Pseudorationality[1]

Adrian M.S. Piper

I want to argue that self-deception is a species of a more general phenomenon, which I shall call *pseudorationality*, which in turn is necessitated by what I shall describe as our *highest-order disposition to literal self-preservation*. By "literal self-preservation," I mean preservation of the rational intelligibility of the self, in the face of recalcitrant facts that invariably threaten it. (The preservation of bodily integrity against physical assault - the familiar, metaphorical sense of "self-preservation" - is a necessary but not sufficient condition of literal self-preservation in this sense.) By a "highest-order disposition," I mean a disposition that constrains any other disposition or motive we may have. Although I have touched upon some of these issues elsewhere,[2] all of this will need to be spelled out and defended at some length. But the basic idea, briefly, is this. Perhaps under Hume's influence,[3] we tend to conceive of theoretical reasoning as a kind of contingent mental operation, conscious or unconscious, which we perform on sentential propositions.[4] Whether or not we perform it is often thought to depend on such contingent factors as training (e.g., whether or not we have had a course in first-order logic), personality (e.g., whether or not we persevere in reasoning when the going gets rough), or the presence or absence of some object of desire we must calculate how to achieve. This Humean conception situates theoretical reason at a considerable remove from the kinds of factors - emotions, dispositions, desires, and so forth - we ordinarily recognize as

---

[1] The term is Kant's. See *The Critique of Pure Reason*, trans. Norman Kemp Smith (New York: St. Martin's Press, 1970), A 311/B 368, A 339/B 397, passim. Work on this paper was partially supported by an Andrew Mellon Post-Doctoral Fellowship at Stanford University, 1982-84. It is excerpted from chaps. 11 and 12 of a longer manuscript in progress, *Rationality and the Structure of the Self* (henceforth RSS). I have benefited from discussion and criticisms of the relevant parts of these chapters by Akeel Bilgrami, Paul Boghossian, Don Loeb, Barry Loewer, David Reid-Maxfield, and Sigrun Svavarsdottir. Jeffrey Evans supplied important criticisms, insights, and psychologically relevant data for an earlier version of this paper. I would also like to thank Amélie Rorty for suggestions and criticisms that have both improved this version and made me aware of how very much more there is to say about the topics discussed, and Brian McLaughlin for his care and patience in reading a number of such versions.

[2] See my "Two Conceptions of the Self," *Philosophical Studies* 48, no. 2 (September 1985):173-197; reprinted in *The Philosopher's Annual* VIII (1985), secs. 3-5.

[3] I discuss Hume's conception of theoretical reason in "Hume on Rational Final Ends" (forthcoming).

[4] I am grateful to Akeel Bilgrami and William Frankena for pressing this view in conversation. I speak of sentential propositions rather than sentences in order to avoid the implication that one must have or use a language in order to be theoretically rational. The significance of this will become clearer in what follows.

capable of causal efficacy. It thus practically forecloses the possibility that theoretical reason might be motivationally effective in behavior. At least it is difficult to imagine how anything so seemingly remote from causation could be.

Some Kantians seem to accept the Humean picture. They assert that moral principles are motivationally effective only for an agent who is fully rational, where "full rationality" just means being moved by the thought of certain propositions.[5] The implication is clear that this stipulation does not purport to approximate the empirical facts of human psychology. Actual human agents who adopt such principles but fail to act on them are then portrayed as suffering from some (perhaps extended) form of *akrasia*.

I think the Humean conception is incomplete: theoretical reason is, as Kant saw, intimately tied to certain necessary conditions of selfhood and agency. However, what I shall claim to be the necessary connection between agency and theoretical reason confronts every such agent with the dilemma of her own imperfection: we cannot possibly make rationally intelligible everything that happens to us or everything we feel and do, without threatening the coherence of that which we *do* think we understand rationally. So we cannot possibly integrate all such events without undermining our agency. Rather than do this, we systematically distort and truncate our understanding, with the help of our rational capacities themselves, so as to achieve the illusion of rationality. This is, as a rough first approximation, what I mean by *pseudorationality*.

We do not strictly *have* to engage in pseudorationality. It is psychologically open to most of us simply to endure the anxiety, confusion, and powerlessness that often accompany reminders of our subjective fallibility. It is in our interests to do this. But reminders of our subjective fallibility are much harder to endure, if being right is more important to us than being genuinely rational; i.e., if we have a favored theory of our experience to vindicate. And they are even harder to endure if what reminds us of our subjective fallibility is our own enigmatic or personally unacceptable behavior, rather than some enigmatic or unexpected event in the world. For our own anomalous behavior poses a more immediate threat to our agency than enigmatic external events, and so calls forth an even more intensified mobilization of the resources of pseudorationality to withstand it.

---

[5] Alan Donagan relies on this assumption throughout his *Theory of Morality* (Chicago: Univ. of Chicago Press, 1977); see especially chaps. 2.3, 7.1, 7.3-4, and 7.6. For a resourceful elaboration and defense of this view, see Christine Korsgaard, "Skepticism About Practical Reason," *Journal of Philosophy* 83, no. 1 (January 1986): 5-25.

This is part of what makes self-deception a difficult and central problem for moral theory: no matter how fully developed or compelling our favored moral theory may be, it is useless to us if we are psychologically incapable of admitting to having violated it. "A conscience," Alice Hamilton observed, "may be a terrible thing in a man who has no humility, who can never say, 'I might be mistaken.'" It may be even worse in one who has sufficiently mastered the philosophical reflex to be able to say this without, nevertheless, entertaining it as a serious possibility. Reflective self-knowledge may therefore seem to be the antidote. But if self-knowledge, i.e., being right about oneself, is morally even more important to one than being right about other things, then the lure of self-deception will be all the more pernicious and compelling.

Kant saw this quite clearly. He saw that he really pressing motivational problem for actual moral agents is not *akrasia*, but rather self-deception. *Akrasia* presupposes that we know our motives, our obligations, and hence our moral derelictions with respect to them. But the more importance we accord to such self-knowledge, the more susceptible we are to self-deception about what our moral obligations are, and whether we have fulfilled them.[6] And of course self-knowledge is morally very important to most of us: we console ourselves with the thought that we may not be morally perfect, but at least we know what we are doing wrong.

Rather than examining the morality of self-deception, I want to consider the relation of self-deception to our favored theories in general, and reach some conclusions that will apply *inter alia* to moral theory.[7] I shall suggest that the greater the consolation we derive from the certainty of self-knowledge, the more susceptible we are to self-deception, because our inability fully to

---

[6] See the footnote to A 551 in *Kant's Critique of Pure Reason* (op. cit.), and the further elaborated claim in Kant's *Groundwork of the Metaphysic of Morals*, trans. H. J. Paton (New York: Harper Torchbooks, 1964), at *Ak*. 407-408. Also see Kant's description of a brand of self-deception at *Ak*. 424-425, and compare it with his characterization of man's natural propensity to evil in *Religion Within the Limits of Reason Alone*, trans. T. M. Greene and H. H. Hudson (New York: Harper Torchbooks. 1960), pp. 27-29. For further remarks on the inevitability of self-deception and the inscrutability of our own motives, see the latter work, pp. 17, 33-34, 46, 56-57, 70, 78, 85, and 90-91. I am indebted to Henry Allison for pointing out to me the importance of Kant's preoccupation with self-deception.

[7] The implication is that a moral theory is a theory like any other, the terms of which do not differ from other theories in their semantic status. In "The Meaning of 'Ought' and the Loss of Innocence" (unpublished paper, 1986), I argue that moral theories are false descriptive theories whose normative force is to be explained by our deeply rooted psychological attachments to them.

satisfy the demands of theoretical rationality requires it. But here, too, we do not *have* to deceive ourselves, any more than we have to engage in pseudorationality more generally. It is psychologically open to us to abdicate the aspiration to inviolable agency or to infallibility or to unalloyed moral rectitude by opting for a policy of *epistemic audacity*. If we are really serious about avoiding self-deception, this is in fact the only choice we have.

Some Unargued Assumptions[8]

1. First, I shall say that an event, object, or state of affairs (henceforth a "thing") is *rationally intelligible* to us if we can recognize it as an instance of some concept. To *recognize* something is to perceive it as familiar, i.e., as the same as or similar to something you've perceived before. If something is in no respect like anything you've perceived before, then you cannot identify it at all.

