Amelie Oksenberg-Rorty

The Deceptive Self: Liars and Layers

Abstract: This paper gives an account of the picture of the self that saves the phenomena of self-deception. On one theory of the self, the phenomena of selfdeception are incoherent: the self as a unified critically reflective rational inquirer cannot deceive itself. On another theory of the self, the phenomena evaporate: the self as a loosely organized system composed of relatively independent subsystems can be conflicted, mistaken, ignorant, compartmentalized. But it does not deceive itself. Our practices as moral agents and rational inquirers are explained by the first theory; our capacities as adaptive survivors are explained by the second. Neither picture can be reduced to the other; neither can be abandoned. The phenomena of selfdeception appear - and are saved - by the superimposition of the two theories.

If any one is ever self-deceived, Dr. Laetitia Androvna is that person. A specialist in the diagnosis of cancer, whose fascinaction for the obscure does not blind her to the obvious, she misdescribes and ignores symptoms that the most junior premediaal student would recognize as the unmistakable symptoms of the late stages of a presently incurable form of cancer. Normally introspective, given to consulting friends on important matters, she uncharacteristically deflects their questions, their attempts to discuss her condition. But again uncharacteristically, she is bringing her practical and financial affairs in order, and though young and by no means affluent, she is drawing up a detailed will. Never a serious correspondent, reticent about matters of affection, she has taken to writing effusive letters to distant friends and relatives, intimating farewells, and urging them to visit her soon. Let us suppose that none of this uncharacteristic behavior is deliberately deceptive: she has not adopted a policy of stoic silence to spare her friends. On the surface of it, as far as she knows, she is hiding nothing. Of course her critical condition may explain the surfacing of submerged aspects of her personality. Self-deception is not always the best explanation of cases of this sort: sometimes people do undergo dramatic changes, changes whose details have a straightforward explanation. But let suppose that Laetitia Androvna's case is not like that. The best explanations of the specific changes in her character or her

Analyse & Kritik 7 (1985), S. 141–161

behavior require supposing that she has, on some level and in some sense, recognized her condition.

Laetitia need not be lying to herself, need not assert what she also believes to be false. She can mislead herself, blind herself, distort or misrepresent her actions, attitudes, perceptions, moods, tastes, without forming a belief in propositional form. And, most effectively, she can focus her attention in ways that subvert what she takes to be her primary attitudes. Without forming a belief about what she is ignoring, she may be avoiding paying attention to what is obvious; but because the avoidance is persistently patterned, she must have at least recognized and scanned the domain in order to determine not to look further. Although the phenomena are much richer, I shall focus on cases of straightout denials of attributable beliefs because these are the hardest cases of self-deception, where the phenomena are most difficult to preserve and to explain.[1]

As is the way with other forms of deception, self-deception multiplies. Not only is Laetitia deceiving herself about her cancer: to maintain this deception she is also deceiving herself about her self-deceptive moves, the significance of her uncharacteristic focusing, deflections, denials. If self-deception involves more than being mistaken or conflicted, it seems (on the face of it) to require some second order attitudes as well: some recognition of the conflict and some ad hoc unprincipled strategies to reconcile the conflict among first order beliefs. If the charge of self-deception is to hold, these second order attitudes should not themselves just be erroneous or conflicted, nor should they be trumped-up ad hoc rationalizing principles.[2]

An account of self-deception should preserve the following, apparantly conflicting intuitions:

1. Self-deception is a disease that only the presumptively strong can suffer. Only those who claim a relatively high degree of integration and self-knowledge, can be self-deceived. The rest of us are ignorant, confused, conflicted, simply mistaken, in situations where we might normally have been expected to be informed, aware, integrated. Like attributions of voluntary action, attributions of self-deception stand on the stilts of presumptions about what is normally within the domain of attention and awareness.

2. On some level, we are often aware of ourselves as deceiving ourselves. As well as being aware of the material which we are denying, - at least enough to perform the patterned deflection of attention -, we are sometimes aware of ourselves as manipulating ourselves in much the same way that we are aware of - and puzzled by - voluntarily acting against what we take to be our judgment about what is best, all things considered, to

do. Very often this uptake on one's own self-deception occurs only slightly after the event, as quick as the blinking of an eye, but with just that time lapse, the blinking of an eye. Often of course it also sometimes occurs much later. But besides discovering some pattern of behavior of which one had not been aware, one also senses that one had somehow been aware of oneself as up to a trick, even though one had not focused on it at the time.

3. Self-deception is characteristically motivated by some form of self-protection, which is not necessarily focused on the self-deceiver, but may be directed to something of close concern to her. While self-deception need not be deception directly for the self, as well as being of the self by the self, still it is standardly in aid of centrally prized concerns. But this casts as wide net, because virtually all non-habitual action is formed by concerns prized by the self, and because those concerns roam freely, sometimes being symbolically masked, so that a relatively trivial matter carries the psychological attitudes that attach to a central concern.

