Verbal Reports and 'Real' Reasons: Confabulation and Conflation

Constantine Sandis

Pre-proofs of paper to appear in Ethical Theory and Moral Practice in 2015

Abstract

This paper examines the relation between the various forces which underlie human action and verbal reports about our reasons for acting as we did. I maintain that much of the psychological literature on confabulations rests on a dangerous conflation of the reasons for which people act with a variety of distinct motivational factors. In particular, I argue that subjects frequently give correct answers to questions about the considerations they acted upon while remaining largely unaware of why they take themselves to have such reasons to act. *Pari passu*, experimental psychologists are wrong to maintain that they have shown our everyday reason talk to be systematically confused. This is significant because our everyday reason-ascriptions affect characterizations of action (in terms of intention, knowledge, foresight, etc.) that are morally and legally relevant. I conclude, more positively, that far from rendering empirical research on confabulations invalid, my account helps to reveal its true insights into human nature.

Key words: reasons, confabulation, verbal reports, action explanation, motivation, experimental psychology.

1. Agential Reasons

In deliberating about what to do we frequently weigh up various considerations for and against a certain course of action. I shall here call any consideration *upon which* one actually acts or refrains from doing so an *agential reason*. So defined, agential reasons are not psychological phenomena such as beliefs or desires but, rather, purported facts about the world: things *that* we believe.[1] I shall not concern myself here with whether or not so-called motivating reasons should be identified with agential reasons (as opposed to, say, mental states). All that matters, for my present purposes, is that the very notion of a consideration one acts upon is a *bona fide* one. This is not to deny that we sometimes confabulate agential reasons when there are none, or that such reasons might not be fictional insofar as it remains meaningful to speak of a person acting upon the consideration that *p* when her belief that *p* is false.

Various features of our psychology, including repressed wishes, desires, beliefs etc. may motivate our actions, but it would be a rookie mistake to identify them with agential reasons. In addition, facts about ourselves which we need not be aware of can explain why we mistakenly took something to be the case, or to favour a particular course of action. Agential reasons may be said to *nest* within such facts.[2] Once we have given an explanation in terms of a person's agential reasons, there will always be, as Jonathan Dancy puts it, 'a possible further question how it was that the agent was the sort of person to be influenced by such features' (Dancy 2000: 173).

---

[1] For competing ways of understanding this commonplace distinction between what one believes and one's believing it see White (1972), Hornsby (1997), and Dancy (2000:121ff.). My argument doesn't hang on the particular details of any such approach.

[2] I'm thinking of explanatory nesting here; normative nesting would require appeal to further agential reasons.

To insist that the facts within which agential reasons are nested are reasons *for which* we acted is to hold what I shall call the *Conflating View of Nesting Reasons* (CVNR):

A reason why A took *x* to be a reason to ϕ is a reason for which A ϕ-d (if she did so).

The conflation is dangerous because it leads to the view that uncovering a nesting reason reveals the 'real' reason for an action, one that is in direct competition with the agent's own verbal reports about her reasons (cf. Stoutland 1998). To be sure, reason-nests serve to render actions intelligible, but they do not do so in virtue of being agential reasons.

Basic facts about human nature make possible all sorts of nesting relations. For example, it is common for incentive factors to awaken drive states, as when the aromatic fumes of the bakery stimulate one's hunger (see Cabanac 1979). Smelling the aroma may thus motivate me to enter the shop and buy a loaf of bread, but it is not my agential reason for doing *s*, regardless of whether or not I am conscious of the mechanisms at play. Likewise, I may be motivated to buy a bottle of a certain brand of rum by the advert portraying beautiful, happy, people on a yacht in some exotic location. My agential reason here will not be the thought that I could in some way be like them but, rather, the purported fact that the rum will be of good quality.

2. Motivational Forces

An explanation of why A took *x* to be a reason for performing a certain action may refer to nesting factors as diverse as her upbringing, a past trauma, or a fact about human nature. These may be further nested within neuro-scientific facts that 'explain why a person is more prone than normal to inhabit certain mental states – e.g. depression' which make them 'more liable to act for a certain kind of reason than someone with a different neuroscientific make-up' (Bennett and Hacker 2003: 29).

Consider Professor Moody, whose essay grading is easily affected by his emotional moods but who is harbouring under the delusion that his grading is always done purely on grounds of actual merit: a certain essay is seen as being stylish and well-researched, another as lacking structure, etc. Moody thus takes himself to be guided by *what* he believes the merit of any given essay is and has the second-order belief that his beliefs are reliably based on stable criteria which he is sufficiently sensitive to. His moods cause him to think that the considerations of merit which he takes himself to be guided by dictate that it is appropriate for him to grade the papers as he does. Moody is clearly self-deceived, but it would be wrong to infer from this that his moods are the 'real reasons' for his actions. Rather, the moods explain why Moody took certain (seeming or actual) considerations to be reasons for grading the essays as he did e.g. why he perceived some reference to Julio Iglesias as being pedestrian, when in some different mood he might have found it imaginative. His agential reasons, by contrast, are not the moods themselves but the considerations which – given his moods – he came to focus on and be moved by. Accordingly, we should distinguish between:

(1) Moods which explain:

(a) Why Moody had certain beliefs about the essays;
(b) Why he took the things he believed to be good reasons for giving a particular set of grades to the essays;
(c) Why he weighed his reasons as he did.

(2) Moody's agential reasons for grading the essays as he did.

Parallel distinctions ought to be made in explanatory narratives featuring the sorts of *situational* attributions described by Milgram (1963), Isen and Levin (1972 & 1975), Gazzaniga, and LeDoux 1978), Doris (1998 & 2002), Wilson (2002), and Ross & Nisbett (2011), or indeed a mix of psychological and situational attributions.[3] Such explanations work at a sub-rational level. By this I don't mean that personal or situational factors do not operate rationally (though this will frequently be the case, to varying degrees), but only that related explanations of what moves people to act as they do are not to be given purely in terms of agential reason. It is at best misleading to say that one's finding a dime in a phone booth is one's 'real reason' for helping a passer-by with directions. Personal and situational attributions explain *why* we take ourselves to have certain reasons for acting.

