Self to Self

Selected Essays

J. DAVID VELLEMAN
New York University

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore,
São Paulo, Delhi, Dubai, Tokyo

Cambridge University Press
32 Avenue of the Americas, New York, NY 10013-2473, USA

www.cambridge.org
Information on this title: www.cambridge.org/9780521670241

First published 2006
Reprinted 2007

A catalog record for this publication is available from the British Library

Library of Congress Cataloging in Publication data

Velleman, James David.
Self to Self : selected essays / J. David Velleman.
p. cm.
Includes bibliographical references and index.
ISBN 0-521-85429-6 (hardcover) – ISBN 0-521-67024-1 (pbk.)
1. Self. 2. Self (Philosophy) 3. Kant, Immanuel, 1724–1804 – Ethics. I. Title.
BF697.V45 2005
126–dc22 2005008114

ISBN 978-0-521-85429-0 Hardback
ISBN 978-0-521-67024-1 Paperback

Transferred to digital printing 2009

10

From Self Psychology to Moral Philosophy

Prescott Lecky's *Self-Consistency* was published in 1945, four years after the author's death, at the age of 48.[1] Subtitled *A Theory of Personality*, the book defended a simple but startling thesis:[2]

> We propose to apprehend all psychological phenomena as illustrations of the single principle of unity or self-consistency. We conceive of the personality as an organization of values which are felt to be consistent with one another. Behavior expresses the effort to maintain the integrity and unity of the organization.

Lecky regarded self-consistency as the object of a cognitive or epistemic motive from which all other motives are derived.[3] "The subject must feel that he lives in a stable and intelligible environment," Lecky wrote: "In a

[1] See the "Biographical Sketch" in the 1961 edition of Lecky's book. At the time of his death, Lecky was employed as an instructor in the Extension Division of Columbia University, having been fired seven years earlier from a faculty position at Columbia College for failing to complete his Ph.D. dissertation.

[2] Lecky (1945), 82.

[3] "One source of motivation only, the necessity to maintain the unity of the system, must serve as the universal dynamic principle" (81). "By interpreting all behavior as motivated by the need for unity, we understand particular motives or tendencies simply as expressions of the main motive, pursuing different immediate goals as necessary means to that end" (82).

This chapter originally appeared in *Philosophical Perspectives* 14, Action and Freedom (October 2000), pp. 349–77. It is reprinted by permission of Blackwell Publishing. Work on this chapter was supported by a fellowship from the John Simon Guggenheim Memorial Foundation and by a matching leave funded by the Philosophy Department and the College of Literature, Science, and the Arts, University of Michigan. For comments on an earlier draft of the chapter, I am grateful to Elizabeth Anderson, Elliot Aronson, Don Herzog, Richard Nisbett, Bill Swann, and Dan Wegner.

world which is incomprehensible, no one can feel secure."[4] The subject therefore constructs an organized conception of his world – an "organization of experience into an integrated whole" – and this organization just *is* his personality, because the effort to maintain its consistency is what gives shape to his thought and behavior.[5]

Central to the personality, so conceived, is the subject's conception of himself. "The most constant factor in the individual's experience," according to Lecky, "is himself and the interpretation of his own meaning; the kind of person he is, the place which he occupies in the world, appear to represent the center or nucleus of the personality."[6] Because the subject's world-view is thus centered on his self-view, his efforts to maintain coherence in the one are centered on maintaining coherence in the other. "Any idea entering the system which is inconsistent with the individual's conception of himself cannot be assimilated but instead gives rise to an inconsistency which must be removed as promptly as possible."[7]

If a person is to maintain consistency in his self-conception, he has to *be* consistent – to think and behave in ways that lend themselves to a coherent representation. That's why the person's conception of his world, and especially of himself, can play the functional role of his personality: it organizes his thought and behavior into a unified whole. Lecky offered the following illustration of how this process works.[8]

> Let us take the case of an intelligent student who is deficient, say, in spelling. In almost every instance poor spellers have been tutored and practiced in spelling over long periods without improvement. For some reason such a student has a special handicap in learning how to spell, though not in learning the other subjects which are usually considered more difficult. This deficiency is not due to a lack of ability, but rather to an active resistance which prevents him from learning how to spell in spite of the extra instruction. The resistance arises from the fact that at some time in the past the suggestion that he is a poor speller was accepted and incorporated into his definition of himself, and is now an integral part of his total personality.... His difficulty is thus explained as a special instance of the general principle that a person can only be true to himself. If he defines himself as a poor speller, the misspelling of a certain proportion of the words which he uses becomes for him a moral issue. He misspells words for the same reason that he refuses to be a thief. That is, he must endeavor to behave in a manner consistent with his conception of himself.

[4] *Ibid.*, 50.

[5] *Ibid.*, 85. See also 90.

[6] *Ibid.*, 86.

[7] *Ibid.*, 136.

[8] *Ibid.*, 103–04.

I regard this as one of the most remarkable passages in twentieth-century moral psychology. On the one hand, it offers an explanation for a pathology that has become especially significant to us – the pathology of being defeated by a negative self-conception. We now look for this particular form of self-defeat not only in children's failure to learn spelling but also, for example, in the perpetuation of racial and sexual stereotypes that are internalized by their victims.[9] On the other hand, this passage also offers, in capsule form, a theory of moral motivation. It says that a person refrains from stealing because he cannot assimilate stealing into his self-conception.

What's remarkable about the passage is that it attributes self-defeat and moral behavior to one and the same motive. A child fails to learn as if on principle, while thieving is, as it were, against his stereotype; or, rather, acting on principle and acting to type are both manifestations of one and the same drive, to maintain a coherent self-conception. Could the question *Why be moral* be so closely related to *Why Johnny can't spell*?

Clearly, Lecky overstated his hypothesis in the passage just quoted. Many psychological factors may go into causing a particular person to spell badly or to refrain from stealing: a self-consistency motive is unlikely to be the only cause or even the primary cause of such behavior. But I would like to believe that this motive can figure among the causes, in roughly the manner described by Lecky; and so I would like to believe that the quoted passage is merely exaggerated rather than false.

My reasons for *wanting* to believe this don't amount to reasons for *believing* it, because they are philosophical rather than empirical. I have presented these reasons elsewhere, in arguing that various philosophical problems about agency can be resolved by the assumption that agents have a motive for doing what makes sense to them.[10] People's having such a motive, I claim, would account for their being autonomous, acting for reasons, having an open future, and thus satisfying our concept of an agent. As a philosopher of action, then, I hope that Lecky is right.

I think that my arguments may be of philosophical interest even if Lecky is wrong, since they show our concept of agency to be realizable, whether or not it is realized in human beings.[11] But experience has taught me that philosophers aren't interested in an account of purely possible agents, and that they tend to regard my account as no more than that,

[9] For recent research on this topic, see Jussim, Eccles, & Madon (1996).

[10] Velleman (1989), (1993), and the Introduction to Velleman (2000c).

[11] See the Introduction to Velleman (1989).

because they find its motivational assumption implausible. Philosophers are generally unwilling to believe that people have a motive for doing what they understand.

I have therefore decided to venture out of the philosophical armchair in order to examine the empirical evidence, as gathered by psychologists aiming to prove or disprove motivational conjectures like mine. By and large, this evidence is indirect in relation to my account of agency, since it is drawn from cases in which the relevant motive has been forced into the open by the manipulations of an experimenter. The resulting evidence doesn't tend to show the mechanism of agency humming along in accordance with my specifications; it tends to show the knocks and shudders that such a mechanism emits when put under stress. But we often learn about the normal workings of things by subjecting them to abnormal conditions; and viewed in this light, various programs of psychological research offer indirect support to my account of agency. I'll begin by reviewing the relevant research, leaving its relevance to my account of agency for the final section of the chapter.