4. Self-deception is a by-product of a set of capacities and abilities that can serve us well: strong compartmentalization, strong focusing and a capacity to ignore distractions, attending to what is significant or important can serve the strategies of self-improvement. If self-deception is distinguishable from compartmentalization, it presupposes that there is a difference between principled focusing formed by rationally grounded general rules on the one hand, and unprincipled, rationally ungrounded but often serviceable and benign blurring on the other.

5. When self-deception is motivated, and when the motivation reveals a flaw of character or of mind for which the agent is responsible, the deception, as well as the flaw is blame-worthy, particularly when it allowes a cruel or vicious person to retain her moral selfrespect. Since self-deception standardly also involves deceiving others, it is insulting. If a person can justifiably be blamed for deceiving herself, self-deception is in some sense voluntary (cf. Rorty 1983).

6. Yet not each and every individual case of self-deception is directly motivated. Like deception, self-deception can generalize and become habitual. Sometimes self-deception is secondary: it can be required to bolster and protect some focal piece of self-deception. (Often it is the scaffolding of supportive, rather than the primary selfprotective moves of self-deception, that seems so irrational, gratuitous.) Sometimes self-deception is a remnant of a habit that remains active in a person's repertoire even after the disappearance of its motivational origins; sometimes it is a by-product of constitionally structured focusing on what is salient. There is also deflected or symbolized self-deception: an apparently unmotivated case can be derived from its motivational sources by an anomalous associative chain.

The phenomena of self-deception seem, on the face of it, paradoxical. How can a rational person deliberately lie to herself? Of course she can contradict herself: but how can she in good faith both acknowledge and deny that contradiction to herself? Those who deny that there are bona fide cases of self-deception need not deny the phenomena of deflected focusing, inconsequent beliefs and actions, problematic lack of self knowledge. The denial of self-deception is a denial that such irrationality involves a person deceiving herself, where the deceiver and the deceived are identical, and the person intentionally accepts what she recognizes is a contradiction or misrepresentation. Those who deny self-deception argue that its attribution is incoherent: a person cannot intentionally believe what she thinks is false, nor can she intentionally be both aware of her beliefs and not aware of them.

But self-deception is incoherent only if one accepts a certain picture of the self. According to a picture that makes the self rationally integrated, automatically scanning and correcting all its beliefs, there can be no cases of strict self-deception. An alternative picture of the self, as a system composed of relatively autonomous subsystems initially seems hospitable to the possibility of self-deception. The second picture demystifies and naturalizes self-deception, and even to some extent explains it, by characterizing the self as a complexly divided entity for whom rational integration is a task and an ideal rather than a starting point. Yet despite its initial hospitality, the second picture also undermines the possibility of strict self-deception, the identity of the deceiver and the deceived. Nevertheless despite the fact that self-deception seems to disappear, we do, and should, take the phenomena of self-deception seriously. After examining the rationale of both pictures of the self, I want to save the phenomena of self-deception. Our ordinary practices presuppose that the self is both a rational integrator and that it is composed of relatively independent subsystems. The classical description of strict self-deception arises from the superimposition of two conceptions of the self. Because the two seem irreconcilable, and because neither can be abandoned, we remain haunted by the sense that despite its evaporation under close scrutiny, strict self-deception does exist.

1. A true self cannot deceive itself

The picture that makes self-deception incoherent is that of the self as essentially unified or at least strongly integrated, capable of critical truth-oriented reflection, with its various functions in principle accessible to, and corrigible by, one another. In this classical picture, the self is oriented to truth, or at least directed by principles of corrigibility that do not intentionally preserve error. The self is of course engaged in many other activities besides those of amassing or attending truths or even minimizing the possibility of error. We are busy avoiding or creating

trouble, worrying about and enjoying our friends and relations, running for office or at least keeping the scoundrels out of office, being dazzled by the seasons and by paintings of the seasons, enhancing and ornamenting our world and our selves, wooing, engaging in, and trying to avoid hierarchical squabbles, despairing of our lives. But the successful exercise of all these activities is made possible by our capacities for critical reflective rationality. And indeed on this first picture these various activities have their sense and their point only if the self is a relatively tightly organized system, whose experiences are registered in cognitions that can be evaluated for their truth and for their rationality.