3. Verbal Reports and Purported Confabulations

Timothy D. Wilson has claimed that 'our explanations are confabulations' and that 'people's reasons about their own responses are as much conjectures as their reasons for other people's responses' (Wilson 2002: 113). His main argument for this stems from earlier experiments he conducted with Nisbett in the 1970s:

> ... passersby were invited to evaluate items of clothing – four different nightgowns in one study (378 subjects) and four identical pairs of nylon stockings in the other (52 subjects). Subjects were asked to say which article of clothing was the best quality and, when they announced a choice, were asked why they had chosen the article they had. There was a pronounced left-to-right effect, such that the rightmost object in the array was heavily chosen. For the stockings, the effect was quite large, with the right-most stockings being preferred over the left-most by a factor of almost four to one. When asked about the reasons for their choices, no subject ever mentioned spontaneously the position of the article in the array. And, when asked directly about a possible effect of the position of the article, 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).

Nisbett and Wilson had undertaken to show that we are largely ignorant of the 'cognitive processes underlying our choices, evaluations, judgments, and behavior' (231) and that 'one has no more certain knowledge of the working of one's own mind than would an outsider with intimate knowledge of one's history and of the stimuli present at the time the cognitive process occurred' (257). They argued further that the 'real reason' for this ignorance is not that we have 'no direct access to higher order mental processes' (232) but that our mental reports are 'based on a priori, implicit causal theories, or judgements about the extent to which a particular stimulus is a plausible cause of a given response' (231). From this concluded:

> ... subjects may have been making simple representativeness judgments when asked to introspect about their cognitive processes. Worry and concern seem to be representative, plausible reasons for insomnia while thoughts about the physiological effects of pills do not. Seeing weight tied to a string seems representative of the reasons for solving a problem that requires tying a weight to a cord, while simply seeing the cord put into

[3] One peculiarity about the situationist attack on virtue ethics (e.g. Doris 1998) is that virtue ethicists can agree that one may act for reasons that are situational features. So the debate is not really about whether we are 'motivated' by external factors or inner dispositions but, rather, whether external factors which we are not conscious of influence our behaviour *more than* 'internal' character traits. Yet the only way one can test for character traits in the first place is through behaviour.

motion does not. The plight of a victim and one's own ability to help him seem representative of reasons for intervening, while the sheer number of other people present does not. The familiarity of a detergent and one's experience with it seems representative of reasons for its coming to mind in a free association task, while word pairs memorized in a verbal learning experiment does not. The knit, sheerness, and weave of nylon stockings seem representative of reasons for liking them, while their position on a table does not (249).

A worry with this analysis is its lack of any distinction between the causes of bodily behaviour and agential reasons for action. Ironically, their argument unintentionally suggests that laypeople might be making just such a distinction. If so they would be right to do so: the position of a pair of stockings on a table is rarely, if ever, a reason for which one chooses them over another pair. It could, however, explain why we mistakenly come to think of them as being smoother etc. What we are fabricating in such a case is not a tale about our agential reasons but one about the quality of the stockings. The subjects are quite right to intuitively distinguish between their agential reasons and the cognitive and conative nests that underlie them. Had they been explicitly asked about the latter rather than about about 'reasons for their choices' they may have given significantly different answers.[4] It is not the subjects who are misled by theory, but the researchers analysing the data. The empirical evidence suggests that it is not laypeople who 'cannot correctly identify the stimuli that produced their response' (ibid: 233), but those theorists who cannot correctly specify what they are asking their subjects to explain.

In a lesser-known paper, Wilson and Nisbett (1978:124-5) speculated that the subjects in the stocking experiment did not have a right-to-left spatial position bias, but a temporal bias which led them to prefer the items they saw last (a bias which would favour late cinema releases at the Oscars). In this paper, they constantly switch between talk of 'reasons for their judgments and choices' (118) and talk of 'what factors influenced their responses' (ibid). For example, they write that '[n]ot a single subject mentioned the position of the stocking as a reason for their choice', furthering interpreting such responses as 'causal analyses' (124).[5] Yet 'to ask a person why he performed a particular behaviour or chose a particular course of action (126) is at best ambiguous between a request for an agential reason and a causal/motivational factor. In addition, Wilson and Nisbett systematically conflate 'reported reasons for' (122) with 'reports of influence', frequently switching between talk of 'reasons for', 'reasons why', 'influence', 'factors', 'explanations', and 'causal analysis'.

Nisbett and Wilson's theories work in the sense that they have been applied (e.g. in super-market product placement) with impressive results. But they do not work for the reasons which they give us. Writing with another collaborator, Wilson later concluded that '...when people are called upon to interpret the events that unfold around them, they tend to overlook or to make insufficient allowance for situational inferences' (Ross and Wilson 1991/2011: 90). Just so, but we must ask ourselves what exactly are subjects instructed to do before they give their verbal reports? Are they asked about motivational factors or agential reasons? As Ericsson & Simon (1980) have argued, it is crucial to explicate the mechanisms by which verbal report data is generated and the instructions it could be sensitive to.

In everyday parlance, the term 'reason for' is sometimes used to refer to the reasons why *x* happened and that it is consequently important to distinguish between a reason/the reason for A's doing *x* and A's reason for doing *x* (See Haidt 2001 & 2012 and Wilson 2002). In light of this, consider the following pairs of propositions:

(i) The reason for her choosing product A was the colour of the packaging.
(ii) Her reason for choosing product A was the colour of the packaging.

[4] Regarding the subjects' alleged denial of 'a possible effect of the position of the article' one would like to know more about the exact question they were asked: an effect on *what* exactly?