Cognitive Dissonance

The largest and most well-known program of research on cognitive motivation is the theory of cognitive dissonance. In the classic demonstration of dissonance, by Festinger and Carlsmith (1959), subjects performed an extremely tedious task and then were asked to tell the next subject that the task was enjoyable. Some were offered $1 by the experimenter for performing this service; others were offered $20. Those who received only $1 for saying the task was enjoyable subsequently came to believe that it *was* enjoyable, whereas those who had received $20 continued to believe that it was tedious. Festinger and Carlsmith hypothesized that the subjects who received only $1 experienced greater "dissonance" between their attitudes and their behavior, and altered their opinion in order to reduce this dissonance.

The effect reported by Festinger and Carlsmith has been replicated hundreds if not thousands of times, but its interpretation remains controversial. Festinger and Carlsmith did not clearly explain their dissonance hypothesis, and others have proposed alternative hypotheses to account for their results.[12]

[12] For recent contributions to the dissonance debate, see Harmon-Jones & Mills (1999).

Aronson's Version of the Dissonance Hypothesis

The clearest version of the dissonance hypothesis was proposed by Elliot Aronson (1968).[13] Aronson argued that the subjects' cognition of their behavior was at odds with what they would expect themselves to have done under the circumstances. What they would expect themselves to have done, having found the task boring, is to say that it was boring; but they found it boring and said that it was interesting. Their cognition of what they had done therefore clashed with the expectation that would naturally follow from their cognition of the circumstances. The subjects changed their opinion of the task, according to Aronson, so that they could change their cognition of the circumstances, rendering it consistent with their cognition of what they had done.

The hypothesis that Aronson thus framed in terms of expectations can also be framed in terms of explanations. Just as the subjects' cognition of having found the task boring would lead to an expectation at odds with their having said that it was interesting, so it would leave them at a loss to explain why they had said that it was interesting. Finding their behavior inexplicable and finding it contrary to expectation would be two aspects of the same cognitive predicament. And changing their opinion of the task would resolve the predicament under either description, by rendering their behavior both explicable and predictable under the circumstances as re-conceived.

This hypothesis relies on two assumptions that are not explicitly stated either by Aronson or by Festinger and Carlsmith. The first assumption, pointed out by Kelley (1967), is that *none* of the subjects knew the full explanation of their behavior.[14]

The pressure that induced these subjects to lie was covert: it was the pressure exerted by the experimental setting and the authority conferred by that setting on the experimenter. People are notoriously unaware of how powerful such pressure can be.[15] Hence the subjects in Festinger and Carlsmith's experiment didn't know why they lied. One group of subjects were offered an explanation designed to seem adequate to them, while the others were offered an explanation designed to seem inadequate. In all probability, $20 would not have been sufficient to induce most of the subjects to lie if it had been offered by a stranger with no authority; but

[13] See also Aronson (1969); Thibodeau & Aronson (1992).

[14] On this point, see also Nisbett & Valins (1972).

[15] See Sherman (1980), in which subjects greatly under-predicted their compliance with a typical dissonance protocol.

$20 was sufficient for the subjects *to believe* that it had been a sufficient inducement for them, whereas $1 was not. Hence some of the subjects but not others were supplied with what seemed like an adequate explanation of their behavior, or an adequate basis on which to expect it.

The second assumption required by the dissonance hypothesis is that the subjects who changed their opinion also deceived themselves about having changed it. The awareness of having retroactively come to believe what they had already said would not have provided them with an explanation of why they had said it, or with a basis on which their saying it could have been expected. The premise required to explain or predict their behavior was that they had believed what they were saying at the time, as they said it. In retroactively coming to believe what they had said, then, they must also have come to believe, falsely, that they had believed it all along.

As supplemented by these assumptions, the dissonance hypothesis says that when people cannot identify the forces that have shaped their behavior, they conjure up forces to make it seem intelligible and predictable – if necessary, by retroactively forming a motivationally relevant attitude and projecting it back in time.[16] This maneuver would appear to be motivated by the subjects' desire for explanatory and predictive coherence in their self-conceptions, a motive of the sort postulated by Lecky. Hence the results of forced-compliance experiments, as explained by the dissonance hypothesis, appear to support Lecky's theory of self-consistency.

A Rival Explanation: Self-Perception

Daryl Bem (1972) has argued that subjects who seem to be motivated by cognitive dissonance are merely interpreting their own behavior as if they were external observers:

> Just as an outside observer might ask himself, "What must this man's attitude be if he is willing to behave in this fashion in this situation?" so too, the subject implicitly asks himself, "What must my attitude be if I am willing to behave in this fashion in this situation?" Thus the subject who receives $1 discards the monetary inducement as the major motivating factor for his behavior and infers that it must reflect his actual attitude; he infers that he must have actually enjoyed the tasks. The subject who receives $20 notes that his behavior is adequately accounted for by the monetary inducement, and hence he cannot extract from the behavior any information relevant to his actual opinions; he is in the same situation as a control subject insofar as information about his attitude is concerned. (pp. 16–17)

[16] See also Nisbett & Wilson (1977).

In this "self-perception explanation," Bem says, "there is no aversive motivational pressure postulated." As described by Nisbett and Valins (1972), "Bem's reinterpretation of dissonance phenomena avoids the use of any motivational concept," and so "the two positions appear to be at a logical impasse": dissonance theory attributes the subjects' change of attitude to "a motivated process" whereas Bem attributes it to "a passive, inferential process."[17]

Subsequent research has shown that the correct explanation for dissonance phenomena is indeed motivational.[18] But I do not want to interpret this research as discrediting Bem's explanation in terms of self-perception. Bem's only mistake, I believe, is in claiming that his self-perception explanation doesn't depend on any motivational postulate. In fact, his self-perception theory postulates the same motive as cognitive-dissonance theory; and so the two theories give coordinate explanations that are mutually reinforcing.[19]

Bem's thesis is that a person often comes to know about his own attitudes in much the same way as we do when observing him from the outside. Bem's formulation of the thesis suggests that a person receives his self-knowledge passively, as if by a process of sensory perception. But this suggestion is superfluous to the thesis – and, indeed, incompatible with Bem's defense of it.

To be sure, our knowledge of a person's attitudes is often obtained by a process that is quasi-perceptual. That is, hearing a person's vocalizations is often inseparable from hearing them as the assertion of a particular proposition, and seeing his bodily movements is often inseparable from seeing them as an effort to attain a particular end. On other occasions, however, we hear a person's voice without hearing what he's saying, or we see his movements without seeing what he's doing; and what he is saying or doing are then matters that we have to figure out.

On the latter occasions, detecting the attitudes behind a person's behavior requires a process that is not passive and perceptual but active and intellectual: it requires a process of interpretive inquiry. We will not undertake that process unless we have the requisite motives. If the meaning of someone's vocalizations or movements doesn't impress itself upon us immediately, we won't bother to figure it out unless we want to – that is,

[17] *Ibid.*, p. 68.

[18] Zanna & Cooper (1974); Zanna, Higgins, & Taves (1976); Cooper, Zanna, & Taves (1978); Higgins, Rhodewalt, & Zanna (1979); Elliott & Devine (1994).