---

[8] For a defense of these, see my *RSS*, chap. 11, and another short paper excerpted from it, "Rationality and the Structure of the Self," presented, with Akeel Bilgrami commenting, to the Association for the Philosophy of the Unconscious, at the Eastern Division Meeting of the American Philosophical Association, Boston, Mass., 1986; the University of Minnesota Philosophy Department, November 1987; the Columbia University Philosophy Department, March 1988; and the Character and Morality Conference, with Nancy Sherman commenting, hosted by Radcliffe and Wellesley Colleges, April 1988; and forthcoming in a volume to be edited by Amélie Rorty and Owen Flanagan.

This discussion in turn attempts to flesh out more systematically some ideas sketched very roughly in sect. 3 of my "Two Conceptions of the Self" (op. cit.) and *RSS*, chap. 12. The notion of the holistic regress and the theoretically rational requirements of horizontal and vertical consistency introduced in the following pages draw heavily on Kant's conception of theoretical reason as developed in the Dialectic of *The Critique of Pure Reason*. See especially A 299/13 355-A 308/B 364, A 322, A 330-332, A 337, B 378-379, B 383, B 387-388, B 437, A 643/13 671-A 669/B 697, "The Regulative Employment of the Ideas of Pure Reason"; compare B 93-94, B105-106 on judgments as functions for unifying our representations. I discuss the interpretation of these passages, and Kant's view of reason more generally, in "Kant's Idea of Reason" (unpublished paper, 1986). In what follows, I do not claim to interpret Kant, but merely to develop some ideas that can be found in Kant's writings. Nevertheless, I shall try to navigate between the Scylla of technical issues in the philosophy of language and the Charybdis of Kant exegesis. My frequent references to Kant are thus intended to provide historical and motivational context for these ideas, not to represent them as what Kant actually meant (nor even, necessarily, what he should have meant).

Second, the requirement of rational intelligibility implies what I shall call a *holistic regress*. This consists in two epistemic facts about us. First, nothing can be rationally intelligible to us in isolation from things to which we recognize it as similar and other things from which we recognize it as differentiated (thus its holism). And second, in order for us to have a concept of the kind of thing some thing or property is, we must have or he able to acquire a host of further concepts of the higher-order kinds of things that kind of thing itself is (thus its regressiveness). For example, if we recognize a thing as red, we must be able to recognize it as a certain color.

Third, the holism of the holistic regress implies that we cannot conceive a thing or property simultaneously as what it is and what it is not, i.e., that all the concepts by which we make sense of the world at a particular moment must simultaneously satisfy the law of noncontradiction.[9] This means that I must conceive all the things and properties that are simultaneously rationally intelligible to me as logically consistent with one another.[10] Call this the requirement of *horizontal consistency*.

Fourth, the regressiveness of the holistic regress implies that I must conceive the higher-order properties by which I recognize something, as logically entailed by the relevant lower-order ones as a matter of conceptual

---

[9] Note that what satisfies the law of noncontradiction is not the relation as we conceive it between things *and* their higher-order properties. So this requirement cannot be expressed by the relation between a predicate letter and the objects that fix its extension, thus:

(1) (x) -(Fx & - Fx).

What is required to satisfy the law of noncontradiction here is rather our concepts of the objects assigned to individual variables, i.e., our concepts of things and properties themselves. Not just sentential propositions, but any rationally intelligible thing I assigned to an individual variable a must satisfy the requirement that

(2) -(a & -a);

i.e., we must conceive it as self-identical. The holistic regress implies that we can recognize things and properties as self-identical only if we can identify them in terms of higher-order properties that are themselves self-identical.

[10] Of course this does not mean that they *are in fact* logically consistent with one another, just that I *must conceive* them as being so.

necessity. Call this the requirement of *vertical consistency*.[11] I shall say more about vertical consistency in section 2 below.

Finally, I shall refer to the property of being an experience I have as the *self-consciousness property* of things I in fact experience, including both things in the external world and my own intentional states. If I could not recognize each of these things as having the self-consciousness property, I could not conceive any of them as my experiences.[12] To conceive an experience as mine is to conceive it as having the character it has partially *in virtue of my nature*. An agent who lacks the concept of the self-consciousness property lacks the recognition of herself as partially responsible for the character of those experiences. She views things as happening *to* her, but not in any part *from* her. Without an implicit recognition of her collaboration in the character of her experiences, she lacks a necessary condition of being motivated intentionally to alter them, i.e., to act. She therefore lacks a necessary condition for motivationally effective agency - but not just in the ordinary sense of being incapable of intentional physical behavior. She lacks it as well in the more pervasive sense in which we ordinarily conceive ourselves actively to *do* things like think, feel, infer, and search our memories.[13] So if I

---

[11] In standard notation, the requirement of vertical consistency would run roughly as follows. Given an individual variable a to which t is assigned, and terms F and G with the extensions P and P^1 respectively,

$$(3)\ Fa \rightarrow [(x)\ (Fx \rightarrow Gx) \rightarrow Ga]$$

It is important not to confuse the requirement of vertical consistency with a claim about the transitivity of relations among predicates generally: not every predicate is of a higher or lower order than every other predicate. Rather, the requirement of vertical consistency is a transitivity claim about the relation between our concepts of the lower- and higher-order properties of a thing, i.e., those that satisfy (3). I am indebted to Wayne Davis for alerting me to notational errors in an earlier formulation of the requirements of horizontal and vertical consistency.

[12] Of course this does not mean that they would not *be* my experiences, just that I could not thus conceive them.

[13] For these reasons, I do not see how Bernard Williams can claim that "When I think about the world and try to decide the truth about it,… I make statements, or ask questions,… [which]… have first-personal shadows,… [b]ut these are derivative, merely reflexive counterparts to the thoughts that do not mention me. I occur in them, so to speak, only in the role of one who has this thought (*Ethics and the Limits of Philosophy* [Cambridge, Mass.: Harvard Univ. Press, 1985], p. 67). If I did not occur in such statements in the role of one who had this thought, I would be unable to act on any thought I had. So I think Williams is too quick to differentiate the "I" of theoretical

am an agent, each thing that is rationally intelligible to me at a given moment must, as a matter of conceptual necessity, instantiate the highest-order concept of an experience I have. Hence this highest-order concept must also satisfy the requirements of horizontal and vertical consistency.

These five premises jointly imply that if we are successfully to make coherent sense of things, including our own actions, we must, in conceiving those things, satisfy the law of noncontradiction in the ways the requirements of horizontal and vertical consistency specify. This is part of the sense in which the requirements of theoretical reason apply not just to sentential propositions but also, and more fundamentally, to those concepts of their constituents that form an agent's perspective at a particular moment: theoretical rationality is to this extent a necessary condition of agency. But then whether an agent is rational or not cannot depend solely upon contingent factors such as training or personality that some normal human agents have and others lack. An agent who is not theoretically rational in the minimal sense to which satisfaction of the requirements of horizontal and vertical consistency commit us cannot make sense of the world at all.

This may seem to be a very strong thesis. For it implies that, at any given moment, we must conceive the things and properties we experience in such a way as to satisfy the requirements of theoretical reason, whether they do so in fact or not. And this, in turn, suggests that we are unable to detect logical inconsistencies in our experience, or at least are extremely averse to doing so. In particular, this thesis suggests that we cannot conceive *ourselves* at a particular moment as simultaneously desiring contradictory objects, or as simultaneously believing contradictory propositions, even if in fact we do.

If this is true, it means that self-deception is just as inevitable as self-consciousness. For in situations in which we may simultaneously hold such contradictory beliefs, it implies that it is psychologically impossible for us to see this. These are the claims I want to defend in the following sections.