[5] Caruthers (2011:329ff.) rightly criticizes Nisbett & Wilson (1977) and Wegner (2002) on placing too much emphasis on *causal* attributions, as opposed to confabulations about the very existence of 'some suitable mental event to serve as the cause'. But we must tread very carefully if we are to maintain, as Carruthers (ibid.: 336) does, that these can include judgments (see n. 9 below).

(a) The reason for the cow's death was BSE.
(b) The cow's reason for dying was BSE.

In assuming that an agent's 'real reason(s)' for choosing *x* is identical to the reason(s) why she chose it, Nisbett and Wilson conflate (i) and (ii). Notice, however, that the difference between (i) and (ii) is identical to that between (a) and (b). Yet under ordinary circumstances (b) makes no sense. Similarly, (ii) makes little sense in the sorts of situations described in the experiments above, though one may well consciously choose to buy a product because they like the packaging. It is as fallacious, then, to infer (ii) from (i) as it is to infer (b) from (a).

Unlike propositions of the form shared by (i) and (a), propositions such as (ii) and (b) are about agential reasons. A person may hide their real agential reason from others, and we may even come to deceive ourselves about the real reasons behind our past actions not long after we have performed them[6] but, at the moment of action, it is our motives, needs, drives, and associations that we are typically ignorant of, not the considerations we act upon. This is not to suggest that we are infallible about our agential reasons at the time of action, only that there is no empirical evidence to suggest that we are systematically deceived about them.[7]

The conflation of motivational facts with agential reasons amounts to what I shall call *The Conflating View of Motivating Reasons* (CVMR):

The reasons *for which* we act are identical to the things which motivate us to act (and vice versa).

The term 'real reason' is often used mischievously by researchers who claim that the discovery of nesting and/or motivating reasons trumps any beliefs we have about our agential reasons, as if the relation between them were that of actual truth vs. mere appearance. The term thus actually serves to indicate a discrete shift of discourse, from issues pertaining to practical reasoning to questions of background psychology.

The experiments of Nisbett & Wilson were anticipated by air dealers such as the 1950s marketing and propaganda 'expert' Vance Packard. In his influential book *The Hidden Persuaders* (which claims to reveal 'what makes us buy, believe – and even vote – the way we do'), Packard writes:

Psychologists at the McCann-Erickson advertising agency asked a sampling of people why they didn't buy one client's product – kippered herring. The main reason people gave under the direct questioning was that they just didn't like the taste of kippers. More persistent probing however uncovered the fact that forty percent of the people who said they didn't like the taste of kippers had never in their lives tasted kippers!

The Color Research Institute ... was testing to see if a woman is influenced more than she realizes, in her opinion of a product, by the package. It gave housewives three different boxes filled with detergent and requested that they try them all out for a few weeks and then report which was the best for delicate clothing. The wives were given the impression that they had been given three different types of detergent. Actually, only the boxes were different; the detergents inside were identical ... In their reports the housewives stated that the detergent in the brilliant yellow box was too strong; it even allegedly ruined the clothes in some cases. As for the detergent in the predominantly blue box, the wives complained in many cases that it left their clothes dirty looking. The third box, which contained what the institute felt was an ideal balance of colors in the package design, overwhelmingly received favorable responses. The women used such words as 'fine' and 'wonderful' in describing the effect the detergent in that box had on their clothes (Packard 1957/2007: 39–40).

[6] For confabulations relating to motivated misremembering see Henkel & Mather (2007).
[7] In section 5 I maintain that any deception cannot be due to straightforward lack of transitive consciousness.

The most interesting thing such experiments reveal is just how deeply our agential reasons may be nested within psychological facts that we have absolutely no awareness of. So while we are indeed strangers to ourselves, as Wilson (2002) puts it in his book title, we've been given no reason to suppose that this is so when it comes to our agential reasons.

In the first example, people who have never tasted kippers claimed that their reason for not purchasing them is that they did like their taste. Packard assumes that since they do not know – indeed it could even be false – that they don't like the taste of kippers, the purported fact that they do not like their taste cannot be their reason for not purchasing them. But this reasoning is fallacious. For while it might well be true that falsehoods cannot explain why people actually act as they do, it does not follow that a person can only act for the agential reason that *p* if they know (or at the very least if it is at least the case) that *p*.

The second example may be dealt with in a similar fashion. Subjects were indeed wrong to think that each of the detergents they tried had a distinct set of effects, and Packard offers a very plausible inference to the best explanation of why they perceived the effects as being distinct when they clearly were not so. But while it is both counterfactually true and psychologically interesting that people often choose to purchase detergents because of their packaging and nothing else, this fact is not in competition with their verbal reports concerning their agential reasons. Experimental manipulations may indeed lead subjects to act on reasons that they would have lacked without the manipulation[8], but the best explanation of why they would, in such cases, choose the fancier box is that they think it contains a better detergent, this supposed fact being the consideration they act upon viz. their agential reason.[9]

Similar conflations may be found in the earlier work of Freud's nephew, Edward Bernays:

> A man buying a car may think he wants it for the purposes of locomotion, whereas the fact may be that he would really prefer not to be burdened with it, and would rather walk for the sake of his health. He may really want it because it is a symbol of social position, an evidence of his success in business, or a means of pleasing his wife. This general principle, that men are very largely actuated by motives which they conceal from themselves, is as true of mass as of individual psychology. It is evident that the successful propagandist must understand the true motives and not be content to accept the reasons which men give for what they do ... what are the true reasons the purchaser is planning to spend his money on a new car instead of a new piano? Because he has decided that he wants the commodity called locomotion more than he wants the commodity called music? Not altogether. He buys a car, because it is at that moment the group custom to buy cars (Bernays 1928/2005: 75 & 77).

Bernays' juxtaposition of true motives with agential reasons suggests that he embraces CVMR. His further identification of the question 'why did he buy a car?' with that of 'why did he buy a car rather than a piano?' suggests a commitment to the more general *Conflating View of Reasons* (CVR):

> The reasons for which we act are reasons *why* our actions occur.