The attraction of this picture is that it makes sense of our claims to be responsible inquirers and believers, capable of systematic critical thought, evaluating our beliefs by evidence and argument and being effectively guided by those evaluations. Such critical evaluations are presumably not merely items in an intellectual autobiography. They stand in a special relation to the formation of beliefs, in principle directly efficacious in determining them. If beliefs form a coherent and rationally grounded system, the self - or at any rate the mind - is so constructed that it can form a coherent system. But if the systematic structure of beliefs also allows for reflexive corrigibility, then rational reflection is not just another item in the sequence of thought: it stands in a special judicial, executive and legislative relation to beliefs and desires. Yet the capacities exercised in critical reflection do not threaten the unity of the self, even though they are themselves logically independent of the beliefs and desires which they take as their objects. Because the level of critical reflection is effectively integrated with the sequence of thoughts which it reflects, a unified rational being cannot lie to itself. Apparent self-deception always turns out on closer examination, to be a case of ignorance or error; or it involves change of belief, or careless and unregulated judgement, or unrecognized conflict. Of course a complex unified self may suffer all these debilities: but it is in principle capable of being aware of its disorders, and in principle capable of at least moving to their correction, if only by acknowledging them and identifying with the capacities exercised in recognition.

Responsible believers can regionalize criteria for validity, without automatically demoting themselves to erratic believers. They can adopt and justify contextually distinctive criteria or thresholds for validity, propriety of evidence. But such justifications must themselves be subject to the conditions of critical rationality. Even when it is inconvenient, we consider ourselves obliged to dispell contradictions, if only by acknowledging them, rationalizing them, or agonizing over them.

Suppose Laetitia denies that she has cancer, and at the same time speaks and acts in such a way that makes sense only if she believes she is dying.

So far, there is nothing mysterious. People often accept conflicting beliefs without realizing they have done so, especially when the conditions of opacity are strong, when it would take unusual acuity to recognize the conflict. The question of whether Laetitia is deceiving herself is then a question of whether she is aware of the conflict, or whether her ignoring it is itself intentional. If she is aware of the conflict and acknowledges it, she need not be self-deceived, providing she follows her principled strategy for suspending or reconciling her beliefs. It would be enough for confess, helplessly, that she is unable to dispell either of her beliefs. (But she could be self-deceived in this very confession.)

Even if she does nothing about the conflict, not even artlessly confessing it, but has a relatively long standing well-founded policy that rationalizes her epistemic inactivity, she need not be self-deceived. It is not irrational or contradictory for thoughts to be guided by fears and wishes, particularly when there is a well-articulated principle that defines acceptible conditions and contexts for doing so. Aware of her tendency to hypochondria, and knowing that as a diagnostic physician she would be ablaze with constant fears of illness, Laetitia might have adopted a general policy of attempting to ignore or at any rate, avoid monitoring, her physical condition. In that case, her failure to acknowledge her symptoms, even her denial of their import, is not in itself irrational or self-deceptive. Nor is it merely a convenient ad hoc policy. Although the policy that rationalizes putative cases of self-deception must be rationally well grounded by the person's standards, it need not be a sound or wise policy for it to be a reasonable one, for that agent.

On this picture, charges of incoherence or irrationality are deflected by attributing rationalizing strategies: the person adopts no intellectual strategy that she has not underwritten. So whatever is done intentionally, is done for a reason that rationalizes it, in the light of the person's other attitudes. The picture of a dominantly truth-oriented self need not be a picture of a self incapable of error, blindness or conflict. It is rather a picture of the self as having or being a structure which gives reflexive rational corrigibility a central regulative and dominant function, taking precedence over other intellectual and psychological goods: richness of associational consequences, a capacity for amazement, joy, reverence, irony, intensity.

The picture of a critical rationality sets a number of conditions on the self. It is 1) a simple <u>unity</u> dominated by rationality; 2) <u>transparent</u>, in that its states are accessible to one another or to a central panoptical scanner; 3) <u>oriented to truthfulness</u>, in such a way that its transparency is organized to maximize truths or at least minimize error; 4) <u>reflexive</u> in that the criteria for rationality can themselves be subject to critical evaluation. But in attempting to explain patterned irrationality - a deviant

disposition to resist the correction of errors or the resolution of conflicts, accompanied by increasingly ad hoc, specialized and improbable rationalizing strategies -, each of these requirements becomes weakened: integration replaces unity; systematic connectedness replaces transparency; rationalizing principles replace truthfulness; the condition of reflexivity becomes a regulative ideal.

1) The unity requirement has increasingly weakened conditions for unification:

a) The self is a simple unity, with access to all its psychological states. This panoptical center is also a judicial and to some extent a legislative center, capable of evaluating beliefs and forming its jugdments accordingly.

b) In the absence of a central scanner, the self is taxonomically organized, so that its mutually accessible, mutually supportive psychological states form an unconflicted system designed to maximize truth.

c) In the absence of a single taxonomic organization, subsystems are harmoniously related in a cooperative way, so that their independent functions do not, in principle, conflict. Should conflicts arise, there is a hierarchical procedure for their resolution.

d) In the absence of a procedure for resolving conflicts, the subsystems are designed so that localized conflict-resolving procedures automatically go into operation when conflict arises.

e) In the absence of such localized conflict-resolving mechanisms, the system is so constructed that it is not destroyed by conflict. It can operate and survive at lower levels of efficiency, energy or directions, because its various functions are either replucable or substitutable when the integrative processes are damaged or depleted.

