretation whenever their behavior is caused by the unconscious processing of System 1. Whereas people have privileged access to the results of system 1 processing, they must use self-interpretation in order to attribute their causes. Whenever behavior results from System 1 processing, people will have first-person authority about what they are doing, but they must rely on fallible self-interpretation in order to know why they are doing it. When people's self-attributions rely on self-interpretation, their responses will sometimes be confabulatory. Explanations of behavior caused by the operations of System 1 therefore do not enjoy first-person authority. However, if people's behavior is caused by the conscious deliberation of System 2, they will

be consciously aware of (at least parts of) their own deliberation and their explanations of their own behavior will be authoritative.

In order to be able to rely upon first-person authority when interpreting others, it is necessary to be able to tell which system is most likely to be responsible for the behavior. According to Fazio and Olsen's (2014, 155) MODE-model, "*M*otivation and *O*pportunity serve as the major *De*terminants of which system is likely to operate." Since System 2 processing "requires some expenditure of effort …, the individual must not only be motivated to engage in the effortful analysis but also have the opportunity (i.e., the time and resources) to do so" (Fazio and Olsen 2014, 155). Sometimes behavior is the result of a mixture of System 1 and System 2 processing, in that System 2 monitors the results of System 1 processing for control. In order to execute such control, an individual must have motivation and opportunity to engage in System 2 reasoning. For instance, "one might be motivated to gauge the appropriateness, or even counter the influence, of an automatically-activated attitude" (Ibid. 156). According to Fazio and Olsen, opportunity "is essentially a gating mechanism" (ibid.). If an individual has the time and resources to engage in System 2 reasoning, an individual can counter or correct for the influence of System 1 processing. However, if the person to be interpreted "is fatigued or cognitively depleted", it is less likely that an individual will be motivated to engage in the conscious and relatively demanding System 2 deliberation (ibid.). Moreover, if the situation "demands an immediate response," there will be little opportunity to engage in conscious deliberation, and it is therefore likely that judgments or behavior are produced by the fast, but unconscious, System 1 (ibid.).

According to the view defended here, people can come to know about their own minds by employing two different methods. The first method requires conscious awareness

and provides mental attributions that enjoy first-person privilege and authority. The second method is through turning our faculty for mindreading others, onto our own selves, as described in Carruthers' ISA-theory. This method, although often more reliable than third-person interpretations, because its attributions are based on richer data, does not enjoy first-person authority. Cases in which people confabulate attitudes are ones in which they are somehow unable to employ the authoritative way of self-attributing mental processes, or where there is something about the circumstances that explains the tendency to confabulate. However, when developing a first-approximation scheme, the presumption of first-person authority is necessary for the ascription of mental content. Interpretation would not be possible without it.

Carruthers (2011, 366) thinks dual-method theories of self-knowledge should be rejected because such theories "need to avail themselves of all of the resources of the ISA theory while at the same time postulating something extra, such as mechanisms of inner sense." Rey (2013, 267) responds convincingly that considerations of simplicity should not make us reject dual-method theory because "[r]edundancy of information from different perspectives can fortify confidence, and it could serve us well sometimes to monitor ourselves 'directly', and sometimes also to see ourselves as others see us—even if one or the other way may sometimes trump the other."

6. Conclusions

The purpose of this article was to explain why confabulation is neither a threat to interpretivism in the social sciences, nor a threat to the presumption of first-person authority in Davidsonian interpretation theory. My argument has utilized Henderson's (1993) distinction between first-approximation schemes and refined interpretation, and the

distinction within dual-systems theory in psychology between the fast, automatic and unconscious processing of System 1, and the slow, controlled and conscious processing of System 2. It also depends upon a dual-method theory of self-knowledge.

When constructing a first-approximation scheme, an interpreter should at first take people's sincere self-attributions of mental states and processes to be true, without need of supplementary evidence, even if error and correction is possible. The presumption of first-person authority is necessary for the attribution of mental content, and for describing the point of view of those who are being studied. When constructing a first-approximation scheme, the justification for the presumption of first-person authority does not depend upon a particular account of the epistemology of self-attributions of psychological predicates. If the interpreter wishes to make fine-grained attributions of mental states, such as beliefs, intentions, desires and emotions, the interpreter must regard his subjects' self-attributions as being authoritative.