Bernays is right to agree with 'the psychologists of the school of Freud who have pointed out that many of man's thoughts and actions are compensatory substitutes for desires which he has been obliged to suppress' (ibid.: 75). The reasons in question, however, are not usually in competition with those offered by the layperson. Rather, they typically serve to explain why people have the agential reasons that they do. In what follows I argue that whether or not any explanatory factor is a real reason for action may be tested counterfactually, and that our everyday verbal reports about agential

[8] Many thanks to an anonymous referee for registering the significance of this.

[9] This is not a case of falsely attributing a judgement to oneself (Caruthers 2011: 336) but of making a false judgement.

reasons typically pass such tests. The exceptions are cases such as hypnosis where the action performed is not obviously intentional.[10]

4. Persuasion and Counterfactuals

It is tempting to think that the agential reasons of verbal reports are epiphenomenal, particularly if one adopts a counterfactual analysis of explanation and/or causation.[11] This temptation must be avoided. Consider the case of fancy packaging. This would not have the effect it does if it did not cause us to believe that the product in contains was good. Fancy packaging only 'causes' people to buy the product it contains in virtue of making them believe that the latter is of superior quality. But the reasons reported by laypeople would only be epiphenomenal in cases in which the quality of the packaging caused the subject to buy the product no matter *what* they took their reason to be. The empirical data, by contrast, confirms that features such as packaging and spatio-temporal location have the effects they do precisely because they alter our beliefs about our practical reasons.

This is not to say that there is no such thing as confabulation, only that we should not all-too-readily postulate the phenomenon without good counterfactual grounds. One such case is provided by Jonathan Haidt in relation to people whose disgust for incest and other taboo violations causes them to maintain related moral judgements after admitting that they cannot offer any good reasons for them:

> In these harmless-taboo scenarios, people generated far more reasons and discarded far more reasons than in any of the other scenarios. They seemed to be flailing around, throwing out reason after reason, and rarely changing their minds when Scott proved that their latest reason was not relevant.
>
> Subject: Um...well...oh, gosh. This is hard. I really - um, I mean there's no way I could change my mind but I just don't know how to - how to show what I'm feeling, what I feel about it. it's crazy![12]

The reasons cited here really are epiphenomenal, but this doesn't entail that the subjects' feelings and emotions are their 'real reasons'. Rather, what is demonstrated is that such people hold certain beliefs for no reason at all. We can provide an emotional explanation of why *they* hold their beliefs, but it won't involve considerations of the form 'incest is bad because it is disgusting'. The subjects have no grounds for their beliefs, only causes. Of additional relevance to the main point of my paper is the fact that groundless belief is different from groundless action in that no belief is held intentionally in the sense in which an action may be performed intentionally. The question of whether beliefs held for no reason are nonetheless intentional does not arise the way in which it does for groundless actions.[13]

5. Reasons and Consciousness

[10] Arguably, the reasons *for which* we act should never be understood as reasons *why* we act. This is because the considerations we act upon maybe false, whereas all genuine (as opposed to merely purported) explanation is by definition factive. When we explain action by citing agential reasons, the explanatory work is done by true propositions about the reasons the agent acted for, and not by the reasons themselves. I argue for this view in Sandis (2013).

[11] See Ruben (2003:ch.6) for a counterfactual theory of *explanation* and Martin (2011:ch.2) for an attack on counterfactual analyses of *causality*.

[12] Haidt (2012:39-40); cf. Aronson (1956) & Haidt 2001). Similar strategies by end-of-the-world cultists whose predictions were disproved are reported in Arosnon (1956).

[13] For more on the difference between belief-orientated and action-orientated confabulations see Hindricks (forthcoming). Hursthouse (1991) offers persuasive examples of intentional actions not performed *for* reasons (but out of habit, anger, etc.) which do not involve confabulation.

More recent empirical studies have yielded new insights into implicit associations that affect our daily actions in ways that we are unaware of, and which can only be controlled indirectly through calculated efforts over extended periods of time.[14] These in turn lead to a range of prejudices or biases that affect our daily interaction with people.[15] Many claims made about implicit bias, however, rest on the sorts of conflations outlined above. Take, for example, the following remarks by Greenwald & Krieger:

> Theories of implicit bias contrast with the 'naive' psychological conception of social behavior, which views human actors as being guided solely by their explicit beliefs and their conscious intentions to act. A belief is explicit if it is consciously endorsed. An intention to act is conscious if the actor is aware of taking an action for a particular reason ... In contrast, the science of implicit cognition suggests that actors do not always have conscious, intentional control over the processes of social perception, impression formation, and judgment that motivate their actions (Greenwald & Krieger 2006: 946).

In a footnote, they add that:

> 'Naive psychology' refers to laypersons' intuitions about determinants and consequences of human thought and behavior, especially their own. Modern treatments were largely inspired by Fritz Heider's book, *The Psychology of Interpersonal Relations* [1958], which initiated systematic investigation of how laypersons' intuitions differ from scientific understanding (ibid.: 967, n.3).

What is naive, however, is the assumption that our implicit biases guide our actions in a way that is in tension with the layperson's reports on her agential reasons. If this were true, implicit associations would be far easier to detect and control. To have an implicit bias for or against *x* is to be disposed to view its features in a more positive or negative light than you otherwise would, but it is the features as we see them that guide our actions *qua* agential reasons.

Consider Frank, who gives a job to Arno instead of Maureen because he believes that, while they are both equally skilled and experienced in all other respects, Arno is a faster typist. Frank's being perfectly aware that this consideration was his agential reason for giving Arno the job is compatible with his being unaware that (i) Maureen is an equally fast - if not faster -typist, and/or more qualified for the job in other respects, and (ii) that his false belief is driven by an implicit gender bias.[16] The claim that Frank gave the job to Arno rather than Maureen, not because he is the better typist but because of Frank's implicit gender bias, is true but potentially misleading. It is true because (a) it is false that Arno is a better typist than Maureen and (b) Frank really does have a gender bias which affected his behaviour. But it is misleading insofar as it implies that Frank was not motivated by the thought that Arno is a better typist, and did not act upon the relevant agential reason.[17] This

---

[14] See Project Implicit: https://implicit.harvard.edu/implicit/demo/selectatest.html. For an example of the use of such studies in popular media see: http://www.businessweek.com/articles/2012-12-13/hidden-bias-why-you-should-pitch-after-lunch (both accessed 27 September, 2014).