2) The transparency requirement has increasingly weakened conditions for accessibility:

a) There is a central panoptical scanner which has direct and immediate access to all psychological states.

b) There is a central panoptical scanner which can in principle initiate a process that has access to any psychological state. But sometimes this method is i) mediated (an intervening process is required for access) and ii) indirect (the content and character of at least some psychological states can only be inferred).

c) In the absence of a central panoptical scanner, psychological states have relevant access to one another by a relatively automatic natural process that allows for an appropriate modification of any state.

3) The truth orientation requirement has increasingly weakened conditions for truthfulness:

a) All other ends and purposes are formed by, and directed towards the formation of true beliefs.

b) All other ends and purposes can be formulated in propositional terms, and the weight given to them is a function of their service to truth-orientation. Not only can all psychological operations be represented cognitively, and in principle propositionally, but they in fact function in their cognitive-propositionalized forms. It is as propositionally formulated reasons that they operate as causes.

c) All psychological operations can be put into cognitive form, which can in principle be propositionalized, in such a way that they can be assessed for their truth and validity.

4) The reflexivity requirement has increasingly weakened conditions for reflexivity.

a) The content and sequence of thought has a reason, which rationalizes it, according to the person's principles in the light of her other attitudes. All general strategies are rationally adopted, and rationally modifiable. Every critical self-assessment is in principle sufficient to produce an appropriate modification in thought.

b) Every thought can be critically assessed for its truth and for the appropriateness of the categorial assumptions implicit in its formation.

Although the first picture of the self is openly normative and reconstructive, it has descriptive and explanatory power: it accounts for the strongly integrative and integrating character of beliefs and attitudes, the intolerance of contradictions and even of strong conflict. But it is radically incomplete as a descriptive theory. The explanation of error, conflict, irrationality, requires psychological causes to fall outside the system of supportive reasons, to a co-existing interferring system. But part of the difficulty is that at least some of the deflecting beliefs and attitudes can sometimes also function as rationalizing causes. So for instance, if Laetitia's denial of her cancer is not rationalized by her policy to avoid the occasions for hypochondria, those denials are caused by a set of attitudes which could in principle at rate also come to serve within the system of her rationalized beliefs. The first theory deals with the problems of this

duplicity of functions - if one may so call it - by separating them, by distinguishing the belief that rationalizes from the belief that causes without rationalizing. So standardly, the apparent paradoxes of self-deception are solved by distinguishing the conflicting beliefs in such a way that they are not longer properly contradictory. It is not the same belief which is affirmed and denied.

Self-deception, along with akrasia and the irrational conservation of emotions, presents problems for the first picture of the self. In attempting to deal with the phenomena of intractible patterns of irrationality, the first picture gradually drifts towards the second. The stronger versions of the unity, transparency, truth-bound reflexivity requirements drift towards the weaker versions; irrationality is rationalized by the ascription of a harmonizing motive, which allows for compartmentilization; the first picture becomes a regulative ideal. When epistemological and moral responsibility are regionalized and relativized - when the capacities for integration, connectedness, truth orientation and reflexivity are attributed regionally and in degrees -, the background assumptions and expectations required for self-deception appear to evaporate. If the integration presupposed by the charge of self-deception becomes relativised and regionalized, then self-deception can be coordinated with those variations: Laetitia would be self-deceived only in those regions where, and to the extent that, she is integrated, truthbound, reflexive. But to what degree must she be integrated, truth-oriented, and reflexive in a region in order to be self-deceived in that region, rather than conflicted, mistaken, or ignorant? This is not an epistemological problem, not a problem of attribution: it is a problem of characterising the structure of the sort of self that is capable of deceiving itself.

2. Nothing is as brilliantly adaptive as selective stupidity. Self-deception is the best cure for melancholia

We would not have survived as the creatures we are if the first picture presented an exhaustive description, if our sole capacities were those of unified transparent critical inquirers. We would not have managed even if the first picture were our central regulative ideal, dominant over all other ideals. The second picture of the self - the complex survival picture - is generally constructed to explain our adaptive strategies rather than our capacities as responsible believers and agents. On this picture, the mind is not a unified, but "rather a problematically yoked-together bundle of partly autonomous systems. All parts of the mind are not equally accessible to each other at all times" (Dennett 1985). The loosening of the integrative bonds of the self moves to its psychological conclusion: relatively independent but integrable subsystems sometimes fail to communicate. Some integrative strategies are local, others generalized and centralized. When there is overlap and replication of functions, there can be tension among the various integrative strategies.