[19] See Fazio, Zanna, & Cooper (1977).

unless we want to understand what he is saying or doing, or to anticipate what he's likely to say and do next. On such occasions, interpretation is an activity that must be motivated.

Unfortunately, social psychologists tend to speak of all interpretation as perceptual, as in the phrases "self-perception," "interpersonal perception," "social perception," and the like. This usage highlights the cases in which interpretation is passive and receptive rather than active and motivated. Even though Bem denies that one is automatically given a knowledge of one's attitudes, he calls the process of acquiring such knowledge "self-perception," and so he is naturally interpreted as describing a process that is passive. The possibility that self-perception might be a motivated activity therefore goes unnoticed.

Yet Bem himself implicitly concedes this possibility, when he says that an observer "might ask himself, 'What must this man's attitude be if he is willing to behave in this fashion in this situation?'" This question expresses the observer's desire to understand and anticipate, a desire without which he wouldn't ask the question or, having asked it, would let it go unanswered. Bem's thesis is that the subject "implicitly asks himself" the same question about himself, because a knowledge of his attitudes is not automatically given to him any more than it is to an observer. To portray the subject as asking this question is to acknowledge that he doesn't passively perceive his own attitudes but must sometimes actively inquire into them. It is therefore to acknowledge a motive for self-inquiry.

Now, an interest in explanation and prediction is the basis of all consistency motivation, according to Lecky. Consistency isn't desired for its own sake; it's desired as the form of the predictable and the intelligible, by a creature who "must feel that he lives in a stable and intelligible environment." Inconsistency isn't intrinsically disturbing; it's disturbing because it stymies comprehension, and "[i]n a world which is incomprehensible, no one can feel secure."[20] A desire for consistency in one's self-conception thus arises, according to Lecky, from the desire to understand and be able to anticipate oneself – the very desire expressed by the question that Bem attributes to the self-interpreting subject.

Hence the motive implicitly conceded by Bem is the same motive that is explicitly postulated by Lecky, the desire for self-knowledge. Bem should not deny the existence of this cognitive motive, since his own theory presupposes it.

[20] These phrases are drawn from the quotations on pp. 224–5.

Lecky's insight was that the desire for self-knowledge can drive either of two coordinate processes. If we want to understand what we do, we can either figure out why we've done things, after we've done them, or we can make sure that we don't do things unless we already know why. The latter process entails doing only what we are aware of having motives or other dispositions to do; and so it amounts to the process of being true to ourselves by acting in accordance with our self-conceptions, the self-consistency process described by Lecky. The former process entails interpreting our behavior after the fact: it is the self-perception process described by Bem. Self-consistency and self-perception are thus two phases of a single activity – the practical and intellectual phases of self-interpretation.

Note that dissonance-reduction lies on the intellectual side of this contrast. When a subject experiences cognitive dissonance, it's too late for him to make his behavior consistent with his self-conception; he has to adjust his self-conception to fit his behavior, which is in the past. Bem and Aronson are thus describing one and the same process, of fitting an interpretive hypothesis to past behavior.

In Aronson's story, the subject begins with an interpretation that doesn't fit – namely, that he believed the experimental task to be tedious – and the resulting discomfort moves him to frame a new hypothesis, that he believed the task to be fun. In Bem's story, the subject appears to have no initial interpretation, and so he isn't motivated by any discomfort. But he is still motivated by a desire for an interpretation that fits his behavior, the desire whose frustration caused the initial discomfort in Aronson's version. The only point of disagreement between Aronson and Bem is whether the subject was dissatisfied with one interpretation before being moved to frame an interpretation that satisfied him.[21]

[21] There may be one other point of disagreement, but it is too small to worry about. Aronson tends to describe the subject as changing his belief about the experimental task; Bem tends to describe him as attributing such a belief to his earlier self. As I have explained, Aronson's version requires the assumption that the subject not only forms the belief in the present but also projects it back into the past, thus arriving at the same attribution as in Bem's version. But Bem's version says only that the subject attributes the belief to his earlier self – which, in principle, he could do without forming the belief in the present. Bem doesn't say that the subject now believes the task to have been enjoyable; what Bem says is that the subject believes himself to have believed what he was saying when he said that the task was enjoyable.

But this in-principle difference between Bem and Aronson makes no difference in practice. Surely, if someone believes that shortly after finishing a task, he believed it to have been enjoyable, then he is likely to believe, in the present, that the task was enjoyable, unless he has some reason for doubting the truth of his earlier belief. In

Thus, the dissonance theorist and the self-perception theorist aren't at a "logical impasse": they are in fact comrades in arms. Both are describing a process of self-interpretation, motivated by a desire for self-knowledge.[22]

If the differences between dissonance theory and self-perception theory are so small, why do their proponents believe that they yield different predictions, and that they are consequently supported by different experimental results?[23] The answer, I think, is that each theory isolates and simplifies one aspect of a large and complicated reality. The reality behind both theories is the holistic process of fitting an interpretation to behavior – a process that is complicated, in the first-personal case, by the possibility of working in the opposite direction, by fitting one's behavior to an interpretation. Each theory treats a single aspect of this process as if it were the whole, thus obscuring the fact that they are, as it were, different ends of the same elephant.[24]

I have already explained how self-perception theorists focus on the case of automatic, passive self-understanding, neglecting cases in which self-understanding is attained through active self-inquiry, motivated partly by the discomforts of reflective ignorance and incomprehension. The

the absence of such a reason, he will probably be unable to attribute the belief to his former self without adopting it. Bem and Aronson agree that what the subject wants is to attribute the belief to his former self, so as to account for his behavior. Bem doesn't mention that the subject will probably have to adopt the belief in order to attribute it to his former self, in the absence of reasons for doubting it. But this omission hardly constitutes an important difference of opinion with Aronson.

[22] In a paper presented to the 1967 Nebraska Symposium on Motivation, Harold Kelley connected dissonance theory and self-perception theory by way of attribution theory, which he presented as containing a "broad motivational assumption," to the effect that attributional processes "operate *as if* the individual were motivated to attain a cognitive mastery of the causal structure of his environment." Commenting on the passage in which this statement occurred, Bem says: "It is an admirable attempt, but the strongest motivation to emerge from this quotation appears to be Kelley's need to understand why he was there" (Bem [1972], p. 45).

This remark is wonderfully self-refuting. Bem claims that the processes discussed by Kelley involve no motivation, despite Kelley's assertions to the contrary. Yet Bem's attempt to discredit the latter assertions ends up confirming them instead. Bem suggests that we shouldn't credit Kelley's assertions about motivation because he made them only in order to satisfy his need to understand why he was addressing a symposium on motivation. The psychological process to which Bem thus attributes Kelley's assertions is one of the very processes whose existence Kelley was asserting – a process driven by the need for self-understanding. What better reason could we have for accepting Kelley's assertions as true than his having made them out of a need to understand himself?

[23] See Bem (1972); Nisbett & Valins (1972).

[24] This view of the debate is suggested by Aronson (1992).

corresponding fault among dissonance theorists is a tendency to focus on individual inconsistencies of particular kinds, neglecting the overall cognitive goals in relation to which inconsistency is undesirable, in the first place. The narrow focus of either theory allows its proponents to state specific algorithms – an algorithm for attributing attitudes, in the one case, and an algorithm for computing total dissonance, in the other. These algorithms do yield conflicting predictions, which would be confirmed by different experimental outcomes. But the algorithms are radically underdetermined by the theories to which they have been attached, and in both cases they are implausible. Attributing attitudes and eliminating inconsistencies are two aspects of the overall process of making sense of the world, a process that has not been and probably cannot be reduced to an algorithm.