[15] See the research papers conducted in conjunction with Project Implicit: http://www.projectimplicit.net/articles.php (accessed 27 September, 201). For a sceptical interpretation of what the tests show see Egloff *et al* (2005). An indication of further debates has been captured by Tierney (2008).

[16] An alternative scenario is one in which Frank honestly avows that he believes Arno and Maureen are equally skilled in *every* respect and takes himself to 'randomly' offer the job to Arno when in actual fact an implicit bias (which can be measured statistically) is at play.

[17] Similar conflations surround Festinger's notion of cognitive dissonance (Festinger 1957:1) and Tamar Gendler's notion of alief (Gendler 2008*a* & *b*).

point is ethically significant, for the description under which Frank's act is intentional is not the description under which it is biased. From a legal and moral point of view, for instance, it is of vital importance that one does not conflate implicit association with unconscious reasons. Consider the question of whether discrimination based on implicit bias was performed for agential reasons relating to race. This could affect whether or not something counts as a hate crime.

Stronger examples of agent confabulation are cases of hypnosis or Korsakoff's syndrome (Sacks 1987; cf. Wilson 2002: 93ff.). These tellingly involve behaviour that is either obviously non-intentional or, at best, a borderline case (Schroeder 2010: 460). The experiments don't demonstrate that we fail to identify our agential reasons, only that we confabulate in cases where there is no reason whatsoever *for* our beliefs, preferences, or behaviour. The hypnotised person does what she does because she has been instructed to do so under hypnosis, but this explanation does not cite her agential reasons. Indeed, there are none, for the hypnotised person does not act upon considerations. To talk of 'real reasons' in such cases, then, is to once again conflate agential reasons for action with reasons why behaviour occurs.

A similar error may be found in the work of those who challenge our ordinary assumptions about free will. In his book *Who's In Charge*, for example, Michael Gazzaniga writes:

> I automatically jump back before I realize why. I did not make a conscious decision to jump, it happened without my conscious consent ...If you were to have asked me why I had jumped, I would have replied that I thought I'd seen a snake. That answer certainly makes sense, but the truth is that I jumped before I was conscious of the snake: I had seen it, but I didn't know I had seen it. My explanation is from post hoc information I have in my conscious system...I jumped way before (in the world of milliseconds) I was conscious of the snake. I did not make a conscious decision to jump and then consciously execute it. When I answered that question I was, in a sense, confabulating: giving a fictitious account of a past event, believing it to be true. The real reason I jumped was an automatic nonconscious reaction to the fear response set into play by the amygdala. The reason I would have confabulated is that our human brains are driven to infer causality. They are driven to explain events that make sense out of scattered facts. The facts that my conscious brain had to work with were that I saw a snake, and I jumped. It did not register that I jumped before I was consciously aware of the snake (Gazzaniga 2012:77; see also McGilchrist 2009:80-82).

As with the hypnosis and Korsakoff's syndrome cases, the situation described above is most plausibly one in which the subject doesn't act *for* a reason at all. He may be guilty of confabulation, but we cannot infer from this that he - let alone his brain - is mistaken about his agential reasons, for there may not have even been any. The vignette certainly doesn't license us to conclude, as Gazzaniga does, that '[w]hen we set out to explain our actions, they are all *post hoc* explanations using post hoc observations with no access to nonconscious processing' or that 'the actions and the feelings happen before we are consciously aware of them' or that 'listening to people's explanations of their actions is ...often a waste of time' (Ibid.: 77-8).

It is also worth noting Gazzaniga's rhetorical over-use of the word 'conscious' here, as if I can only be in charge of my actions if there are such things as *conscious* brains and systems which enable me to give *conscious* consent to my behaviour by making *conscious* decisions which I consciously execute. Gazzaniga seems to assume that all thought and action is either completely conscious or hidden from us. But while it is true that one has to be conscious in the intransitive sense in order to act for agential reasons, it doesn't follow that one needs to be conscious of any particular thing (i.e. have consciousness in the transitive sense), let alone of one's agential reasons. To be conscious of something is to have your attention gripped by it White (1964), yet we normally act knowingly without thinking so much about things.

The mistake of drawing lessons concerning the explanation of everyday actions from experiments relating to behaviour that is not performed for reasons is also present in Gazzaniga's earlier work on confabulations and commissurotomy (Gazzaniga 1998) and by other work on direct brain stimulation (e.g. Hirstein 2009: 177). These are all extensions of the fallacy of assuming there is

a common denominator between everyday cases and those involving error and illusion.[18]

A related worry is that of choice blindness, in which subjects are asked to make a preference choice between two pictures and subsequently fail to detect that the picture they were looking at did not match the one they had chosen mere seconds before, offering reasons for preferring the wrong one that appear to be epiphenomenal (Johansson et. al. 2008; Hall & Johansson 2009). Here the confabulation lies in attempting to justify - or at least explain - a choice that was never made, or a preference that was never had. Again, we must ask ourselves whether these experiments really demonstrate that we frequently fail to be conscious of our agential reasons, or whether we confabulate because we had no agential reason at all. If the latter, our failure to understand ourselves is not a failure of consciousness but of expectation. We simply presume, assume, or otherwise convince ourselves that we must have had an agential reason, and so confabulate it. It is highly plausible that we do so in order to make ourselves feel better.[19] But this *telos* is not an agential reason, for we do not (unconsciously or otherwise) act upon the *consideration* that this will make us feel better. The drive to feel better acts as a (counterfactually robust) motive, not an agential reason.