Fragile creatures who survive in highly differentiated, changing environments must be able to discriminate between subtly different sorts of dangers and opportunities without being too sensitive to adaptationally irrelevant changes. Though for some purposes a central panoptical monitor is adaptive, we are well served by autonomous and automatically triggered subsystems. Survival is served by psychological and physical plasticity, replication and specialized differentiation. Plasticity and replication allow substitutability of functions in cases of damage; diversification and differentiation allow for relatively automatic unmonitored highly specific responses. Capacities and habits that can sometimes undermine strongly integrated rationality are highly useful: compartmentalization and detachment, regionalized and variable thresholds of sensitivity, vagueness of the kind used in self-directed rhetorical persuasion. A certain kind of selective insensitivity, blind persistence, an unresponsive stupidity, creative denseness have enormous benefits. The more sensitive the creature, the more highmindedly rational, the more vulnerable it is to disorientation and debility by attack at one central point. (A creature whose functioning depends solely on critical rationality, a vulnerable critical rationality impaired by lack of sleep, let alone by the flu, is well served regionally specific, automatically activated habits.)

What are the attractions of being a creature capable of being self-deceived? The structures and capacities that permit self-deception as a tangential consequence enable us to manipulate ourselves in situations of indeterminacy. In the interests of generating a self-fulfilling policy, we intentionally and even deliberately mislead. We can speak to ourselves as the friendly neighborhood demagogue, cannily conning ourselves to believing that we can do things that are only distantly or marginally within our repertoire. Because confident belief can be a crucial causal factor in enabling us to acquire those competence, it is useful to trade on the ambiguous forms of speech acts, rhetorically eliciting beneficial actions and responses in contexts where truth is deliberately though perhaps temporarily held in abeyance or set aside (cf. Fingarette 1969; Rorty 1980; Pears 1984). Trading on the fact that declarative sentences normally assert beliefs, we use them to induce beliefs: on one reading, such sentences express vague intentions, perform non-truth-functional rituals. On another reading, they assert presumptively true beliefs. It is by playing between the ambiguity of these two readings, that we can induce the beliefs that serve us well. "You can do it, you can speak German", we tell the world at large, intending ourselves as part of the audience, "I can do it". Since "can" means "it is logically possible that ...", such possibilities are cheap; one can accumulate them wholesale, storing them in the warehouses of the mind. Trading on the triviality of claims about possibilities is one way of generating the confidence required to move towards what is only remotely in one's repertoire. Believing that I can speak German is a good way of loosening my tongue to speak when bumbling into pre-proto-crypto-German

is one of the royal roads to German. If we were careful to avoid deceptive manipulative strategies we would be restricted, unable to act energetically and loyally beyond our initial means. Strictly starkly truthful people are often chained, self-chained. (Yet this capacity to be one's own rhetorician also permits the immoralist to convince herself that she is noble.)

Devoting energies to many of our projects often involves a careful shift of perspectives, a refusal to see matters sub specie aeternitatis, ignoring the relative unimportance of our various enterprises and projects, shelving doubts and hesitations, setting aside the larger corrective perspective that makes focused attention enlarge what is directly present. Sometimes this is done by compartmentalizing, focusing so sharply that what is on the fringes of attention becomes blurred, out of sight or irrelevant. We sometimes do this with the nagging knowledge that if we did pay attention to the larger context in which our enterprises and projects occur, their importance to us would be diminished. Writing philosophy papers, devoting ourselves to political causes, taking our students seriously do not, of course, require self-deceptive manipulation. But it helps. Self-deception is an effective, if temporary, cure for melancholia. Of course such a person might adopt a general principle about the propriety of such focused attention that makes such manoeuvres nondeceptive strategies that are presumptively rational, rather than rationalized by an ad hoc policy of turning a blind eye. In such cases, selfmanipulative focusing or compartmentalizing blinding no longer involves being selfdeceived. But generally, the habits of compartmentalization and conveniently useful deceptive focusing are stronger than the principles that would correct them. Even when we know better, we retain inapropriate habits of thought, hiding the behavioral traces that reveal carefully preserved myopia, the consistently canny averted gaze. There is no irrationality or paradox in accepting a policy of causing oneself to forget what one judges maladaptive or unwise to remember. What is difficult is accounting for the simultaneous noticing and not-noticing that applying such a policy seems to require: enough scanning is required to recognize that one should not keep looking. This is just the sort of activity that gives hospitality to self-deception.

The beneficial inertia of belief in the face of counterevidence also opens the possibility of self-deception. The utility of conservation does not, of course, lie in the particular case, but in the general practise of resisting (over) sensitive criteria for the revision and modification beliefs. Latitudinarian believers develop strongly entrenched intellectual habits that work best when they operate relatively automatically and unreflectively. Susceptibility to self-deception is the unintended but predictible cost of the benefits of such psychological strategies. Of course in principle, a person can tell where decent compartmentalization leaves off and indecent self-deception begins, and so in principle, self-deception can be avoided. But when the cost of constant alert scanning is greater than the benefits

of the unmonitored application of latitudinarian principles, there is a natural slide from sensible strategies to dangerous self-deception. Given a choice between being able to exercise those psychological strategies and being prey to self-deception, it is not unreasonable to retain the capacities and strategies. Adopting a policy that has self-deception as a predictible unintended consequence does not mean adopting a self-referentially absurd general policy legitimizing self-deception.