Literal Self-Preservation

2. Suppose, then, that you are in New York, making your way down West Broadway, where anything may happen, and you suddenly encounter - what? It is large, mottled gray, prickly, shapeless, undulating, and it moos at

---

deliberation as necessarily impersonal from the "I" of practical deliberation as necessarily personal. I have tried to show elsewhere ("Moral Theory and Moral Alienation," *Journal of Philosophy* 84, no. 2 [February 1987]:102-118) that impersonality in deliberation is usually a function of psychological factors, not moral or philosophical ones.

you. You have at the disposal of your current perspective certain concepts of higher-order properties that might enable you to recognize this entity - street sculpture, advertising gimmick, genetic mutation, three-martini lunch hallucination, tropical plant, Mayor Koch, etc.; but it is not immediately evident which one would suffice in these circumstances, or if any of them would. It is tempting to think that this is just the sort of case that belies the necessity of the requirement of vertical consistency to rational intelligibility. For in this case, it may seem, you must know at least that you have encountered a gray blob, even though you don't know what higher-order kind of gray blob it is.

But reconsider. If it is unclear which of those higher-order properties now at your disposal would enable you to recognize this entity, any of them might. If it is unclear whether any of them would, none of them might. If none of them did, concepts that would enable you to recognize this entity would not form part of your current perspective. In this case, you could not be said to experience this entity at all. If it is unclear whether any of the concepts constitutive of your current perspective would enable you to recognize this entity or not, then you can in fact neither identify this entity as a kind with which you are already familiar nor differentiate any such kind from it. The recalcitrance of this entity to identification in terms of the properties currently at your cognitive disposal calls into question *all* the concepts that form your current perspective: if they do not clearly fail to identify this entity, neither can they clearly succeed in identifying any other. So if you cannot now ascertain whether this entity is a three-martini lunch hallucination, a tropical plant, or Mayor Koch, you cannot ascertain whether it is a gray blob or not, either; or whether, if it is not, anything else could be.

This conclusion may seem to be too strong. Surely, it might be objected, it does not follow from the fact that you do not know what something is that you therefore do not know what anything is. Indeed it does not. But the preceding narrative does not address the question what or how you *know*, or even what propositional *beliefs* you have, but rather a presupposition of both of those questions. It addresses the question whether, if you cannot successfully *recognize* something you experience in terms of the concepts at your disposal, you can successfully recognize anything else you encounter at the same time; and concludes that the answer is no. If you cannot recognize something in terms of the concepts at your disposal, you cannot identify it as having the properties of which you have those concepts. In this case, propositional beliefs about, and *a fortiori* propositional knowledge of, that thing are impossible.

Again it might be objected that it does not follow from the possibility that your identification of one thing is incorrect that your identification of

everything else is called into question. Indeed it does not; yet again the objection misses the point. The preceding narrative does not address the question whether your identification of something is *correct* or not, or, therefore, the question of the fallibility of your other identifications, but rather a presupposition of both of those questions. It addresses the question whether, if you fail to make something you experience *rationally intelligible* relative to everything else you experience at the same time, you can succeed in making anything else rationally intelligible at that time, despite this one failure. The preceding narrative concludes that the answer is no. If you cannot recognize something as the same as or different from something else, then you cannot identify that second thing relative to it. Hence the question whether you have identified either one of them correctly or not does not arise.

But this may seem unsatisfactory. For it *must* be in some sense possible for us to recognize an unfamiliar thing in terms of its lower-order properties, independently of our ability to identify it in terms of higher-order ones; otherwise how could we ever come to recognize and eventually categorize unfamiliar things at all? The implication would seem to be that if we were truly to adhere to the requirement of vertical consistency, we could never learn anything new. I want to defer addressing this valid objection until later in the discussion, after I have developed a more fine-grained taxonomy of agent's perspectives than that which we now have. At that point I shall suggest that it is, indeed, much more difficult for some agents than for others to learn anything new, about themselves or anything else, even if, like the gray blob on West Broadway, it is staring them in the face.

That you do not experience what is rationally unintelligible to you at all is why it would be a mistake to take the conclusion of the preceding narrative to be that your encounter with a gray blob on West Broadway will necessarily plunge you into madness. The preceding narrative has been intended, rather, to suggest that we have a very deeply ingrained, motivationally effective aversion to rational unintelligibility, because it threatens the rational coherence of the self as having that experience. We have already seen that an agent must, as a matter of conceptual necessity, finally be able to conceive of everything that happens to her consciously as her experience, in order to conceive of herself as capable of altering what happens to her, and so in order to exercise her agency. The preceding narrative shows us that, conversely, an agent who cannot conceive of her experiences in a rationally intelligible form cannot conceive of them as her experience at all, and so similarly lacks the agency necessary to change them. An agent who experiences events that she cannot make rationally intelligible in terms of the concepts that constitute her perspective at a given moment *loses her perspective on those events*: she confuses them with others, and all of them with herself. That is, she confuses all of those events with the rationally intelligible cognitive and conative events that

constitute her perspective at a given moment. But a self that confuses unintelligible external or internal events with itself loses the ability to distinguish those events from itself, and with it the ability to defend its rational integrity against them, and so, finally, the ability to act intentionally in response to them. That is why your most likely initial reaction to encountering a gray blob on West Broadway will be neither madness nor annoyance, but instead temporary cognitive and conative paralysis.

It is this cognitive and conative paralysis, and the loss of unified selfhood and agency it threatens, that motivate us either to render a perceived conceptual anomaly rationally intelligible at any cost - even at the cost of plausibility, accuracy, and truth - or else to suppress the perception altogether. I shall refer to these strategies for preserving selfhood and agency against the threat of disintegration as *pseudorational* strategies. Below I shall try to say in greater detail in what these consist, and why they should be viewed as integral to theoretical rationality.

First consider the objective of these strategies, i.e., the preservation of rationally integrated agency - or, as I shall call it, *literal self-preservation*. As we have seen, literal self-preservation just is the preservation of the rational intelligibility of our experience in the form necessary for agency, i.e. as self-conscious experience. We have also seen that this, in turn, requires that the ways we conceptualize our experiences satisfy the requirements of horizontal and vertical consistency, however else they may differ. These requirements, I have suggested, are the familiar requirements of theoretical reason applied to the substantive and predicative constituents of declarative propositions we occurrently believe. This means that literal self-preservation is, in effect, preservation of theoretical rationality as motivationally overriding in the structure of the self. Theoretical rationality is *motivationally overriding* in that it constrains and is a necessary condition of any other motive an agent may have. For without it, there would be no agent to he motivated to perform any particular action whatsoever. So literal self-preservation must be a biologically fundamental disposition that any such action - and any more particular motivation for it - must presuppose.[14] Since it enables us to preserve

[14] For this reason, it would be a mistake to confuse this disposition with a *desire* for literal self-preservation; the very idea is incoherent. For however we understand the notion of a desire - whether as an occurrent, internal event, or as a "pro-attitude" as do, for example, Donald Davidson ("Actions, Reasons, and Causes," in *Readings in the Theory of Action*, eds. Norman S. Care and Charles Landesman [Bloomington: Indiana Univ. Press, 1968]) and Alvin Goldman (*A Theory of Human Action* [Englewood Cliffs, N.J.: Prentice-Hall, 1970]), or as itself a disposition to experience certain occurrent, internal events under certain circumstances (as do Richard Brandt and Jaegwon Kiln in

the consistency of the highest-order concept of our selves as having our experiences, and so constrains all other motives we as agents can have, I shall describe it as a *highest-order disposition to literal self-preservation*.

That literal self-preservation, i.e., the preservation of the theoretical rationality of the self, is in this sense motivationally overriding in the structure of the self, may help to explain why, when confronted by a conceptual anomaly, we are more inclined either to suppress it from consciousness altogether or to distort or truncate the concepts constitutive of our perspective in order to accommodate it. For the alternative would be passively to acquiesce in the threat of unintelligibility, disorientation, and ego-disintegration that such an anomaly represents. By definition, such an inclination towards biological self-defeat could have no survival value whatsoever.

That literal self-preservation has biological survival value implies, of course, that it has *value*, i.e., that it is, for us, a normative good. But we have just seen that literal self-preservation just is the preservation of the rational

---

"Wants as Explanations of Actions," in Care and Landesman, *Readings in the Theory of Action*), eds., or as a merely theoretical construct (as does David Lewis in "Radical Interpretation," *Philosophical Papers, Volume I* (New York: Oxford Univ. Press, 1983) - a desire must be in any case something that some agent has. So it cannot be the same as that disposition that is required by the continuing existence of that agent to begin with. The disposition to literal self-preservation must be presupposed by any desire an agent has, because it must be presupposed by motivationally effective agency. If it is a necessary presupposition of desire, it cannot be the same as desire.