Whereas the presumption of first-person authority has an a priori status when constructing a first-approximation scheme, first-person authority is not necessarily the most likely empirical hypothesis in refined interpretation. By combining dual-systems theory of mental processing with dual-method theory of self-knowledge, we can see why explanations of social behavior and social phenomena might deviate from how the agents themselves account for their own actions. People do not have conscious access to the fast and automatic processing of System 1, and will therefore rely upon unconscious and biased self-interpretation in order to make sense of behaviors that are caused by such processing. In refined interpretation, whether or not subjects are authoritative when reporting on their mental lives, depends upon the epistemology of their self-reports. I have assumed that

conscious deliberation or conscious monitoring gives rise to epistemic privilege and therefore justifies taking subjects to be authoritative when reporting on their own mental lives. Whether or not social actors have engaged in conscious deliberation or conscious monitoring before acting is an empirical question. Fazio and Olsen's (2014) MODE-model offer plausible criteria for when interpreters can rely upon first-person authority in refined interpretation, and when they cannot do so.

Although explanations in the social sciences need not be rationalizing, the descriptive and interpretative endeavor, which is necessary in order to identify and provide evidence for confabulation is governed by a principle of first-person authority. Theories about the human tendency to confabulate require epistemic support from interpretations, which are guided by a presumption of first-person authority. Explanations of confabulation thus depend on prior interpretations. If the above is correct, then we should distinguish between interpretation and explanation in the social sciences. Interpretation is necessary for descriptive purposes, and for uncovering the meaning of social action. If actions are consciously controlled, the interpretation of those actions will also offer explanations. However, explanations of actions do not need to be rationalizing. My position is therefore not a naïve form of interpretivism in which the social scientist should simply take an individual's self-attributions at face value. It is consistent with describing social actors as being wrong, confused or deluded. I conclude that it is possible to incorporate the psychological findings on confabulation, without abandoning the main tenants of interpretivism and Davidsonian interpretation theory.

Acknowledgements

I would like to thank the reviewers of this article for very useful comments. I would also like to thank Edmund Henden, Lars Klemsdal, Erik Lundestad, Lars Nyre, Cato Wittusen and the audience at the Conference of the European Network for the Philosophy of the Social Sciences (ENPOSS) 2016 for their advice and remarks on a previous draft of this article.

References

Balsvik, Eivind. 2003. *An Interpretation and Assessment of First-Person Authority in the Writings of Philosopher Donald Davidson.* Lewiston, Queenston, Lampeter: The Edwin Mellen Press.

Carruthers, Peter. 2011. *The Opacity of Mind. An Integrative Theory of Self-Knowledge*. Oxford: Oxford University Press.

Davidson, Donald. 1970. Mental Events. Reprinted in his *Essays on Actions & Events*. Oxford: Clarendon Press 1980: 207-225.

_____. 1975. Thought and Talk. Reprinted in his *Inquiries into Truth and Interpretation*. Oxford: Clarendon Press 1984: 155-70.

_____. 1976. Hume's Cognitive Theory of Pride. Reprinted in his *Essays on Actions & Events*. Oxford: Clarendon Press 1980: 277-290.

_____.1984a. First Person Authority. Reprinted in his *Subjective, Intersubjective, Objective.* Oxford: Clarendon Press 2001: 3-14.

_____.1984b. Expressing Evaluations. Reprinted in his *Problems of Rationality*. Oxford: Clarendon Press 2004: 19-37.

_____. 1991. Three Varieties of Knowledge. Reprinted in his *Subjective, Intersubjective, Objective.* Oxford: Clarendon Press 2001:205-220.

Fazio, Russel H. and Olson, M. A. 2014. The MODE model: Attitude-Behavior Processes as a Function of Motivation and Opportunity. In Sherman, J. W., Gawronski, B. & Trope, Y. (eds). *Dual-Process Theories of the Social Mind*. New York: Guilford Press.