The second picture of the mind - the picture that captures the beneficial functions of a self composed of relatively independent subsystems - need not deny the phenomena of critical reflexive rationality: it is after all, also highly servicable. Latitudinarian policies that allow for compartmentalizing and self-mesmerizing capacities are compatible with a strongly centralized or integrated self. Indeed such policies would be adopted on rational grounds. While the second picture need not deny the importance of critical rationality, it does deny the dominating centrality of the capacities that are exercised in truth-orientation. As is the way with the political analogue of this model, the organization that maximizes general adaptability and safety appears to increase the possibility of internal conflict and dissociation: not all of the various capacities and strategies directed to survival are easily harmonized, let alone coordinated with or dominated by rationality. The advantages of local autonomy in a bureaucracy that also has a central monitoring allows conflicts among subsystems. There is a trade-off point between the impoverishment and vulnerability that attend centralized control and the erratic unweildiness of a system composed of relatively independent and plastic subsystems. What is wanted is an appropriate rather than maximal level of internal integration.

We have concentrated on describing the advantages of creatures composed of relatively independent subsystems. But we have not yet characterized the second picture of the self in any detail. The second picture has, roughly speaking, two versions:

1) The self is analyzed into relatively complete homuncular subsystems.[3]

2) The self is subdivided into subsystems that are themselves composed of increasingly simple independent subsystems, eventually reaching a level of relatively mechanical subpersonic specialized functions. According to this view, intentionality begins with simple relatively pre-conscious discrimination, and ranges through increasingly complex forms to self-conscious and self-corrective propositionalized beliefs. Since these routines of intentional activity can occur relatively independently of one another, intentionality can be a matter of degrees.

We can distinguish: a) pre-intentional physiologically based discrimination (e.g. discrimination between light and dark, sensitivity to heat. While such

discriminative response come to be integrated in the categorial intentional system, they also simultaneously continue to function pre-intentionally, at a relatively automatic, physical level.); b) pre-logical categorial discrimination which, while itself too vague to be propositionalizable, can be specified by general descriptions (e.g. mood responses to colors); c) propositionalizable interpretive descriptions of events and situations (e.g. seeing a stimulus as dangerous); d) propositionalized interpretations of situations and events ("This is a cumulus storm cloud."); e) critically evaluated propositionalized interpretations of situations and events ("In these climatic conditions, a cumulus storm cloud gives a 76% probability of rain."); f) reflexively critically evaluated propositionalized interpretations of situations and events (An evaluation of the statistical laws which are used to predict weather conditions under specific climatic and geographical conditions).

Some psychological states might standardly function as carriers of information without performing any functions that require their 'having the information they bear': though conforming to some conditions of intentionality, such protological states would not fully conform to conditions of rationality (van Gulick and Stich's subdoxastic states; cf. van Gulick 1982; Stich 1984). But those states, and the information they bear, can also be functionally connected with highly propositionalized intentional states. While by some standards of rationality, the functions of such subsystems preserve the intentionality of the mental, they would be defective by other standards. Similarly, pre-intentional states that also function in an intentionalized form, might be arational by some standards, irrational by others. For the explanation of self-deception, it does not matter whether some intentional activity is subconscious or whether there is some conscious intentional activity that is subdoxastic. What matters is that some psychological states apparantly fuse a number of different functions, with intentions whose rationality is measured by quite different standards. The relations among the various intentional subsystems affect the relation between the nonrational and the rationalizing functions of intentional actions. On this version of the second picture, there is nothing unusual about possible conflicts among the various grades of intentional 'takes' on the same event or situation. At least some types of self-deception might involve two distinctive and independent intentional processessing of the same event or situation, in cases where there might also be a presumption that at some level the two processes were also coordinated.

On this picture of the self, a person's activities - including intentional, voluntary and even purposive actions - need not arise from any particular set of motives: they can arise directly from constitutional structures without motivational intervention. Selective focusing and compartmentalization occur whenever the conditions for triggering the operations of the relevant subsystems obtain. A particular irrational action might have several

distinctive aetiological sources: overdetermination allows a psychological state or an action to be nonmotivatedly constitutional along one aetiological line, and to be motivated along another. In principle, constitutionally generated and motivationally based activities might sometimes conflict. There is no particular difficulty explaining the frequency and persistence of what is, on the first picture, simply classified as irrationality. On the first picture a good deal of our thought and behavior is simply classified as erroneous, conflicted, ignorant. But when irrationality is patterned, and when purportedly rational beings show unexpected resistence to correction, we need an explanation. We want to know not only how it occurs, but why it is such a fashionable indoor sport. Why is it often so highly patterned? And does the pattern explain the attractions it has for us?