Consider an experiment by Snyder and Ebbesen (1972), billed as "a test of dissonance theory versus self-perception theory."[25] This experiment modified a standard dissonance protocol by making salient to the subject either his initial attitude, or the behavior inconsistent with that attitude, or both. Snyder and Ebbesen claimed that making the subject's attitude salient to him ought to increase his awareness of dissonance between it and his behavior, thereby increasing his tendency to alter the attitude, if dissonance theory were correct; whereas if self-perception theory were correct, making the subject's attitude salient ought to discourage him from attributing a different attitude to himself. Snyder and Ebbesen reported that their results favored self-perception theory in this respect.

But the "prediction" that Snyder and Ebbesen derived from dissonance theory depends on a very narrow view of the circumstances. To be sure, if the subject's belief that a task was tedious is made salient to him, then he will be more aware of its inconsistency with his statement that the task was interesting.[26] But his initial belief would also be inconsistent with a subsequent *belief* that the task was interesting, and this potential inconsistency will also be impressed on him by the salience of the former belief. Once he is made aware of believing that the task was tedious, he cannot come to believe that it was interesting without acknowledging that he has changed his mind, for no apparent reason. Thus, even as the discomfort associated with his initial belief is intensified, so is the discomfort to be expected from the alternative.

[25] Discussed by Bem (1972), 31–33.

[26] For ease of exposition, I speak here as if Snyder & Ebbesen applied their modifications to the dissonance experiment of Festinger & Carlsmith (1959). In fact, they modified a different dissonance experiment, but the differences aren't relevant in this context.

What, then, does dissonance theory predict that the subject will do? Surely, the theory cannot make a definite prediction in a case so underdescribed. What the subject will do depends on circumstances that will vary from one subject to another, since he will seek the most coherent view of the situation *all things considered.* Among the things he'll have to consider will be such questions as what the task was, specifically, and how offensive to his personal tastes; whether he is the sort of person to lie, or the sort of person to be unsure of what he likes; whether he identifies with the experimenter or with his fellow subjects; and so on. If dissonance theorists think that they have an algorithm for predicting how such questions will be resolved, they are mistaken. But their critics are also mistaken if they think that the failure of some particular algorithm entails the failure of the theory.

Another Rival Explanation: Self-Enhancement

I have now argued that self-perception theory tacitly presupposes the same cognitive motive that is explicitly postulated by dissonance theory, and that these theories can appear to yield conflicting predictions only if formulated with more precision than their shared theoretical basis can support. But self-perception theory is not the only attempt to re-interpret the evidence gathered in dissonance research. Others have explained that evidence by postulating motives that clearly aren't cognitive.

According to dissonance theory, the Festinger-Carlsmith subjects came to believe what they had said in order to escape a specifically cognitive predicament, of being unable to explain their behavior, or of finding it contrary to expectation. But they might instead have come to believe what they had said in order to escape the appearance of having been irrational, in having said it for no good reason. In that case, their change of mind would have aimed to rationalize their past behavior rather than to remedy their current state of reflective ignorance or incomprehension; and it would thus have aimed at removing not a cognitive problem but a threat to their self-esteem as rational agents. The effects of forced compliance have therefore been taken by other psychologists to indicate a motive for attaining a favorable view of oneself rather than for maintaining consistency with one's actual self-view – a motive of self-enhancement rather than self-consistency.[27]

[27] See, e.g., Steele & Liu (1983). Aronson (1968) rightly points out that this self-enhancement hypothesis can be subsumed under the hypothesis of cognitive dissonance. People tend to conceive of themselves as rational agents, and their perception of having acted with insufficient justification will be inconsistent with this self-conception,

Of course, dissonance theory is not committed to denying the influence of the former motive. Everyone prefers not to look foolish, and this preference may well be implicated in the forced-compliance phenomena. Dissonance theory is merely committed to asserting the influence of an additional motive, a motive to avoid that which is inexplicable or contrary to expectation. But if the phenomena cited in support of dissonance theory can be completely explained by a desire not to look foolish, then the theory will have lost most of its empirical support.

Attribution Effects

A number of experimenters claim to have found dissonance effects that cannot be explained as instances of self-enhancement.[28] But I am less interested in dissonance *per se* than in the cognitive motive for reducing it – a motive that, as we have seen, can drive not only dissonance-reduction but self-perception and other attributional processes as well. I therefore prefer to draw further evidence from phenomena that don't clearly involve dissonance but turn out to involve the same cognitive motive.

Self-Verification

One such phenomenon has been explored by William B. Swann, Jr., under the label "self-verification."[29] Swann has shown that people tend to seek, credit, and retain feedback that confirms their actual self-conception, even if that conception is negative. Thus, for example, people tend to choose and feel committed to partners who view them as they view themselves, for better or worse.[30] When interacting with someone who appears to view them differently, they tend to behave in ways designed to bring him around to their view, even if it is unflattering.[31] They also lend more credence to his feedback about them, and are more likely to

producing a higher-order dissonance that is cognitive. Yet there is a difference between a desire to avoid seeming irrational and a desire to avoid the inconsistency of believing that one is rational while also believing that one has behaved irrationally.

[28] See, e.g., Prislin & Pool (1996); Stone, Cooper, Wiegand, & Aronson (1997).

[29] For reviews of Swann's research, see: Swann (1983); Swann (1985); Swann and Brown (1990); McNulty and Swann (1991); Swann (1996).

[30] Swann, De La Ronde, and Hixon (1992); Swann, De La Ronde, and Hixon, (1994); Swann, Pelham, and Krull (1989), Study 3. See also Swann and Predmore (1985) and the research by Swann and B.W. Pelham reported in Swann (1986), pp. 419–20.

[31] Swann and Read (1981b), Investigation II; Swann and Hill (1982).

remember that feedback, if it confirms their conception of themselves, favorable or unfavorable.[32]

Because these tendencies are associated with negative as well as positive self-conceptions, they cannot be explained by a desire for self-enhancement.[33] Yet they don't exactly confirm the Leckian hypothesis. What Swann and his colleagues have found are biases in people's collection and interpretation of feedback from others. These biases may well be motivated by a self-consistency motive such as Lecky postulated. But Lecky hypothesized that this motive would lead people to confirm their self-conceptions directly, by behaving in ways that verified those conceptions. Lecky's hypothesis was that people who think of themselves as poor spellers would not just choose friends who think of them as poor spellers, too, but would actually tend to spell poorly in order to be true to themselves. Swann's research does not demonstrate a tendency toward such direct, behavioral self-verification.

There is ample evidence for behavioral self-verification of positive self-views, especially in children. This evidence is less than conclusive, because it can be explained, at least in part, by a motive for self-enhancement; but it is nevertheless worth reviewing, since it coincides in interesting respects with Lecky's views.

Miller, Brickman, and Bolen (1975) compared attribution and persuasion as means of modifying the behavior of elementary school pupils. In one experiment, they compared the littering behavior of children who had repeatedly been told that they *ought to be* tidy (persuasion) with that of children who had repeatedly been told that they *were* tidy (attribution). Both groups decreased their rate of littering, in comparison with both their own prior rate and that of a control group that was offered no messages on the subject. But the effects of attribution were significantly greater and lasted significantly longer than those of persuasion. Children who had been told that they were tidy showed a sharp and lasting decrease in their rate of littering, whereas children who had been told that they ought to be tidy showed only a moderate decrease in littering and then returned to littering at the same rate as the control group. In another experiment, the same researchers found a similar difference in the effect of attribution and persuasion on children's performance in arithmetic.