For the same reason, it would be a mistake to suppose that the preservation of the theoretically rational intelligibility of the self might be a mere means to the satisfaction of some further desire. There can be no doubt that the disposition to literal self-preservation has *at least* instrumental value, since it is a necessary precondition of any of the ends an agent adopts, and so *a fortiori* of those she actively tries to achieve at a given moment. But it also *precludes* the adoption of any ends or beliefs that are themselves inconsistent with the rest of her experiences. For example, consider the hero of Henry James' "The Last of the Valerii," a young Roman count of ancient lineage who unearths a pagan statue on his family's estate. The statue evokes in him the desire to engage in ancient and, to him, completely inexplicable Dionysian rituals. He finds himself compelled to perform these rites nightly from dusk until dawn. Tormented by impulses that, although harmless, are to him completely unintelligible and inconsistent with the other desires and habits that characterize him as a modern European, he has the statue reburied almost immediately, rather than utilize his wealth and freedom to indulge these anomalous impulses. The biologically fundamental disposition to literal self-preservation requires the suppression not only of external but of internal events that violate the requirements of horizontal and vertical consistency, on pain of cognitive and collative paralysis - or, at worst, madness. I have profited from discussing these objections with Louis Loeb and Paul Guyer.

intelligibility of one's experience, i.e., satisfaction of the requirements of horizontal and vertical consistency. This means that what we often refer to as descriptive or explanatory coherence is itself a normative good - one we must achieve to some degree before we can even strive to achieve any other. Now in some ways this may seem to be an impossible task. We are continually assaulted, if not by the presence of gray blobs, by other internal and external experiences that test the psychological strength of the self to withstand them, or its cognitive flexibility to accommodate them within the constraints of rational intelligibility. And we must take it as a given that we can neither withstand all such events - on pain of the fate that frequently befalls ostriches who bury their heads in the sand - nor accommodate all of them - on pain of the fate that befalls overladen computers, whose simulated cognitive psychoses bear a touching resemblance to our own. On the other hand, theoretical reason is all we have for coping with such cognitive assaults. So the requirements of theoretical reason must be systematically attenuated and bent somewhat in order to do so. Its consistency requirements must remain in force, but be made easier to satisfy; the stringency of those requirements must be upheld, yet tempered by rational loopholes. The result is not rational intelligibility in the sense described above, but rather pseudorationality. Let us now examine how the strategies of pseudorationality function.

Pseudorationality

3. The pseudorational strategies I want to target are three in number: denial, dissociation, and rationalization respectively. There are probably others, but I believe these three are primary, for reasons that will become clear in what follows. But briefly, *denial* is our imperfect attempt to satisfy the comprehensive requirement of rational intelligibility, whereas *dissociation* is our imperfect attempt to satisfy the requirement of horizontal consistency, and *rationalization* our imperfect attempt to satisfy the requirement of vertical consistency.

We have already seen how what I shall refer to as *denial* might operate in cases in which the arsenal of concepts constitutive of an agent's perspective is completely inadequate to identify a conceptual anomaly: if one has no concepts even remotely appropriate for coping cognitively with the thing in question, one will simply fail to register that thing as an experience one has. Here the preservation of rational intelligibility, i.e., literal self-preservation, requires that one remain oblivious to the thing's presence.

But now contrast this case with one in which denial is required, not in order to preserve the rational intelligibility of one's experience as such, but rather the rational intelligibility of *a certain interpretation or theory of* one's

experience. The distinction can be limned as follows. I may be able to make sense of everything I experience *as my experience*, without trying or being able to make sense of it as confirming the *theory* that, say, it's a jungle out there or that everything happens for a reason or that I am a serious person, or some more sophisticated theory of human nature or the physical world or myself. In the first case, the rational intelligibility of my experience is a function of its horizontal and vertical consistency relation to the highest-order concept of the self-consciousness property *simpliciter*: all the experiences I have are mutually consistent with one another relative to the concept of their being my experiences. This is the only highest-order concept that unities all of them. I shall describe someone who conceives her experience in this way as a *naïf*.

The naïf lacks what I shall describe as a *personal investment* in any particular theory of her experience. I shall say that an agent A is personally invested in something t if (1) t's existence is a source of personal pleasure, satisfaction, or security to A; (2) t's nonexistence elicits feelings of dejection, deprivation, or anxiety from A; and (3) these feelings are to be explained by A's identification with t. A *identifies with* t if A is disposed to identify t as personally meaningful or valuable to A.[15] Since the naïf lacks a personal investment in any particular theory of her experience, she is less prone to encounter genuine conceptual anomalies. For the only requirement something must meet in order to be rationally intelligible as one of her experiences is that it must be the kind of thing she *can*, in fact, experience. Included among the lower-order concepts that constitute the naïf's perspective are the commonsense observational ones of size, shape, color, etc., we all share. But absent from that perspective are higher-order concepts that qualify and restrict the scope of those observational concepts to any particular theory of the kind of thing one can in fact experience. Since the naïf lacks higher-order pet theories that restrict what qualifies as, say, contemporary art, the possible mutant effects of radioactive fallout, Mayor Koch's attention-getting devices, etc., that a mottled gray, mooing blob on West Broadway might violate, she has less cause to suppress recognition of such a blob than you or I. But we already know, from folklore and history as well as from personal experience, that naïfs, and children, often see many things, not just the emperor's sartorial desolation, that the rest of us systematically overlook.

Contrast the second case, in which I do have a personal investment in some favored theory of my experience.[16] This case is different. For here the

---

[15] I discuss the definition of personal investment at greater length in "Moral Theory and Moral Alienation"' (op. cit.).

[16] Does it matter whether my favored theory of my experience is normative or explanatory? The preceding remarks suggest not. Any powerful explanatory theory

rational intelligibility of my experience is a function of its horizontal and vertical consistency relative to two higher-order concepts, the mutual relation of which may vary. First, there is the concept of things as experiences I have, and second, there is the concept of things as confirming my favored theory of my experience. What is the relation between these two? There are at least three possibilities.

We have already seen (section 2) that the second concept could not dominate the first without violating a necessary condition of agency. Of course this does not mean it cannot dominate the first, period. By an *ideologue*, I shall mean someone who regards her experience as an instantiation of her theory, rather than the other way around. She thus has a sense of mystical inevitability about herself, as an impersonal force in the world that, like other such forces, behaves in the ways her theory predicts. The ideologue may seem to have the concept of the self-consciousness property, in that she recognizes things that happen to her as experiences she has. But in fact this recognition is hollow, because she does not, in so doing, recognize things as happening to her precisely in that form *in virtue of her nature*. Instead, she thinks her experience has the character it does in virtue of the forces, specified in the theory, that determine her nature. And she interprets her own active responses to that experience in similarly impersonal terms, not in terms of personal motivations to alter it. Because the ideologue accepts no responsibility for the particular character of her experience, in fact she does not fully grasp the concept of the self-consciousness property. Hence she abdicates a necessary condition of motivationally effective agency: her thoughts, feelings, and impulses are to her a series of *aha-Erlebnisse*, forced upon her by her situation; and she is, to varying degrees, propelled into action by impersonal internal forces that are beyond her intentional control.

For the ideologue, a conceptual anomaly is intolerable. By threatening the rational intelligibility of her favored theory of her experience, it threatens, so far as she is concerned, not only the rational intelligibility of that experience itself, but thereby the rational intelligibility of the universe and her predestined place in it. Because she regards her own experience as an instance

---

also prescribes a way things are supposed to, i.e., *should* work under ideal conditions, and so contain a normative component. And any full-blooded normative theory also explains a way things *would* work if conditions were, in fact, ideal, and so contains an explanatory component (this point is developed at greater length in my paper "The Meaning of 'Ought' and the Loss of Innocence," and with reference specifically to Kant's Categorical Imperative in my "Kant's Idea of Reason"). Part of what we do by attempting to make things rationally intelligible in the terms given by our favored theory of our experience is to assess the extent to which the real measures up to the ideal - or, to put it in Hegel's infamous terms, the extent to which the actual is rational.

of her theory, rather than the other way around, it is not open to her to rethink her perspective on the world as independent of that world itself. Her perspective is such that she views it as fully determined by that world, in the ways specified by the theory that purportedly describes it. To undermine the theory, then, is to undermine everything at once. For the ideologue, conceptual anomalies do not exist. I shall say more shortly about some more subtle pseudorational mechanisms by which they are made to disappear.