It is worth stressing, at this point, that agential reasons are not the sorts of things that one may or may not be conscious of. This is because the things we consider are not to be found in our minds, be it via introspection or any other means. While we are often motivated by attitudes that we are unaware of, to the extent that it makes sense to talk of their propositional or representational content as agential reasons we acted upon - as opposed to considerations we may have failed to register - the question of consciousness simply doesn't arise. *Pari passu*, it is a mistake to talk here of an introspection illusion as Emily Pronin (2009) and Peter Carruthers (2011: 326-33). Agential reasons - reasons *qua* considerations - are not mental states, though we may act upon considerations that are *about* our mental states (Dancy 2000). Consequently, when we are mistaken about our reasons this is not a case of failed introspection. *Pace* Nisbett and Wilson, then, mistakes about agential reasons cannot be introspective ones.

The faulty guesses that people make in trying to explain their thought processes have been termed 'causal theories' by some researchers. But when we talk of the considerations we acted upon we are not providing any kind of theory, let alone a causal one. A person who is not conscious of some taboo behavioural disposition of his may well be motivated to discriminate against others who manifest it because he is too ashamed to admit to himself that he shares the disposition, but this does not mean that his avowals are not things that he believes and acts upon. The mistake of thinking otherwise is even shared by critics of Wilson et. al such as Nielsen & Kaszniak (2007) who argue that people's causal explanations are not merely a matter of *a priori* theory but also frequently involve of privileged information access.

Nisbett and Wilson originally distinguished between mental *contents* (thought to include feelings) and mental *processes*, arguing that while introspection gives us access to the former, the latter remain hidden. From this they concluded not only that 'people's lack of insight into the causes of their everyday responses' but also that they consequently ' do not know reasons for their feelings, judgments, and actions' (Wilson 2002: 103-4). This was thought to be so on the grounds that:

> people have priviledged access to a great deal of information about themselves, such as the content of their current thoughts and memories and the object of their attention...But these are mental *contents*, not mental *processes*', the former being thought of as 'results' of the latter. The view entails that while we may know what we think and do we don't know *why* this is so if we are ignorant of the causal processes that produced the 'contents' in question (ibid.: 105).

[18] For a critical assessment of the fallacy across various debates relating to perception, belief, and action see Dancy (1995).

[19] I owe this point to an anonymous referee for the journal.

I have tried to show that is the very idea of agential reasons as 'contents' that misleads us into thinking that they are things which we might come to know via introspection.[20] In later work, Wilson allows (for reasons unrelated to the main thrust of this paper) that 'the distinction between mental content and process is not very tenable' (ibid). This leads him to replace it with one between the 'adaptive unconscious' and the 'conscious self', concluding that 'people do not have conscious access to the adaptive unconscious, their conscious selves confabulate reasons for why they responded the way they did' (ibid). I have already noted some worries about overemphasizing consciousness. To this we might now add Richard Moran's concerns about first person authority being based upon epistemic evidence (Moran 2001:91; cf. Tanney 2002:314-16). We don't discover what we think via self-directed mindreading. We don't know what our agential reasons are because we have better access to them (be it introspective or otherwise) but because it is our business to make up our minds about what to think and do and to do so is to weigh up considerations and settle.

Carruthers (2011:325ff.) argues that 'our access to our own attitudes in general is interpretative rather than transparent'. But agential reasons are not attitudes and it makes no sense to talk of accessing them. Even if we assume (in my opinion, faslely) that beliefs and desires are mental states[21] and that Moran is wrong to say that we do not access these via introspection, the point about agential reasons still stands: they are not bits of information which we may succeed or fail to be conscious of. [22]To realise that oneself or another has a reason to do *x* is not to make a mental discovery but to relate normative facts to some situation. A view which lies somewhere between Carruthers' and my own is that of Neil Levy. Like myself, Levy (2011: sec V) argues against those who think we are systematically mistaken about the reasons *on* which we act, but he does so by appeal to a consciousness and awareness of our reasons, which he further maintains is important for moral responsibility (Levy 2013 and 2014*a*&*b*). A related difference between Levy's argument and my own is that he conceives of what I have been calling agential reasons as 'mental contents', whereas I am resistant to this metaphor. Indeed, it is this difference that enables him (and prevents me) from talking of consciousness and awareness.[23] So while we hold similar positions, we do so on radically different grounds.

7. Conclusion

There is much of great value that psychology and cognitive science can teach us about motivation. What these theories cannot do, however, is reveal our 'real' agential reasons. These rarely need uncovering, though we may come to misremember, or altogether forget them. Applications of a flawed theories may - up to some point - nonetheless have the very results they predict, in deviant (but non-accidental) ways. [24]

The various conflations I have sought to outline cause researchers to misunderstand the way their own theories function, rendering psychological theories void of certain psychological insights. In *misinterpreting their own discoveries, they pass over their true value. In light of all this, it is perhaps*

[20] For this reason I cannot endorse the letter of Peter Goldie's claim that there are 'influencing factors' on a person's mind that 'are not themselves part of the content of his mind', though I wholeheartedly agree with his contention that 'he proposed marriage to her at the party because he was drunk' is an explanation which offers 'a cause but...not a reason for...not a consideration' (Goldie 2012:20).

[21] But see Collins (1987).

[22] A former consideration may nonetheless leave motivational traces which influence us unconsciously. I owe this point to István Zárdai.

[23] For a counterfactual attack on Levy's view - and by extension by own - see Vierkant (forthcoming). But Vierkant's challenges can only be sustained if one refuses to distinguish between agential reasons and other motivational factors

[24] Kuhn (1962). This much is unquestionably true of Bernays (1928/2005), Packard (1957/2007), and Thaler and Sunstein (2008/9). For the relation of rhetoric to motives see Burke (1950).

unsurprising then, that researchers like Gazzaniga proclaim that 'psychology itself is dead' and that 'everyone but its practitioners knows about the death of psychology' (Gazzaniga 1998:xi-xii). But psychology hasn't grown old, or useless, or reached the end of its lifespan. If it is dead it is because it has been murdered - or at the very least assisted in its suicide - by experimental science. This is a great loss, because while psychology without experimental data is empty, experimental data without psychology is blind.