[32] Swann and Read (1981b), Investigation III.

[33] See Swann (1986); Swann, Pelham, and Krull (1989); Swann, Hixon, Stein-Seroussi, and Gilbert (1990); Swann, Stein-Seroussi, and Giesler (1992); Jussim, Yen, and Aiello (1995).

Children told that they *were* skillful and highly motivated in arithmetic showed a greater and more long-lasting improvement than children told that they *ought to be* skillful or motivated.

These experiments compared favorable attributions with injunctions, which did not have a similarly favorable tone and might even have been interpreted by the children as presupposing an unfavorable attribution instead. (Why would teacher exhort us to be tidy if we weren't in fact untidy?) Perhaps, then, the experiments demonstrated, not an interesting motivational difference between attributions and injunctions, but an utterly unsurprising difference between positive and negative reinforcement. Grusec and Redler (1980) sought to rule out this alternative explanation by comparing favorable attributions with equally reinforcing praise offered for the same behavior. Children who had won marbles in a game were induced to deposit some of them in a collection bowl for poor children, whereupon they were either praised for doing so (reinforcement), told that their doing so showed that they liked to help others (attribution), or given no feedback at all (control). The children were then left to play the marble game on their own, while an experimenter observed through a one-way mirror to record how many marbles they placed in the collection bowl. Finally, the children were given colored pencils as a reward for their participation in the experiment and told that they could deposit some of them in a box for classmates who had not participated. Among 8-year-olds, the treatments were equally effective in increasing donations of marbles, but only attribution increased the donation of pencils. The 8-year-olds generalized their increased helpfulness to a new situation only if they had heard themselves described as helpful. These results were confirmed in subsequent sessions with the same children.[34]

Grusec and Redler gathered additional, developmental evidence by repeating the marble-and-pencil experiment with older and younger children. Neither treatment had any effect on 5- and 6-year-olds, while they

[34] On a later occasion, the 8-year-olds were induced to help a different experimenter prepare materials for building toy houses, and they were again given praise, an attribution of helpfulness, or no feedback. They were then left alone and allowed to choose between playing with a toy or continuing with the helpful task, while an experimenter observed through a one-way mirror. Only attribution showed an effect on their tendency to help. One or two weeks later, these children were given an opportunity to donate drawings and craft materials to hospitalized children. Although total donations were too few for a full statistical analysis, more donations were received from the attribution group than from either of the others.

were equally effective on the 10-year-olds. Grusec and Redler hypothesized that the former subjects were too young to understand the implications of trait attributions, whereas the latter were sufficiently mature to extend the attributions on their own, without hearing them from the experimenters.

This developmental hypothesis was subsequently bolstered by research applying Freedman and Fraser's (1966) "foot-in-the-door" technique to children in the same range of ages. Eisenberg *et al.* (1987, 1989) rewarded children with prize coupons for participating in an experiment, and then induced some of them to donate part of their winnings to the poor. By eliciting this first donation, the experimenters had gotten a "foot in the door," designed to help them elicit further sharing behavior. Eisenberg *et al.* found that children were susceptible to this technique only if they were old enough to demonstrate an understanding of trait stability; and then their susceptibility was correlated with an independent measurement of their motivation toward self-consistency. These results suggest that the technique depended on the children's motivation to behave in accordance with self-attributions of helpfulness or generosity induced by their first donation.[35]

All of these experiments seem to show subjects being true to themselves by behaving in ways that verify self-attributions. Indeed, some of the experiments seem to confirm Lecky's claim that such behavioral self-verification accounts for moral behavior, while others seem to confirm his corresponding claim about academic performance. Hence attribution research with children is at least consistent with Lecky's views on the connection between being a bad speller and not being a thief.

Unfortunately, these findings involve the attribution of positive traits, and so they can in principle be explained by a motive of self-enhancement.[36] The children who heard themselves described as tidy or helpful may have come to regard tidy or helpful behavior as a way of earning that favorable description rather than as a way of making sense to themselves in light of it. Of course, a self-enhancement motive would not necessarily account for the difference in effectiveness between

[35] For research connecting the foot-in-the-door effect to self-consistency motivation in adults, see Kraut (1973) and Goldman, Seever, & Seever (1982). For contrary findings, see Gorassini & Olson (1995). For other experiments in which children show a tendency to verify attributions, see Jensen & Moore (1977), Toner, Moore, & Emmons (1980), Biddle *et al.* (1985), and McGrath, Wilson, & Frassetto (1995).

[36] The same is true of Jensen and Moore (1977); Toner, Moore, & Emmons (1980); Goldman, Seever, & Seever (1982); and Kraut (1973).

attribution and praise, or for the observed correlations with the development of trait-based self-understanding or with independently measured levels of motivation for self-consistency. But these phenomena may be too subtle to determine a choice between rival explanations.

What would confirm the existence of a cognitive motive is evidence that people tend to verify self-conceptions that don't enhance their self-esteem. Some researchers have therefore attempted to demonstrate a tendency to confirm negative self-conceptions.

In the classic experiment of this type, Aronson and Carlsmith (1962) asked subjects to identify the pictures of schizophrenics from among pictures that had in fact been randomly cut from a Harvard yearbook. Since subjects had no grounds for questioning the feedback they received about their rate of success, that feedback could be manipulated by the experimenters. Some subjects were led to believe that they were being consistently successful or unsuccessful; others were led to believe that they were scoring a long string of failures followed by a short string of successes, or a long string of successes followed by a short string of failures. All subjects were then given an opportunity to re-do the last set of items, on which some of them had seemed to take a turn for the better or the worse. Those who had seemed to take such a turn changed more of their answers, even if the turn they had taken was for the better. They thus appeared to prefer scoring consistently poorly to scoring inconsistently – as if trying to confirm the self-conception that they had formed during their initial string of failures.

Unfortunately, efforts to duplicate this result have met with only intermittent success.[37] One possible explanation, proposed by Swann (1986), is that most of the attempts at duplication have tested the effects of artificially induced self-conceptions about one's ability at a previously unfamiliar task. Yet the tendency to verify a self-conception appears to depend on the degree of certainty with which that conception is held.[38] Hence these experiments may not have induced self-conceptions with the degree of certainty required to produce an observable effect. And, indeed, the most persuasive replication of Aronson and Carlsmith's result was in subjects who had been found to hold negative overall self-views with relative certainty.[39] For these subjects, success was inconsistent

[37] See the review in Dipboye (1977).

[38] Swann & Ely (1984); Maracek & Mettee (1972); see also Setterlund & Niedenthal (1993).

[39] Maracek & Mettee (1972).

not only with an immediately prior series of failures but with a well-entrenched conception of themselves. Even so, this line of research cannot be regarded as clearly demonstrating the presence of self-consistency motivation.

Self-Attribution of Emotion

The research summarized in the previous section is inconclusive partly because it focuses on self-conceptions of personal traits. These traits are often conceived in evaluative terms, and so attributing them to oneself often yields a self-conception that is clearly favorable or clearly unfavorable. A tendency to verify favorable self-conceptions can always be explained by a motive of self-enhancement; and whatever cognitive motive there is to verify self-conceptions may not be sufficiently strong to prevail reliably over the desire to falsify them when they are unfavorable. Hence the self-attribution of personal traits is unlikely to produce clear evidence of self-consistency motivation. A more likely source of such evidence is the self-attribution of motives or emotions, which – unlike traits of character – are often evaluatively neutral.