Gazzaniga, Michael S. 2011. *Who's in charge? Free will and the sciences of the brain*. New York: Harper Collins Publishers.

Giddens, Anthony. 1976 [1993]. *New Rules of Sociological Method.* 2nd edition. Cambridge: Polity Press.

Goldman, Alvin I. 2006. *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading.* Oxford: Oxford University Press.

Hall, Lars; Johansson P., Sikström, S., Tärning B., Lind, A. 2006. Reply to commentary by Moore and Haggard. *Consciousness and Cognition*, 15: 697-699.

Hall, Lars; 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.

Henderson, David K. 1993. *Interpretation and Explanation in the Human Sciences*. Albany: State University of New York Press.

Hopkins, James. 1999. Wittgenstein, Davidson, and radical interpretation. In Hahn, Lewis Edwin (ed.). 1999. *The Philosophy of Donald Davidson*. The Library of Living Philosophers. Vol. XXVII. Chicago and LaSalle, Illinois: Open Court.

Johansson, Petter, Hall, L., Sikström, S., Olsen, A. 2005a. Failure to detect mismatches between intention and outcome in a simple decision task. *Science*, Vol 310, Issue 5745: 116-119.

Johansson, Petter, Hall, L., Sikström, S., Olsen, A. 2005b. Supporting online material for "Failure to detect mismatches between intention and outcome in a simple decision task." *Science*, Vol 310, Issue 5745: 14 pages, not numbered.

Johansson, Petter, Hall, L., Sikström, S., Tärning 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.

Johansson, Petter, Hall, L., Chater, N. 2012. Preference change through choice. In Dolan, R. & Sharot, T. (eds.) *Neuroscience of Preference and Choice: Cognitive and Neural Mechanisms.* Elsevier, Science direct: 121-141.

Kahneman, Daniel. 2011. *Thinking, fast and slow.* London: Allen Lane, Penguin books.

Nisbett, Richard E. 2015. *Mindware. Tools for smart thinking.* New York: Farrar, Straus and Giroux.

Nisbett, Richard E. and Wilson, Timothy D. 1977. Telling More Than We Can Know: Verbal Reports on Mental Processes. *Psychological Review* 84: 231-259.

Peters, Uwe. 2014. Interpretive sensory-access theory and conscious intentions. *Philosophical Psychology,* Vol. 27, (4): 583-595.

Petitmengin, Claire, Remillieux, A., Cahour, B., Carter-Thomas, S. 2013. A gap in Nisbett and Wilson's findings? A first-person access to our cognitive processes. *Consciousness and Cognition* 22: 654-669.

Rey, Georges. 2013. We are not all 'self-blind': A defense of a modest introspectionism. *Mind & Language,* Vol. 28: 259-285.

Scaife, Robin. 2014. A Problem for Self-Knowledge: The Implications of Taking Confabulation Seriously. *Acta Analytica*, Vol. 29: 469-485.

Schwitzgebel, Eric. 2010. Acting contrary to our professed beliefs or the gulf between occurrent judgment and dispositional belief. *Pacific Philosophical Quarterly*. Vol. 91, No. 4: 531-553

Siewert, Charles. 1998. *The Significance of Consciousness*. Princeton: Princeton University Press.

Sherman, J. W., Gawronski, B. & Trope, Y. (eds). 2014. *Dual-Process Theories of the Social Mind*. New York: Guilford Press.

Taylor, Charles. 1985. Self-interpreting animals. In his *Human agency and language*. Cambridge: Cambridge University Press: 45-76.

Wilson, Timothy D. 2002. *Strangers to Ourselves. Discovering the Adaptive Unconscious.* Cambridge, Massachusetts: The Belknap Press of Harvard University Press.

Author Biography

Eivind Balsvik is a Senior Lecturer at the Faculty of Social Sciences, University of Oslo, Norway. His principal research interests are related to questions concerning rationality, interpretation theory and research ethics. He has also worked on the philosophy of Donald Davidson and theories of self-knowledge.