Like the ideologue, the character I shall describe as the *solipsist* also attempts to make all her experiences rationally intelligible, relative both to her favored theory and to the concept of the self-consciousness property. But by contrast with the ideologue, the solipsist reverses the relation between them, for she recognizes her favored theory, and its confirmation by her experiences, as itself an experience she has. So even if her theory that, say, it's a jungle out there or that she is a serious person does, in fact, make all of her experience rationally intelligible, she conceives it as doing so *in virtue of her nature*, i.e., as itself an experience she has: her favored theory is subordinate to the highest-order concept of the self-consciousness property. Because of this order of priorities, the solipsist's investment in any such theory can never be more than tentative, and her attitude toward it never more than pragmatic. If the theory makes sense of what is *already* rationally intelligible as her experience, well and good. If it is undermined by a conceptual anomaly, then it is to be modified or replaced. But this is merely to restate what we already know about solipsists, namely that they are, on the one hand, inclined to skepticism about higher-order explanatory theories, and on the other, fondly attached to the observational data those theories are recruited to explain.

Like the naïf, the solipsist has less trouble with conceptual anomalies than the ideologue. For since she recognizes even her favored theory to have the character it does in virtue of her nature, her personal investment in it cannot be so absolute as to blind her to the possibility of its - and her - limitations. And since she lacks such an overriding personal investment in her favored theory, its modification or eventual replacement by a theory better able to accommodate the existence of gray blobs is more a matter of regret than anxiety or panic. Finally, since her conceptualization of her experience as hers takes priority over her conceptualization of it in terms of any such tentatively held theory, she is, like the naïf, freer to recognize a gray blob simply for what it is.

The figure for whom the relation between the concepts of the self-consciousness property and of her favored theory as confirmed by her experience presents a genuine dilemma is one I shall call the *dogmatist*. Luckily, this dilemma is one the dogmatist is unusually well-equipped to solve. For the dogmatist, the relation between these two concepts is one of uneasy parity: both are of the highest order in the dogmatist's perspective;

neither is subordinate to the other. The dogmatist both conceptualizes all of her experience as hers and conceptualizes it as instantiating her favored theory. The dogmatist would not deny that her experiences have the particular character they have in virtue of her nature. Nor would she deny that they have that character in virtue of the truth of her favored theory. Rather, the dogmatist would congratulate herself on the good fortune of being so constituted that the way she experiences the world is, in fact, the way it is. Thus for the dogmatist, these two concepts are materially equivalent.

The notion of being personally invested in one's favored theory about the world has special poignancy in the case of the dogmatist. For the dogmatist is someone who does derive very great pleasure, satisfaction, and security from her favored theory of her experience (indeed, the dogmatist may feel instinctively that it is the only genuine source of security to be had); and these feelings arc, of course, to be explained by her identification with her favored theory. But notice that the higher-order priority she gives to her favored theory implies her identification with it in an even stronger sense than that required by the definition of personal investment. Her favored theory of her experience is not just personally meaningful or valuable to her; it *is* her at the deepest level of self-identification. For as we have just seen, she assumes that the way she experiences the world is, in fact, the way her theory depicts it; and that this, in turn, is the way it is.

Conceptual anomalies that threaten or undermine the rational intelligibility of the dogmatist's favored theory are correspondingly anxiety producing. For in so doing, they undermine the dogmatist's conception of her own experience, and the rational intelligibility of that experience itself. Thus the dogmatist is like the naïf and the ideologue (and unlike the solipsist), in that all three are made more susceptible to rational self-disintegration by their unqualified attachment to the concepts that constitute their perspectives. But the dogmatist is like the ideologue (but unlike the nail and the solipsist), in that the personal investment of both in favored theories of their experience constricts the scope of their experience, and so brings the threat of rational self-disintegration that much closer. Because the favored theory with which the dogmatist strongly identifies restricts the range of 'concepts by which to make sense of those realities, her perspective on them is correspondingly less open-ended, more rigid, and therefore more fragile. The constriction and fragility of the dogmatist's perspective creates more occasions on which she may encounter conceptual anomalies, to the extent that her favored theory excludes more from its scope of rational intelligibility: modern art, ESP, the inscrutable cultural Other, avant-garde styles of self-presentation, play, astrology, jokes, interpersonal theater, agitprop cultural subversion, and her own delinquent impulses must be either explained (or explained away) by

her theory or eke consigned to conceptual oblivion. It is for the dogmatist, as for the ideologue, then, that the gray blob on West Broadway may present a real problem.

We can now distinguish three circumstances in which denial may be an expected response to the presence of a conceptual anomaly; only the last is, strictly speaking, a *pseudorational* response. First, it may function, as it does for the naïf, to eradicate from consciousness something that is anomalous relative even to the most comprehensive and flexible concept one has, namely the concept of something as an experience one has. Something that is not recognizable in these terms is by definition conceptually inaccessible, and so is not a candidate for rational intelligibility in the first place.

Second, denial may function as it does for the solipsist, who is in a loose sense a member of the scientific community, in that her favored theory of her experience has been tested, confirmed, and consensually validated to some extent by that community (macroscopic determinism might exemplify such a theory). Something that is a conceptual anomaly relative to such a theory still may he an experience one has, but the weight of consensus and scientific method militate against acknowledging it as such. Under these circumstances, the anomaly may be a candidate for rational intelligibility, but the solipsist's skepticism, plus the weight of theoretical reason itself, is against it. Here again, denial is consistent with the requirements of theoretical reason.

Third, denial may function as it does for the dogmatist or the ideologue, whose theories may or may not receive consensual validation, but whose theoretical biases in either case would not survive disinterested critical scrutiny. To determine this here we may, but need not, appeal to rational method. We have a commonsense, lay criterion for distinguishing that which is so obscure or genuinely enigmatic as to be rationally inaccessible, from that which is intersubjectively obvious, namely, *third-person disinterested recognition*. If a third party, similarly equipped both culturally and cognitively, but lacking the dogmatist's personal investment in her favored theory, can make the thing rationally intelligible relative to her own perspective, whereas the dogmatist cannot relative to hers, then the dogmatist's difficulty is not that the thing in question is conceptually anomalous, but that her favored theory is just too restrictive or parochial to accommodate it. In this case, her denial of the thing in order to preserve the rational intelligibility of her theory is a pseudorational strategy.

Because she identifies her experience with her theory, rather than conceiving her experience as subordinate to it, the dogmatist has, in addition to pseudorational denial, cognitive resources for meeting such challenges that the ideologue lacks. We have already seen that because the latter lacks a necessary condition of agency, she lacks the conception of herself as actively *doing* things like thinking, inferring, and searching her memory. This is not to

say that she does not do these things at all; just that she does not conceive herself as doing them. Hence by contrast with the solipsist, the ideologue does not conceive herself as capable of revising or rethinking her favored theory - or, by contrast with the dogmatist, as capable of rearranging it to fit the facts.

The dogmatist has the same cognitive resources for conceptually rearranging things as she had for arranging them in the first place, in order to satisfy the two consistency requirements of rational intelligibility. And she is more highly motivated to do so by the fragility and constriction of her theory and her self-protectiveness toward it. That is, the dogmatist does not just have a biologically fundamental disposition to render her experiences horizontally and vertically consistent, as the rational intelligibility of those experiences requires. In addition, she has a contingent but central *desire* to render her experiences horizontally and vertically consistent, relative to the requirements and constraints of her favored theory of those experiences. And the more parochial her theory, the stronger this desire must be. Thus the requirements of horizontal and vertical consistency afford the dogmatist the option of two more subtle pseudorational strategies, in addition to blanket denial, for dealing with conceptual anomalies. And her natural disposition to satisfy these requirements, together with her personal investment in her favored theory, motivates her to exercise those strategies.