In the classic experiment on such attributions, Schachter and Singer (1962) recruited subjects for an experiment billed as testing the effects of a vitamin on vision. Two thirds of the subjects were injected with adrenaline, labeled as the vitamin; one third were injected with a saline solution but told it was the vitamin as well. One of the adrenaline-injected groups was informed that the drug would cause symptoms of arousal – trembling hands, racing heart, and so on. The others were not warned of any side effects.

Each of the subjects then moved on to the next activity, at which he was ostensibly joined by a fellow subject, who was in fact a confederate of the experimenters. For half of each group, the confederate became increasingly angry at the next activity; for the other half, the confederate became giddy and playful. The experimenters then observed the extent to which the subjects were influenced by the confederates' behavior. Those who had received a placebo, and those who had received adrenaline and been warned of its side effects, were influenced significantly less than those who had received adrenaline without being warned. The latter group showed a marked tendency to behave as if they were angry or giddy, depending on how their fellow subjects were behaving.

Schachter and Singer hypothesized that the subjects in this group interpreted their arousal as anger or euphoria, according to the suggestion provided by their fellow subjects, and then enacted the emotion that they had attributed to themselves. So interpreted, the experiment showed that people have a tendency to behave in accordance with the motives that they *believe* themselves to have – which would be a tendency toward self-consistency.[40]

The Schachter-Singer results have repeatedly been called into question on methodological grounds.[41] But the underlying hypothesis has been confirmed in experiments of a significantly different design.

Zillman, Johnson, and Day (1974) arranged for subjects to be angered by someone and then to engage in vigorous exercise. Some of the subjects were given an opportunity to retaliate against their provoker shortly after exercising; others were given the same opportunity after a longer interval. The latter group retaliated more intensely than the former. Zillman and his colleagues hypothesized that the subjects' retaliation expressed the degree of anger that they perceived themselves as having; and that the excitatory effects of exercise were correctly interpreted by the first group but misinterpreted by the second as heightened anger.

This hypothesis was subsequently tested by Cantor, Zillman, and Bryant (1975), who asked subjects to report, at intervals following exercise, whether they still felt its excitatory effects. By measuring the subjects' levels of excitation at the same intervals, these experimenters detected an initial phase during which the effects of exercise continued and were perceived as continuing; a second phase during which the effects of exercise continued but were not perceived as such; and a third phase during which these effects had disappeared both objectively and subjectively. During each of these phases, erotic materials were shown to one third of the subjects, who were asked to report their degree of sexual arousal. Subjects exposed to erotica during the first phase reported no greater arousal than those exposed during the third phase; but those exposed during the second phase reported greater arousal than the others. Thus, arousal that was not attributed to exercise appears to have been misattributed to the erotica, supporting the hypothesis of a similar misattribution in the previous experiment.

[40] Actually, Schachter's (1964) theory of emotion implies that people actually *have* the emotions that they believe themselves to have, provided that they are in fact aroused or excited. This feature of Schachter's theory is philosophically problematic, but I won't discuss it here.

[41] Most recently by Messacappa, Katkin, & Palmer (1999).

Zillman (1978) therefore concludes that what led the previous subjects to behave angrily was a self-attribution of anger.[42] These experiments suggest that people tend to manifest not only what they're feeling but also what they think they're feeling. Note that there is no competing explanation of this tendency in terms of self-enhancement, since people are unlikely to regard anger as a self-enhancing attribute. The most likely explanation is that people tend to behave consistently with their self-attributions, being true to themselves in precisely the manner envisioned by Lecky.

Summary

This review of dissonance and attribution research has yielded two tentative conclusions. The research appears to show, first, that we tend to act in accordance with the motives and traits of character that we conceive of ourselves as having. The research is also consistent with a second

[42] For related experiments, see Brodt & Zimbardo (1981); Olson (1990); and the research of Berkowitz, discussed later.

Bem (1972) points out that experiments of this form often show more effect on the subjects' behavior than on their reported attitudes. In the Schachter-Singer study, for example, the misattribution condition was more strongly correlated with a tendency to join in the angry or giddy behavior of a fellow subject than with a tendency to report anger or giddiness. Bem argues that if the behavior were caused, as hypothesized, by the subjects' self-attributions, then the self-attributions ought to have been more strongly correlated with the experimental manipulations, not less. Bem therefore concludes that such experiments support self-perception theory, according to which the subjects' self-attributions were based on their behavior rather than vice versa.

Yet there are many ways of accounting for the weaker effect on attitudes than on behavior, even under the hypothesis that the attitudes came first in the order of causation. After all, the problematic correlations were observed, not with the attitudes themselves, but rather with the subjects' reports of those attitudes. Any gaps in the process of articulating self-attributions could therefore account for the results. Suppose, for example, that the subjects attributed anger or euphoria to themselves but not in so many words, or not in words at all. Perhaps they had mental images of those emotions (say, images of facial expressions or bodily postures), which they immediately associated with particular kinds of behavior, but to which they were not equally quick to attach names. Their self-attributions would then have been less reliable in prompting self-reports than in prompting behavior.

Even if we grant that Bem is right about the order of causation, his interpretation of the results would still support the hypothesis of cognitive motivation. A plausible explanation of the results, even as interpreted by Bem, is that the subjects behaved emotionally in order to facilitate the emotional attributions that would render their feelings intelligible. Feeling aroused, they sought a self-conception that would explain why, and they consequently behaved in ways that would make such a self-conception applicable to them. If their behavior was thus designed to facilitate the attribution, then it preceded the attribution in the order of causation; but it would still have been motivated by a cognitive interest in the self-understanding that the attribution would provide.

conclusion, that this tendency is due to a cognitive motive, to find ourselves explicable and predictable.

In the past I have argued that creatures endowed with such a motive would satisfy our ordinary concept of an agent in the respects that often seem to make that concept seem unsatisfiable. Creatures so motivated would have futures that were open in a sense sufficient to afford them choices or decisions;[43] they would be the causes rather than the mere vehicles of behavior;[44] they would be guided by the normative force of reasons for acting;[45] and they would find such force in principles requiring them to be moral.[46]

I will not repeat these arguments here. What I'll attempt instead is to highlight pieces of the psychological literature that already point the way toward the philosophy of action. This work by psychologists tends to support a philosophical theory like mine.

Philosophical Implications

Some psychologists have gestured toward the philosophy of action in the course of discussing self-consistency motivation. Zillman, for example, having concluded that self-attributions of emotion can influence behavior, goes on to speculate that their influence makes for the difference between automatic manifestations of emotion and emotional actions that are under voluntary control. In Zillman's view, an emotion involves some basic motor responses, which can be reinforced, suppressed, or redirected by the subject's interpretation of them. The basic motor responses belong to "the primitive heritage of man," which we share with the lower animals, and they are not under voluntary control; their modulation by the subject's self-interpretation manifests his "rational capabilities," by which he controls his behavioral response.[47]

Berkowitz draws a similar distinction between impulsive and purposive aggression. In collaboration with Turner (1974), he manipulated the degree of anger that subjects attributed to themselves toward a particular person, thereby modifying the intensity of the "punishment" that they inflicted on that person, though not their aggression toward a third party. Berkowitz emphasizes that the attribution-governed aggression observed

[43] Velleman (1989a).

[44] Velleman (1992b) and Introduction to 2000c.