Acknowledgments

This paper was written with funding from the Spanish Ministry of Science and Innovation Consolidor-Ingenio Schemes *Philosophy of Perspectival Thoughts and Facts* (CSD2009-00056). It was presented at the *European Society for Philosophy & Psychology Conference*, Institute of Philosophy, Senate House, London, (August 28, 2012), the *International Symposium on Alternatives, Belief, and Action*, University of Valencia (15-16 Nov, 2012), The *XXIII World Congress of Philosophy* in Athens (4-10 Aug, 2013), & the *Everyday Reasons Workshop* at the Erasmus University of Rotterdam (14-15 Nov, 2013). Many thanks to all the organisers and participants , especially Maria Alvarez, Santago Amaya, Jan Bransen, Lilian O' Brien, Giuseppina D'Oro, Frank Hindricks, Neil Levy, Guido Loehrer, Christopher Lumer, Michael McKenna, Carlos Moyas, Carlos G. Patarrayo, Sergi Rosell, Katrien Schaubroeck, Scott Sehon, George Sher, Maureen Sie, Helen Steward, Karsten Stueber, Susanne Uusitalo, Tillman Vierkant, and Arno Wouters. Particular thanks is due to David-Hillel Ruben and Tobies Grimaltos for their nuanced responses to my paper at the London and Valencia conferences, respectively, and to István Zárdai for his comments on an earlier draft. Finally, I'm extremely grateful to Eric Wiland and an anonymous referee for extremely helpful reviewer reports.

References

Aronson, E. (1956), *When Prophecy Fails* (New York, NY: Harper).

Bennett, M. R. and Hacker, P. M. S. (2003), *Philosophical Foundations of Neuroscience* (Oxford: Blackwell).

Burke, K. (1950), *The Rhetoric of Motives* (New York, NY: Prentice-Hall).

Cabanac, M. (1979), 'Sensory Pleasure', *Quarterly Review of Biology*, Vol. 54, 1–29.

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

Collins, A. W. (1987), *The Nature of Mental Things* (Notre Dame Ind.: University of Notre Dame Press).

Davidson, D. (1963), 'Actions, Reasons, and Causes', *Journal of Philosophy*, Vol. 60: 685–700. Reprinted in Davidson (2001a: 3–19), to which any page numbers given refer.
—— (1976), 'Hempel on Explaining Action', *Erkenntnis* Vol. 10, 239–53. Reprinted in Davidson (2001: 261–75), to which any page numbers given refer.
—— (1982), 'Paradoxes of Irrationality', in (eds) R. Wollheim and J. Hopkins, *Philosophical Essays on Freud* (Cambridge: Cambridge University Press), 289–305. Reprinted in Davidson (2004: 169–87), to which any page numbers given refer.
—— (2001), *Essays on Actions and Events*, 2nd revised edition (Oxford: Clarendon Press).
—— (2004), *Problems of Rationality* (Oxford: Clarendon Press).

Dancy, J. (1995), 'Arguments from Illusion', *The Philosophical Quarterly*, Vol. 45, No. 181 (Oct), pp. 421-438.

—— (2000), *Practical Reality* (Oxford: Oxford University Press).

Doris, J. M. (1998), 'Persons, Situations, and Virtue Ethics', *Noûs*, 32, 504–30.
—— (2002), *Lack of Character: Personality and Moral Behavior* (Cambridge: Cambridge University Press).

Egloff, B., Schwerdtfeger, A., Schmukle, S. C. (2005), 'Temporal Stability of the Implicit Association Test-Anxiety', *Journal of Personality Assessment*, Vol. 84, No. 1, 82–8.

Ericsson, K. A. & Simon, H. A. (1980), 'Verbal Reports As Data', *Psychological Review*, Vol 87(3), May, 215-251.

Festinger, L. (1957), *A Theory of Cognitive Dissonance* (Stanford CA: Stanford University Press).

Festinger, L. & Carlsmith, J. (1959), 'Cognitive Consequences of Forced Compliance', *Journal of Abnormal and Social Psychology*,58, 203-210.

Gazzaniga, M.S. (1998), *The Mind's Past* (Berkeley CA: University of California Press).
_____ (2012), *Who's In Charge* (London: Constable & Robinson).

Gazzaniga, M.S, and LeDoux, J.E. (1978), 'Awareness of influences on one's own judgments: The roles of co-variation detection and attention to the judgment process', *Journal of Personality and Social Psychology*, 52, 453-63.

Gendler, T.S. (2008*a*), 'Alief and Belief', *Journal of Philosophy*, Vol. 105, No. 10, 634–63.
—— (2008*b*), 'Alief in Action (and Reaction)', *Mind & Language,* Vol. 23, No. 5, 552–85.

Goffman, E. (1990), *The Presentation of Self in Everyday Life*, new ed. (London: Penguin).

Goldie, P. (2012), *The Mess Inside: Narrative, Emotion, & the Mind* (Oxford: Oxford University Press).

Greenwald, A.G. & Krieger, L.H., 'Implicit Bias: Scientific Foundations', *California Law Review*, Vol. 94, No. 4 2006, 945-6).

Haidt, J. (2001), 'The Emotional Dog and its Rational Tail: A Social Intuitionist Approach to Moral Judgement', *Philosophical Review*, Vol. 108, No. 4, 814–34.
—— (2012), *The Righteous Mind: Why Good People Are Divided by Politics and Religion* (London: Allen Lane).

Hall, L, & Johansson, P. (2009), 'Choice Blindness: You Don't Know What You Want', *New Scientist*, Issue 2704 (18 April).