From now on, in discussing these two further pseudorational strategies, shall speak not just about the dogmatist but also about *us*. This is not because I think anyone who is likely to read this essay is purely and simply a dogmatist in the sense in which I have described one. Obviously, the naïf, the ideologue, the solipsist and the dogmatist are all equally caricatures, abstracted from more complex agents whose dispositions and perspectives may change from moment to moment, and who are capable of exhibiting the characteristics of each. But I do think that anyone likely to read this essay probably does have a favored theory of her or his experience, however nascent or inchoate, a theory in which she or he is, to varying degrees, personally invested. So I hope to be analyzing cognitive phenomena that all of us will recognize.

Our disposition to satisfy the requirement of horizontal consistency supplies us with the pseudorational strategy I shall dub *dissociation*. Recall that horizontal consistency requires us to conceive all our experience at a given moment as mutually logically consistent, i.e., as satisfying the law of noncontradiction. Relative to a favored theory of that experience, this is to require, first, that the theory be horizontally consistent, and second, that all our experience be recognizable in the theory's terms. A conceptual anomaly is then by definition anything that defies recognition in these terms. In

dissociation, the anomaly is then identified in terms of the negation of some or all of the concepts that constitute the theory; thus the horizontal consistency of our experience is preserved.

This is one juncture that separates the dogmatist from the solipsist. The solipsist's tentative investment in her theory allows her greater detachment from it, which enables her more easily to rethink or revise it in order to accommodate what appears to be a conceptual anomaly. By contrast, the dogmatist's personal investment and self-identification with her theory makes her reluctant to abdicate or modify it, and inclines her to construe her theory, and therefore the events and phenomena it explains, honorifically, as normative goods. Relative to these, the negation of her theory that a conceptual anomaly represents is to be dismissed not only as intrinsically alien and inscrutable but therefore as *insignificant*, without value, and so unworthy of further attention.

Reconsider, for example, the gray blob on West Broadway. There are, obviously, a variety of ways of making sense of this entity, and we have considered some of them. But it is equally easy to construct a rather arid theory of one's experience in which there is simply no room for such things: a theory, say, in which there are two sexes, three races, a circumscribed set of acceptable roles and relations among them, an equally circumscribed set of acceptable norms of behavior, dress, and creative expression, and a further division of the human race into those who observe these standards and those who do not. Not only gray blobs but much else that is of interest, not just in our contemporary subcultures but in other ones as well, will then fall outside the pale of this theory. Again, someone with a personal investment in such a theory similarly will tend to dissociate such phenomena from the realm of the meaningful and important, and consign them instead to the status of intrinsic and uninteresting conceptual enigma (assuming that these perceived enigmas do not allow their existence to be denied altogether).

Yet a third way of dealing pseudorationally with conceptual anomalies is what I shall describe as *rationalization*, a degenerate form of vertical consistency. Recall that vertical consistency requires us to preserve transitivity from the lower-order concepts by which we identify something to the higher-order ones they imply. Relative to a favored theory of our experience, this is to require, first, that the lower- and higher-order concepts of the theory be vertically consistent, and second, that any experience recognizable in terms of its lower-order concepts instantiate the relevant higher-order ones as well. Now any theory even ostensibly worth its salt must include, among its lower-order concepts, the observational concepts by which we commonsensically interpret our experience - of shape, color, size, and so forth however otherwise parochial that theory may be. But this means that even a parochial theory of one's experience can exclude through its lower-order concepts only

genuine conceptual anomalies, of the kind that might trouble the nail or the solipsist. It cannot exclude gray blobs simply by fiat.

This may explain the valid objection, noted earlier but not addressed, to the case of the gray blob on West Broadway as originally narrated. Surely, we felt, if we have the lower-order concepts of grayness, shapelessness, mooing things, and so forth, we can recognize the thing in question as a gray blob, even if we cannot say what higher-order kind of gray blob it is. Indeed, parochial theories were characterized as precisely those that made into conceptual anomalies things that were well within the range of rational intelligibility from a theoretically disinterested perspective. The need for rationalization arises because the commonsense rational intelligibility of these things at lower conceptual orders puts pressure on the theory's higher-order concepts to accommodate them, on pain of violating the requirement of vertical consistency and so of revealing the conceptual inadequacy of the theory. The dilemma for one who is personally and dogmatically invested in such a theory is that she must accommodate the anomaly without seeming to revise the higher-order concepts of her favored theory; this dilemma is what separates the dogmatist from the solipsist. It is for the dogmatist that rationalization is of greatest use: it is the process by which one stretches, distorts, or truncates the customary scope of instantiation of the higher-order concepts of one's theory, in order to accommodate the recalcitrant phenomenon within the theory's scope of rational intelligibility. More generally, *rationalization* consists in applying a higher-order concept too broadly or too narrowly to something, ignoring or minimizing properties of the thing that do not instantiate this concept, and magnifying properties of it that do.

For example, consider once more the gray blob on West Broadway. Again it is easy to imagine a theory of a particularly self-righteous and sour-minded sort, according to which this blob is, like much else on West Broadway, nothing but one more capitalist plot to poison the minds of the unsuspecting masses and fill the coffers of media devils. The beauty of any favored theory of one's experience is a boon for the personal investor in particularly parochial ones, namely the versatility of its constituent concepts. Pseudorationality, if not genuine rationality, is an available resource for literal self-preservation for even the most dogmatic and narrow-minded among us. For as Humpty Dumpty knew, we are free to use concepts in any way we like.

Self-Deception

4. Now, finally, I shall try to cash out my claim that self-deception is a particular kind of pseudorationality. In particular, I shall argue that self-deception is pseudorationality about a particular kind of theory in which we have a personal investment, namely our personal self-conception.

First I want to show that not all dogmatic pseudorationalizers are self-deceivers. Consider a cult member. A cult member self-identifies with a dogmatic and parochial theory of her experience, a theory in which her degree of personal investment necessitates denial, dissociation, or rationalization of dissonant data in order to preserve the rational intelligibility of her experience. Nevertheless, such an individual might be completely *selfless* in the sense that her pseudorationality is motivated solely by her dogmatic allegiance to the theory and not by considerations of personal vanity or self-esteem. She might, indeed, simultaneously exhibit all the beneficent virtues to a particularly high degree: devotion to others, compassion, generosity, humility, modesty, and so forth; virtues that lead us to deplore all the more their being squandered in the service of the dogmatic theory that deludes her. To call her selfless is not to say she lacks a self, for it is precisely the virtuous characteristics of the self she expresses whose waste we deplore. Rather, it is to say that her self-identification with her favored theory is not itself motivated by self-aggrandizing considerations. While she defends herself by pseudorationally defending her theory, the defense of her theory is not intended to redound to her own greater glory. Conversely, although an assault on her theory is an assault on the rational coherence of her self, she does not perceive such an assault as an *insult* or as denigrating her own value. Her responses to such an assault include anxiety and panic, not rancor or resentment. That the cult member's personal investment in her theory is to he explained by her selfless self-identification with it, *but not* her self-aggrandizement by it, underwrites the intuition that this case is, indeed, most naturally described as a case of *delusion*, not self-deception. To identify it as a case of self-deception would be conceptually peculiar.

The implications are two. First, although all self-deceivers are dogmatic pseudorationalizers, not all dogmatic pseudorationalizers are self-deceivers. The cult member has everything it takes to be a pseudorationalizer, but lacks a certain feature conceptually necessary to being identified as a self-deceiver. Second, therefore, self-deceivers are dogmatic pseudorationalizers of a certain kind: they are pseudorationalizers with a personal investment in a certain *kind* of dogmatic theory, namely one with two mutually dependent parts. The first, explicit part is a dogmatic and parochial theory of their experience, of the sort already discussed. The second part, however, is often left implicit: it is a theory of who they are, how they behave, and how they relate socially to

others. For the self-deceiver, this second part of the theory is the source of the vanity and self-aggrandizement the cult member was shown to lack.