[45] Velleman (1996) and Introduction to 2000c.

[46] Velleman (1989b), Part Four.

[47] Zillman (1978), pp. 356–57. See also Cross & Markus (1990); Wegner & Bargh (1998).

in this experiment was purposive rather than impulsive. "[I]mpulsive acts," he says, "are automatic, stimulus-elicited responses to the external situation governed primarily by associative factors and relatively unaffected by cognitive processes."[48] By contrast, purposive aggression is subject to cognitive governance:[49]

> The present results generally support [my] cognitive analysis of purposive aggression. Emotionally aroused people seek to attack a particular target when (a) they interpret their internal sensations as "anger," and (b) they believe this specific target had been the cause of their feelings. As indicated in this study, the intensity of the subjects' desire to hurt a particular person, reflected in the intensity of the punishment given him, arose from their perceptions of the strength of their anger and their belief that this person had been the one who had provoked them.

Berkowitz goes on to explain this mechanism in terms of a cognitive motive toward self-consistency:[50]

> Looked at from a larger perspective, the findings also provide yet another demonstration of the search for cognitive consistency. We want our actions to be in accord with our emotions, as we understand them, and apparently we are also disturbed if these feelings do not seem to be warranted by the causal incident. The emotion as well as the behavior must be consistent with our other cognitions.

The idea that behavior becomes purposive or intentional when it is regulated for self-consistency can be traced back to the early days of self-consistency theory. Six years after the publication of Lecky's treatise, Carl Rogers published "A Theory of Personality and Behavior" offering a similar postulate:[51]

> *Most of the ways of behaving which are adopted by the organism are those which are consistent with the concept of self*. . . . As the organism strives to meet its needs in the world as it is experienced, the form which the striving takes must be a form consistent with the concept of self. . . . The person who regards himself as having no aggressive feelings cannot satisfy a need for aggression in any direct fashion. The only channels by which needs may be satisfied are those which are consistent with the organized concept of self.

[48] *Ibid.*, p. 176.

[49] *Ibid.*, pp. 186–87. See also Berkowitz (1987).

[50] *Ibid.* See also Berkowitz (1987).

[51] *Client-Centered Therapy: Its Current Practice, Implications, and Theory* (Boston: Houghton Mifflin, 1951), Chapter 11, 507–08. For other theories in the Leckian tradition, see Snygg & Combs (1959); Kelley (1967); Korman (1970); Epstein (1973), (1981); Andrews (1991); Nuttin (1984).

To this Leckian postulate, Rogers added the following piece of action theory:[52]

> *Behavior may, in some instances, be brought about by organic experiences and needs which have not been symbolized. Such behavior may be inconsistent with the structure of the self, but in such instances the behavior is not "owned" by the individual. . . .* In such instances the individual feels "I didn't know what I was doing," "I really wasn't responsible for what I was doing." The conscious self feels no degree of government over the actions which took place.

According to Rogers, then, only behavior that is regulated for self-consistency is experienced as intentional action, for which the subject takes responsibility. Hence the difference between mere behavior and intentional action – the difference, as Wittgenstein put it, between my arm's rising and my raising it – may be due to the intervention of a self-consistency motive.

Carrying Out an Intention

Other psychologists have filled in the self-verification process with steps that correspond to steps in the production of intentional action, as it is ordinarily understood. They have pointed out that people must have not only a conception of their motives but also a conception of what they are doing out of those motives – for example, that they are "retaliating" against someone, or that they are "donating" to the poor. There is evidence that the latter conception also tends to influence their behavior, thereby playing the role of an intention to act.

Wegner, Vallacher, and colleagues (1986) led subjects through a sham experiment involving a clerical task, and then asked them to complete a questionnaire about the degree to which various descriptions applied to the activity in which they had just participated. Some of the suggested descriptions were designed to test whether the subjects conceived of the activity in low-level, mechanical terms, such as "making marks on paper," or high-level, explanatory terms, such as "participating in an experiment."

[52] *Ibid.*, p. 509. In this quotation, philosophers of action will detect a resemblance to Harry Frankfurt's theory of autonomy (1988c, 1999c). Like Frankfurt, Rogers believed that whether behavior amounts to an autonomous action depends on its relation to the self. But if Rogers had expressed his view in these terms, he would have been using the psychologist's sense of 'self', which refers to the self-conception; whereas Frankfurt uses the term in a philosophical sense referring to the core or essence of the person. Rogers thus resembles Frankfurt partly by courtesy of ambiguity. (For psychologists who prefer a Frankfurtian conception of the self, see Deci & Ryan [1991].)

The last seven items were designed to suggest either altruistic descriptions ("helping people study psychology," "aiding the experimenter") or egoistic descriptions ("getting a better grade in psychology," "earning extra credit"). The experimenters then left the room, to allow the subject-pool coordinator to distribute a questionnaire about the subjects' preferences among future opportunities to participate in research. Among the opportunities offered, one was described in altruistic terms, and another in egoistic terms.

The experimenters found that subjects who initially conceived of the prior activity in low-level terms were more likely to adopt the suggested high-level descriptions and also expressed a higher preference for future opportunities described in similar terms. In other words, subjects who could be induced to think of their current participation as "helping" were more inclined to "help" in the future, whereas subjects who could be induced to think of their current participation as "getting ahead" were more inclined toward future opportunities to "get ahead."

This experiment can be interpreted as demonstrating a "foot-in-the-door" effect; but in this case the effect appears to be mediated by act-descriptions rather that trait- or motive-attributions. The experimenters got their foot in the door by enlisting a subject's participation in one experiment, and they were then able to elicit his willingness to participate in another, but only by getting him to conceive of the second under the same description as the first.[53] This dependence was the same for egoistic as for altruistic actions. Vallacher and Wegner (1985) therefore remark, "although egoism and altruism can represent opposing forces in everyday life, they arise from similar action identification processes."[54]

On the basis of this and related experiments, Wegner and Vallacher have proposed a theory of action identification. The first principle of their theory is "that people do what they think they are doing," by selecting a "prepotent act identity," or act description, and then instantiating it in their behavior.[55] The result, according to Wegner and Vallacher, is that

[53] Kraut (1973) links the foot-in-the-door effect to the attribution of traits, such as "charitable" and "uncharitable," rather than to act-descriptions. But my view is that attributions of traits, motives, and acts are themselves linked, under the principle of self-consistency. That is, someone who conceives of himself as angry finds it consistent to conceive of himself as retaliating; someone who conceives of himself as uncharitable finds it inconsistent to conceive of himself as donating to charity; and so on.

[54] Vallacher & Wegner (1985), p. 143.

[55] Wegner & Vallacher (1986), p. 552.

people usually *know* what they're doing, because they are doing what they think.[56]

This principle can readily be interpreted as describing an intermediate step in the self-verification process described here.[57] We can imagine, first, that the cognitively motivated agent selects a "prepotent act identity" consistent with the motives and other dispositions that he conceives himself to have. Conceiving of himself as angry, he thinks of doing something consistent with anger, such as retaliating; conceiving of himself as generous, he thinks of doing something consistent with generosity, such as making a donation. He thereby maintains the coherence of his self-conception. When he goes on to do what he is thinking, we can regard him as taking the next step in the same process. For we can imagine that he does what he's thinking *in order to* know what he's doing, given that whatever he thinks he's about to do is the thing that he would consequently know about, if he did it. The agent thinks of doing something that fits his self-attributed motive, and then he does what fits this self-attribution of action, so that his self-conception is consistent with itself and with his actual behavior.