On this second picture of the self, self-deception is readily assimilated into failure of integration among systems that are standardly coordinated. Even the capacities for critical rational reflection are subdivided into subsystems. Patterns of inference, calculation, 'stepping back to evaluate evidence' are analyzed as themselves arising from a large variety of constitutional, psychological and cognitive habits. Because psychological and intellectual activities are performed by relatively independent subsystems, there is no difficulty in explaining how a person can believe contradictions, can be aware and not aware of herself as holding contradictory views, can adopt conflicting policies. The phenomena of self-deception are naturalized and demystified on all varieties of the second picture. A self constructed from a set of subsystems might well be expected to be host to conflicting beliefs and conflicting strategies. 'Self-deception' becomes a natural by-product of functional structures and strategies. While the phenomena of self-deception are irrational and incoherent on the first picture of the self, they are standard normal operating procedure - the basic equipment of the self - on the second picture. Because there is no assumption that the system is constantly either selfinformed or even informed about its condition, no defensive regression of deception about the first level strategies is required. Since Laetitia need not be aware of the aetiology of her beliefs, she might well have persistent, patterned unfounded or malformed beliefs without being implicitly aware of her condition.

Sometimes self-deception occurs as a by-product of strongly entrenched patterns of focused attention. Just as there are patterns of dominance in visual attention (e.g., red over grey, irregularly moving over stationary objects), so too psychological salience follows strongly entrenched patterns that do not always reflect a person's beliefs and priorities. There is a general correlation between a person's considered priorities about what is important and the patterns of her attention: magnetising attitudes - fear, anger, love, competitiveness, eroticism, reactions to power - are connected

with what is important to us, and to our well-being. But the strength of those generalized habits of attention can over-ride considerations that are appropriate to specific particular situations and events. When that happens a person may not be able to use the material at the periphery of her attention, material of which she is one sense aware, to correct magnetised attention. Being aware of something does not occur at a single glance, at an instant. It takes place over time; it integrates distinctive actions of focusing, scanning, refocusing, reconstructing a series of interpretations derived from shifting the foreground and the background of attention. Standardly, a person can and does use the information that falls outside the center of a single focus to correct distortions of attention that arise from intensive focusing.

When a person is afraid, or absorbed in love or grief, or concentrated on some form of hierarchical combat, she can fail to integrate the relevant material that is the periphery of her strong attentive focusing. Sometimes such patterns of salience are strongly constitutionally based; sometimes their origin lies in the person's individual history. Sometimes their import is direct and obvious; sometimes it is indirect, encoded by idiosyncrayic associations. Sometimes it is directly motivated; sometimes it is a by-product of functional but unmotivated psychological structures or habits. When a person's attention is strongly riveted, she does not scan the periphery or background as often or as long a she normally does. Still, she knows in a general way what is there, and may even know that it is relevant to her primary concentration; she may even know that it provides a corrective to her beliefs and attitudes about what captures her attention so strongly. In principle, she is capable of redirecting her attention, of absorbing and integrating the peripheral material: such salience need not be pathologically obsessed. When psychological attention is strongly magnetised, what is not salient can seem subjectively unimportant, even when the person is aware that the pattern of her magnetised attention does not reflect her general attitudes, all things considered. Though salience and importance (particularly importance for truthfulness) are strongly correlated, they can vary independently. It is this feature of psychological structure that makes both self-deception and akrasia possible.

Self-deception is sometimes a variety of akrasia of attention or focusing. But why would we call this self-deception, rather than, say, being conflicted or obsessed or mistaken? When someone who can normally voluntarily redirect her attention in situations of this kind 1) identifies with, or underwrites the magnetised attitude and its consequences, and 2) denies the existence or the relevance of the corrective surrounding material, and when 3) there is evidence that she has understood the material she denies, then there are strong grounds for attributing self-deception rather than conflict, error or obsession. When magnetised patterns of salient attention involve attitudes that are basic to our sur-

vival and our sense of ourselves - as our fears, loves, griefs and competitions surely do -, the cannot be simply exorcised as external influences. And since they are central to actions for which we take responsibility, they cannot readily be excluded from the realm of the voluntary. Because they are crucial for normal functioning and constitute an important part of a person's self-image, their influence on our thoughts and attention are actions, they are things we do. If we are deceived by them, we are deceived by ourselves.