This second part of the theory is not to be confused with the self-consciousness property. The latter is merely the concept of one's self as having one's experiences; the former is a substantive conception of the *kind* of self one is, for example, that one is a serious person. I shall refer to this as the agent's *self-conception*. The agent's self-conception includes the properties she thinks accurately describe her psychologically, socially, and morally, and the more complex principles she thinks govern her behavior and relations with others at a given moment. Any agent may have a self-conception, and not all self-conceptions function as does the self-deceiver's.

A self-conception, the unstated second part of the self-deceiver's theory, is mutually dependent with the first, in that the validity of the first is a necessary and sufficient condition, in the self-deceiver's eyes, of the validity of the second. This is because, typically, the first part, the dogmatic theory of her experience, includes in it honorific status for persons of the kind she conceives herself to be. According to this analysis, then, a self-deceiver is a pseudorationalizer who conceives of herself as a good and valuable person if and only if the dogmatic theory of her experience she espouses is the correct one. Nazis, racists, sexists, anti-Semites, and other elitists of various kinds are all obvious examples of individuals we would identify as (at the very least) self-deceived according to these criteria. But there are many other dogmatic theories of one's experience that may function similarly to align one on the side of the angels, as it were, depending on one's social values. It may be that, held by the right agent, any such theory may, in that agent's eyes, confer on her the exalted status of being holier than thou.[17]

Now one implication of the foregoing characterization of self-deception as a species of pseudorationality is that a certain familiar analysis of self-deception, as believing that not-p because one wants to, even though one knows in some sense that p, is inadequate to the psychological facts. For if the familiar analysis is right, either we must continually vacillate between believing that p and believing that not-p, adjusting our current perspective, favored theory of our experience, and self-conception accordingly in order to preserve horizontal and vertical consistency, which is psychologically

[17] I doubt the difficulty of imagining alternatives to this way of thinking about oneself. For example, one might derive a great deal of self-esteem from being an academic, because one enjoys teaching and research, and believes one can make a valuable social contribution by engaging in them, without thereby supposing that academics, and so oneself, are any more important or valuable in the total scheme of things than janitors or secretaries or postal clerks.

implausible, or else our personal investment in believing that not-p must lead us pseudorationally to deny, dissociate, or rationalize p in order to maintain the belief that not-p. In this case, I would argue, it is not true that we also "in some sense" believe or know that p. For to have any such belief would presuppose the rational intelligibility of p that our pseudorational mechanisms are designed to obliterate.

The second implication of the foregoing characterization is that, even if we could be said "in some sense" to believe or know that p while believing not-p because we want to, as the familiar analysis would have it, this analysis could not in any case provide a sufficient condition of self-deception. For according to the familiar analysis, we would have to identify the cult member as self-deceived, which, as I have suggested, seems conceptually peculiar. In addition, one's desire to believe the falsehood not-p must be, specifically, a desire for self-aggrandizement, to which belief in the falsehood is a means. This is to argue that in addition to deception of the self by the self, self-deception intrinsically involves deception *about* the self that deceives.[18]

Is there any pseudorationality recognizable as self-deception that does *not* involve self-aggrandizement as a motive? I doubt it, but remain open to persuasive counterexamples. Consider two kinds of case, nonpersonal and personal. First the nonpersonal case: suppose I have a personal investment in the theory that it's a jungle out there. Also suppose, for the sake of argument, that this theory is false. My investment in it then may be explained either by the generally oppressive experiences I and most everyone else seem to be having or by the tact that this theory excuses my own failures and moral turpitude. Only in the second case does it make sense to describe me as self-deceived. Now take the personal case. Suppose I have a personal investment in the theory that my spouse is a good person. Again suppose this theory to be false. Again my investment in it may be explained in at least one of two ways: either by my spouse's resourcefulness in maintaining an appearance of virtue and guilelessness, which elicits my love and respect, or by the fact that my recognition of his moral turpitude would reflect negatively on my conception of my own tastes, preferences, and susceptibility to moral corruption. If my spouse is recognizably a bad person, then either I have vicious tastes - say, a fascination with evil - or else I am morally unconcerned by the close proximity of evil. Again it seems to me that only in the latter case does it make sense to describe me as self-deceived.[19] Hence self-deception

---

[18] Also see Amélie O. Rorty, "Belief and Self-Deception," *Inquiry* 15 (1972): 387-410. Rorty has since repudiated this view.

[19] Of course there are further, large questions about whether or not, in the absence of vicious tastes, one can be said to love a person one recognizes as unregenerately had,

does not depend on the nature of the theory in which one has a personal investment but rather on the motive that causes the investment. My claim is that it always involves a desire to buttress another theory, namely an honorific self-conception.

Now I want to consider a case that is identifiable as one of self-deception according to these criteria, and test the capacity of the foregoing analysis to explain it. Take the hero of André Pieyre de Mandiargues's *The Margin*. Sigismond, while on a business trip in Barcelona, has received an ominous letter from the servant of Sergine, his wife. As he begins to open the letter, his eyes alight on these sentences: "She ran to the wind tower. She climbed the spiral staircase. She threw herself from the top. She died right away." He decides not to read the letter just yet, and puts it in a prominent place on his hotel dresser. For the next three days, he drifts through the streets of Barcelona, reveling in its museums, architecture, and unsavory nightlife. Some of his experiences recall to him with disgust his dead father's depravities. Often he finds himself imagining Sergine's sturdily impassive reactions to the situations he encounters, responding as he imagines she would, and reminiscing fondly about episodes in their life together. Every morning he returns to his hotel room, naps, notices the letter, and goes out again. Sometimes he thinks about the letter there in his hotel room while engaged in very different pursuits. His revelry is gradually brought to a halt as his companion of the night deserts him, his pleasures grow stale, and the image of the unopened letter becomes more persistent. Finally he returns to the hotel, and opens and reads the letter, to learn that his only child, Elie, has drowned in an accident, and that Sergine, immediately upon discovering this, has committed suicide. He quits his hotel, drives away from Barcelona, and pulls over to the side of the highway, where he, too, commits suicide by shooting himself in the heart.

Now on the familiar analysis of self-deception, we would be forced to describe Sigismond's state during his three days of revelry and dissipation as one in which he in some sense knew that Sergine had committed suicide, but convinced himself that she had not, because he loved her and did not want her to abandon him, and so both believed (perhaps unconsciously) that she had and believed that she had not. But this just seems completely inadequate to handle the complexity of the case. He may not have wanted her to commit suicide, but surely this desire would ordinarily motivate him to ascertain whether she had or not, and, if so, why. And if he believed she had, why did he spend three days partying in Barcelona before committing suicide himself?

---

and in general about what our commitment to recognizably and incorrigibly morally flawed others consists in. I am indebted to Brian McLaughlin for this example.

I would suggest a different analysis. First, the functioning of the pseudorational mechanisms themselves: the sanguinity of Sigismond's perspective is violated by the intimation of tragic news about his wife, in the form of the letter. He pseudorationally *denies* this intimation, with the help of the distractions and novelties his stay in Barcelona provides. Relative to the fragile and studied innocence of his perspective, he regards the physical presence of the letter on his hotel dresser as a potential threat, which he pseudorationally *dissociates* as an inscrutable, enigmatic object that regularly intrudes on his guilelessness, only to he repeatedly dismissed. The exhaustion of his resources for denial forces him to confront the contents of the letter, in the hope of integrating it into the sanguine perspective he has, with the aid of these pseudorational mechanisms, so tenuously maintained. This proves to be impossible. Sigismond's avoidance of the contents of the letter is not predicated on his unconscious knowledge of its contents, but rather on his cognitive inability to make its contents rationally intelligible relative to the constraints of his perspective. These contents are threatening to him not because he already knows what they are but because he cannot find the conceptual resources for figuring out what they are without violating the dogmatic assumptions in which he is personally invested.

Second, the personal investment that motivates Sigismond's pseudorationality: it is very hard to understand the point of Sigismond's pseudorational behavior without knowing the self-conception its presence threatened. After all, he cares deeply about Sergine; why wouldn't he hasten to find out whether the phrases in the letter actually referred to her, and, if so, what had motivated her suicide? The implication is that it could not have been news of Sergine's suicide alone that he was avoiding. Without reference to his self-conception, it is similarly difficult to understand why the contents of the letter lead him to commit suicide himself. After all, his affection for his son, Elie, was rather distant to begin with; and although Sergine's suicide must be a terrible blow, he obviously is not without resources for containing his loneliness. The implication is that it was not just the combination of his wife's and his son's deaths itself that led him to this end. Without reference to the self-conception in which Sigismond is personally invested, we cannot quite understand why he has been so energetically motivated to deceive himself in the first place.