Henkel, L.A.& M. Mather (2007), 'Memory attributions for choices: How beliefs shape our memories', *Journal of Memory and Language* 57 (2): 163–176.

Hindricks, F. (forthcoming), '(How) Does Reasoning Fail to Contribute to Moral Judgment? Rationalizations,Dumbfounding, and Disengagement'.

Hirstein, B. (2009), 'Confabulation' in (eds.) Bayne, T. Cleeremans,A., & Wilken, P., *Oxford Companion to Consciousness* (Oxford: Oxford University Press).

Hornsby, J. (1997), 'Thinkables', in (ed.) M. Sainsbury, *Thought and Ontology* (Milan: Franco Angeli).

Hursthouse, R. (1991), 'Arational Actions', *The Journal of Philosophy*, Vol. 88, No. 2 (Feb), pp. 57-68.

Johansson, P., Hall, L., & Sikstrom, S. (2008), 'From Change Blindness to Choice Blindness' *Psychologia* 51 (2): 142–155.

Levin, P. F. and Isen, A. M. (1972), 'Effects of Feeling Good and Helping: Cookies and Kindness', *Journal of Personality and Social Psychology*, 3, 384–8.

Levy, N. (2011), 'Expressing Who We Are: Moral Responsibility and our Awareness of Our Reasons for Action', *Analytic Philosophy*, Vol 52, Iss 4 (Dec), 243-261.
_____ (2013), 'The Importance of Awareness', *Australasian Journal of Philosophy*, Vol. 91, Iss 2, 211-229.
_____ (2014*a*), 'Consciousness, Implicit Attitudes and Moral Responsibility', *Noûs*, Vol 48, Iss 1 (March), 21-40.
_____ (2014*b*), *Consciousness and Moral Responsibility* (Oxford: Oxford University Press).

Martin, J.L. (2011), *The Explanation of Social Action* (Oxford: Oxford University Press).

McGilchrist, I. (2009), *The Master and His Emissary: The Divided Brain and the Making of the Western World* (New Haven, CT: Yale University Press).

Milgram, S. (1963), 'Behavioral Study of Obedience', *Journal of Abnormal and Social Psychology*, Vol. 67, No. 4, 371–8.

Moran, R. (2001), *Authority and Estrangement: An Essay on Self-Knowledge* (New Jersey: Princeton University Press).

Nielsen, L., & Kaszniak, A.W. (2007), 'Conceptual, theoretical, and methodological issues in inferring subjective emotional experience: Recommendations for researchers', in (eds.) J.J.B. Allen & J. Coan, *The Handbook of Emotion Elicitation and Assessment* (New York: Oxford University Press), 361-375.

Nisbett, R. E. and Wilson, T. D. (1977), 'Telling More Than We Can Know: Verbal Reports on Mental Processes', *Psychological Review*, Vol. 84, No. 3, 231–259.

Ogilvy, D. (1963),*Confessions of an Advertising Man* (New York: Atheneum).

Packard, V. (1957/2007), *The Hidden Persuaders*, new (50th anniversary) edition (New York: IG Publishing).

Peters, R.S. (1958), *The Concept of Motivation* (London: Routledge and Kegan Paul).

Pronin, E. (2009), 'The Introspection Illusion', in (ed.) M. P. Zanna, *Advances in Experimental Social Psychology*, Vol. 41 (Burlington: Academic Press), 1-67.
Reeves, R. (1962), *Reality in Advertising* (New York: Alfred A. Knopf).

Ross, L. & Nisbett, R.E. (2011), *The Person and the Situation*, rev. 2nd edition (New York: Pinter & Martin).

Ruben, D-H. (2003), *Action and its Explanation* (Oxford: Clarendon Press).

Sacks, O. (1987), *The Man Who Mistook His Wife For a Hat, and Other Clinical Tales* (New York: Harper and Row).

Sandis ,C. (2012), *The Things We Do And Why We Do Them* (Basingstoke: Palgrave Macmillan).
_____ (2013), 'Can Action Explanations Ever be Non-Factive?' in (eds. B. Hooker, M. Little, & D. Backhurst), *Thinking About Reasons*: *Themes From the Philosophy of Jonathan Dancy* (Oxford University Press), 29-49.

Schroeder, S. (2010), 'Wittgenstein' in O'Conner and Sandis, *A Companion to the Philosophy of Action* (Oxford: Wiley-Blackwell), 554–61.

Stoutland, F. (1998), 'The Real Reasons' in J. Bransen and S. E. Cuypers (eds), *Human Action, Deliberation and Causation* (Dordrecht: Kluwer Academic Publishers), 43–66.

Sutton, R. & Douglas, K. (2013), *Social Psychology* (Basingstoke: Palgrave).

Tanney, J. (2002), 'Self-Knowledge, Normativity, and Construction', in (ed. A. O' Hear) *Logic, Thought, and Language*, Royal Institute of Philosophy Supplement 51 (Cambridge: Cambridge University Press), 37-55; reprinted in her *Rules, Reason and Self-Knowledge* (Cambridge MA: Harvard University Press, 300-321) to which any page numbers refer.

Thaler, R. H. and Sunstein, C. R. (2008/9), *Nudge: Improving Decisions About Health, Wealth and Happiness*, revised edition (London: Penguin).

Tierney, J. (2008), 'In Bias Test, Shades of Gray', *New York Times*, 17 November.

Vierkant, T. (forthcoming), ' Explicit reasons, implicit stereotypes and the direct control of the mind'.

White, A.R. (1964), *Attention* (Oxford: Basil Blackwell).

Wilson, T.D. (2002), *Strangers to Ourselves: Discovering the Adaptive Unconscious* (Cambridge MA: Belknap Press).

Wilson, T.D. & Nisbett, R.E. (1978), 'The Accuracy of Verbal Reports About the Effects of Stimuli on Evaluations and Behavior', *Social Psychology*, Vol.41, No.2, 118-131.

u