Wegner and Vallacher suggest that the agent's "prepotent act identity" is in fact an intention to act: in doing what he thinks, the agent is carrying out an intention.[58] This suggestion enables us to map the self-verification process, as now elaborated, onto the process of intentional action as ordinarily understood. Described in theoretical terms, the process goes like this: first, something arouses the agent's anger, which already involves some behavioral dispositions; then the agent interprets his arousal *as* anger and thinks of what, in light of it, would make sense for him to do; finally, the agent's anger and his thought of behaving angrily jointly cause the corresponding behavior – the behavioral impetus of the one being regulated for consistency with the other by the agent's motive for making sense. But now we can redescribe the same process in ordinary language, by attaching the term 'motive' to the agent's anger and the term 'intention' to his thought of behaving angrily. Thus redescribed, the process goes like this: the agent forms an intention that's consistent with his motive; and then he acts, under the impetus of his motive, as regulated for consistency with his intention. The theory of self-verification can thus be seen to coincide with our ordinary understanding of intentional action.

[56] Ibid., p. 568. See also Velleman (1989), Chapters 1 and 2.

[57] See also Aronson, 1992, p. 307.

[58] Vallacher & Wegner (1985), pp. 6–11.

Acting for Reasons

If the agent's doing what he is thinking constitutes the carrying out of an intention, then what about the preceding step, in which he thinks of doing what would make the most sense? To which phase or aspect of an action, as ordinarily understood, does that earlier part of the self-consistency process correspond?

Wegner and Vallacher allude to this step in a further principle, which says that people ordinarily seek to identify their behavior at a "high" or "comprehensive" level, representing their underlying motives and ultimate goals. Wegner and Vallacher describe this tendency as a "search for meaning in action"[59] or "a human inclination to be informed of what we are doing in the most integrative and general way available."[60] An act-description will be "integrative," of course, insofar as it incorporates the motives and traits that the act expresses and in light of which it will make sense. Hence the "search for meaning" posited by Wegner and Vallacher coincides with the agent's search for an act-description that makes sense in light of his self-conception.

The process of adopting and then instantiating integrative act-descriptions resembles – or, in fact, may just *be* – a process of enacting a coherent narrative.[61] Consider Trzebinski's (1995) discussion of self-narratives, which makes them sound like Wegner and Vallacher's act identities:

> Constructing self-narratives is the mode of searching for a meaning.... To find meaning, and more often just to maintain meaning and avoid disruption of the ordered world, an individual has to move in a specified way within the narrated events. In this way the active schema... not only directs the individual's interpretations of on-going and foreseen events, but also pushes him toward specific aspirations, decisions, and actions. By particular moves within the events an individual elaborates, fulfils, and closes important episodes in the developing self-narrative. Personal decisions and actions are inspired by, and take strength from self-narratives – devices for meaning searching.

A self-narrative can thus provide the meaningful act-descriptions that enable the agent to understand what he's doing. When he instantiates one of these narrative act descriptions, he performs an action that "elaborates, fulfils, and closes" an episode in his self-narrative, so that his behavior is intelligible as part of the story.

[59] Wegner & Vallacher (1986), pp. 555–56.

[60] Vallacher & Wegner (1985), p. 26.

[61] See Velleman (1993).

I suggest that the narrative background on which the agent draws, in order to fashion an integrative act description, is material that would ordinarily be called his reasons for acting – the circumstances, motives, and other considerations that make one action rather than another the sensible thing to do. I therefore suggest that adopting an integrative act description amounts to forming an intention on the basis of a reason, and that enacting such a description amounts to acting for the reason on which the intention was based.

Philosophers have long noted a distinction between doing something that one has reason to do, on the one hand, and doing it *for* that reason, on the other.[62] One can do something that one has reason to do without necessarily doing it *for* that reason, because one can fail to be appropriately influenced by the reason that one has. Philosophers have therefore sought to analyze the influence that a reason exerts when one acts for that reason.

The traditional assumption among philosophers is that a reason for acting must include the expectation of a desired outcome, and that this expectation influences the agent by appealing to his desire for the outcome expected. Elsewhere I have argued that the influence exerted by an agent's expectation of a desired outcome does not satisfy our concept of the influence exerted by a reason.[63] Indeed, the assumption that expectations of desire- or preference-satisfaction have the normative force of reasons is itself in need of justification.[64]

In my view, an agent is influenced by a reason, and his action is consequently performed *for* that reason, when he is influenced by a representation of the action that makes it intelligible to him. Naturally, this representation may make the action intelligible precisely by setting it in the context of his desires and expectations, but his reason for the action consists in this cognitively attractive representation of it rather than in the desires and expectations to which it alludes. A reason is a *rationale*, in light of which an action makes sense to the agent, and promoting a desired outcome is one such rationale.

If I am right, then the search for an integrative act description to instantiate, or a meaningful story to enact, is in fact a search for an action supported by reasons. And an act identity has "pre-potency" insofar as it

[62] E.g., Davidson (1980).

[63] Velleman (1992a), (1992b), (1996); "Introduction" to Velleman (2000c).

[64] Velleman (1993); Korsgaard (1997).

satisfies this search, by serving as a rationale. When the agent does what he is thinking under an integrative act description, or "fulfills and closes" an episode in his self-narrative, he is doing what philosophers call acting for reasons. The upshot is that the steps of finding and acting for reasons correspond to successive steps in the self-verification process, the process of being true to oneself.

Conclusion

Nuttin has illustrated the resulting theory of practical reasoning as follows:[65]

> Consider a son who is tempted to lie to his parents in order to be able to accompany his friends on a vacation despite the anticipated opposition of his father. The son must evaluate and determine the extent to which he is able to integrate within one structure the two conflicting components: his image of himself and the lie to his parents as he perceives it in the present behavioral context. Is he able to take up, subsume, or accept that concrete type of lying within his dynamic self-concept? The strength of the tendency to accompany his friends will be one of the factors determining the degree of distortion of the self-image that can be tolerated by the personality. The degree of inner consistency within the subject's personality will be another factor. In some people, there will be no difficulty at all in subsuming the lying behavior in the self-concept; in other people, accepting such a lie within their own personality functioning will not be possible. In the latter case, the subject is not "willing" to lie in the present behavioral context.

In my view, the conflict between the boy's desire to accompany his friends, on the one hand, and his need for a coherent self-conception, on the other, is a conflict between inclination and practical reason. If the boy finds a way to reconcile the lie with his self-conception – a story to tell himself about telling the lie, which would amount to a rationale for telling it – then his practical reason condones telling the lie, and he is consequently "willing" to tell it. But if he cannot reconcile telling a lie with his self-conception, then his need for self-understanding opposes his telling it, and this opposition embodies the restraint that practical reason places on his inclination to lie.

As Vallacher and Wegner point out,[66] this same process can lead to immoral as well as moral behavior. Or, as Lecky suggested, it can lead to bad spelling. What a Leckian moral philosophy will need, then, is an

[65] Nuttin (1984), p. 187.
[66] Quoted at note 54.

account of why a conception of oneself as honest is more rational than a conception of oneself as dishonest; or, for that matter, why a conception of oneself as a good speller is more rational than a conception of oneself as a poor one, given that one will do as one conceives. I have attempted such an account elsewhere.[67] Here I have tried to connect the underlying moral philosophy to an empirical basis in the psychological research that Lecky inspired.

[67] See Velleman (1989a), Part III.