The description of the case provides evidence for what this self-conception is. We know, for example, that he feels both attracted and repelled by the thought of his own father, and that he does not give a thought to his own son's safety after receiving the letter. We also know that he is, on the one hand, deeply attached to his wife, and on the other, untroubled by occasional, casual betrayals of her. Although his recollections of her include no demonstrative expressions of her love or affection for him, we know that he

assumes that site is attached to him as well, and ignorant or tolerant of these dalliances. We can say, then, that he has a deep personal investment in the conception of himself as Sergine's beloved and of their bond as intimate, loving, and durable, and that he views his extramarital activities as unproblematic and is untroubled by Sergine's likely reactions to them. We also know that he feels some distaste for, or at least detachment from, the role of father, and is emotionally indifferent toward his son.

That this self-conception is pseudorational is suggested, first, by the distance and impassivity of Sergine's responses as Sigismond has recalled them. They do not provide evidence of her emotional attachment to him at all. His assumption that she does love him is sustained by *rationalization*, by misconceiving her imperviousness as itself the way she expresses her love for him. This rationalization enables him falsely to assume that she loves him, because she does not correct it by telling him explicitly that she does not.

Second, the pseudorationality of Sigismond's self-conception is evinced by Sergine's having committed suicide immediately upon Elie's death. For the implication is clear that without her son, Sergine's life is no longer worth living; and her husband, despite his attentions to her, does not make it so. Sergine's suicide nullifies by a single act the importance of his commitment to her as he conceived it, and thereby his value and importance in his own eyes. It is not simply the combination of her suicide and his son's death that drives Sigismond to suicide, but the now-inescapable realization that he meant so little to her that his love provided her with no consolation or further reason to live. In demonstrating through her suicide that he provided *her* with no reason to live, Sergine has taken away *his* reason to live. Sigismond is goaded to suicide by the realization that his self-conception as the valued and beloved object of her devotion was false. This is the truth that he went to such lengths to avoid, that Sergine's suicide makes inescapable, and that makes his own suicide inescapable as well.

What makes Sigismond a self-deceiver, then, is not just that he manages to avoid unpleasant truths because he prefers not to know them, as the familiar analysis would have it. What makes him a self-deceiver is his self-aggrandizing self-conception, sustained by denial, dissociation, and rationalization: by a studied obliviousness to the conclusive, tragic evidence of his wife's indifference; by dissociation of the letter that contains it; and by rationalization of the earlier unresponsiveness to him that otherwise would have indicated it. His personal investment in his pseudorational self-conception is self-deceptive because it enables him to avoid recognition of who he really is.

But why is it in general so important for the self-deceiver to avoid self-knowledge? I would suggest that this is to be explained, quite simply, by the

self-deceiver's personal investment in her self-conception, in conjunction with the disparity between that self-conception and what the pseudorationalized evidence in fact suggests to be a less exalted truth. Earlier I suggested that our highest-order disposition of literal self-preservation made the horizontal and vertical consistency of our favored theory of our experience tantamount to a normative good, and disposed us to ascribe to it, and to the things it explains, an almost honorific status. I also argued that a particularly fragile or parochial theory elicits an even more intensely self-protective desire to preserve it, proportional to one's personal investment in it. For these reasons, the self-deceiver is particularly recalcitrant and impervious to any attempts of her own to survey and critically revise her own pseudorational self-conception. Her investment in it is too great, and increases not only with its fragility, but with the bogus value it confers on her. I think that this is why the project of convincing a self-deceiver that she is self-deceived often seems such an exasperating and futile one: the self-deceiver has not only the rational intelligibility of her experience, but her self-conception as a valuable person, to protect.

But the same vigilance and self-protectiveness that leads the self-deceiver so strenuously to avoid self-knowledge leads her to value it all the more. For of course her pseudorational self-conception would become a source of intense humiliation to her if it were revealed to be false. The revelation that one is not as nice, smart, or popular as one thought is a shaming experience in which one's deficiencies are exposed to the ridicule of the cruelest and most unsympathetic spectator of all. To avoid this revelation, one must be very humble on principle, like Uriah Heep, very vigilant, like St. Augustine, or, like the self-deceiver, very resourceful in one's commitment to truth. As Sigismond's case suggests, self-deception, and pseudorationality more generally, requires energy, perseverance, an inquiring mind, a good grasp of the data, and a deep desire for epistemic rectitude. In order to avoid the humiliation of self-discovery, the self-deceiver needs not only to excise the damaging evidence that portends it but also to believe that the pseudorational mechanisms by which she does so themselves rather bespeak her honesty, sincerity, and perspicacity. Thus may self-reflection and a commitment to truth supply a disguise for pseudorationality for the self-deceiver. Her pseudorational self-conception, then, provides not only a source of bogus value for the self-deceiver, but the illusion of a limited but impregnable scope of personal infallibility that enhances it. This is what I meant when I suggested, at the beginning of this discussion, that the self-deceiver would rather be right than rational.

Now against such self-deception, as well as other forms of pseudorationality, philosophers of a Humean persuasion, such as Henry Sidgwick, John Rawls, Richard Brandt, Stephen Darwall, and of course, David

Hume himself[20] have urged a palliative, i.e., vivid reflection on the relevant data in a calm and composed setting. But if the mechanisms of pseudorationality function as I have suggested, the Humean palliative may in many cases amount to little more than ineffectual bootstrap-pulling. For the whole point of exercising our pseudorational resources is to *restrict* what counts as relevant data to the psychologically and theoretically palatable. If the self-deceiver, and the pseudorational agent more generally, had appropriate conceptual access to these data in the first place, vivid reflection on them would he unnecessary. For the self-deceiver, vivid reflection on the relevant data is an occasion for pseudorationality, not an antidote to it.

What hope is there, then, for the self-deceiver - and, indeed for us all - to avoid or ameliorate self-deception, if reflective self-scrutiny is ineffective? I shall close this discussion by advocating a thoroughgoing policy of what I shall call *epistemic audacity*. By this I mean, simply, having the courage of one's convictions; being willing to test one's favored theory of one's experience more generally, as well as of oneself, against circumstances or aspects of one's own behavior that one perceives as challenging or threatening it. For we do, at least, have conceptual access to these observational data. We are all familiar with the sinking of the stomach, increased heart-rate, or tightening of the throat that motivates us to ignore such behavior, or turn away from such circumstances, or dismiss them summarily as unimportant or without value, or explain them uneasily in familiar terms that nevertheless do not seem entirely to fit. Perhaps *these* are the data that genuinely deserve our reflection, more than any peculiar to the circumstances in question: the anxiety and discomfort that accompany intimations of our confusion, fallibility, or inadequacy. The suggestion is that most of us can stand much greater doses of these feelings than we may think, and might be better off in the end for doing so. The real threat, of course, is the cognitive and conative paralysis, or self-disintegration, or madness broached earlier. But a little madness is not necessarily a dangerous thing, if it forces us to rethink and restructure the dogmatic theory that crippled our vision in the first place.

---

[20] See David Hume, *A Treatise of Human Nature*, ed. L. A. Selby-Bigge (Oxford: Clarendon Press, 1968), bk. 3, sec. 3, p. 603 (and my "Hume on Rational Final Ends" for a systematic discussion); Henry Sidgwick, *The Methods of Ethics* (New York: Dover, 1966), Bk. I, chap. 1, pp. 13-14, and chap. 8, p. 101, passim; John Rawls, *A Theory of Justice* (Cambridge, Mass.: Harvard Univ. Press, 1971), chap. 7, sec. 64, p. 417; Richard Brandt, *A Theory of the Good and the Right* (New York: Oxford Univ. Press, 1979), chap. 1.1, pp. 11-13, chap. 4, pp. 111-113: and Stephen Darwall, *Impartial Reason* (Ithaca: Cornell Univ. Press, 1983), chap. 8, pp. 85-86, 91-93.