We return to the two superimposed pictures of the self. On the second picture, a set of subsystems that include patterns of visual and psychological attention, with weighted salience (fear, anger, love, eroticism, competition, hierarchy and domination) serve long range survival, even though they do not always express or reflect our individual commitments and priorities. But justifying beliefs and attitudes require a process of integration whose normative perspective is not reducible to the set of subsystems that compose the second picture of the self. Without the first picture, the strength of justification or validation of beliefs and attitudes can only be a function of the efficacious or causal strength of the contributory subsystem. In order to make sense of the self as actively integrating, evaluating and correcting the perspectival distortions that occur from strongly magnetised focusing, the first picture of the self must be independent of the second. Both pictures are required; neither can be subsumed by the other.

But it now seems as if this second picture of the naturalized self as a strategic survivor has demystified self-deception so thoroughly that it has evaporated. Starting out by saving the phenomena, we seem again to have lost them. The picture of the self as a loosely confederated system of subsystems, which includes the various activities of critical rationality without giving them any dominant centrality, loses self-deception: it has abandoned the identity of the deceiver and the deceived. The left hand is misguiding the right hand, the neck is averting the gaze of the eyes. But the eyes do not both see and not see, not exactly the same things in the same way at the same time. If self-deception is incoherent and impossible on the first picture, it is also lost on the second. The phenomena of self-deception again turn out to be nothing more than ignorance, conflict, non-integration or compartmentalization.

3. Self-deception is a disease only the presumptuous can suffer

Have all our attempts to preserve self-deception failed after all? I think not. We have, and require, two pictures of the self. On the picture of a rationally unified person, self-deception is not only irrational but incoherent. The ascription fails in some way. On the second picture, the

phenomena of self-deception are unproblematic ... but they are misleadingly characterized as the self deceiving itself.

Yet we seem convinced that strong self-deception does nevertheless exist, that it resists evaporation and reduction. We certainly think we can recognize self-deception in others, and we strongly suspect it in ourselves. Why then do we persist in thinking that there must be genuine cases of self-deception?

The reason that we are convinced that there is self-deception is that we cannot imagine how to renounce either of the two pictures of the self. The classical picture of the integrative rational self is the picture that makes sense of our attempts to systematize our beliefs, our attempts to integrate subsystems even when they are relatively independent. Even the weakest form of that picture, taken as a regulative ideal, is essential to thinking of ourselves as responsible agents and responsible believers. Those committed to the scientific enterprise are committed to attempting to avoid false beliefs, correcting them where possible, suspending judgment when necessary. Similarly, responsible agents, like responsible believers, are committed to consistent plans of action, interconnected by a system of reasoning. Whatever may be actually the case, those who hold themselves responsible must believe that the capacities for critical rationality do not merely form one subsystem, having no particular privilege over others.

There is a presumption that their role in the formation of beliefs, attitudes, actions is not only a direct function of their relative psychological strength: it reflects the power of their normative justification. The requirements of responsibility and rationality are not satisfied by assigning the subsystem of critical rationality the weight it would have on a principle of 'one subsystem, one vote'. Even for those who take epistemic and moral responsibility to be regionally a matter of degrees, the capacities of critical rationality must be prima inter pares, presumed to have centrality and dominance in the areas, and to the degree responsibility is assigned. Even when it is inconvenient, we consider ourselves obliged to dispell contradictions, if only by rationalizing or agonizing over them.

We can't imagine what it would be like to give up this picture of the self. Who would be the we who would consider whether, in the interests of truth and accuracy, we should renounce the pretentions of a regulative principle of rationality? Even characterizing the self as a set of subsystems seems to introduce a system, distinct from other systems-of-subsystems. After all, there are all sorts of subsystems: which do, and which do not, fall into the rough area of the self? In any case the self is, after all, a biological organism, a body that lives or dies, thrives or fails as one entity. Even when the subsystems do not always work together, even when they actually conflict, still at a minimal level, they are all either alive together

or dead together. Organic interdependence provides a presumptive basis for psychological integration.

But from the point of view of the second picture, such a victory is far too easy: what is at issue is not the existence, but the character and structure of organic interdependence. Any serious version of the first picture must introduce other capacities besides those of critical rationality: a creature whose only beliefs and motives are derived from the principles of critical rationality would be a very boring and short-lived creature. On the second picture, an organism that lives or dies as a whole, can be a loose confederation of autonomous subsystems, some of which can be effectively dead while the whole survives. Indeed the naturalized survival picture of the self describes and explains why we are often so patently and persistently arational and irrational despite our integrative efforts at rationality. Relying on the details of modular theories of all kinds, the second picture explains the functions of the sort of compartmentalization that is hospitable to self-deception and other forms of irrationality.

Each of the pictures purports to represent the important claims of the other. From the point of view of the first picture, the second of the self as a complex survivor fails to account for the centrality of integrative processes, and the centrality of critical rationality in the integrative processes. While it is complex, and while complexity can lead to error and conflict, the self cannot be deeply or radically divided. There couldn't even be a division of labour (let alone conflict) among subsystems unless there were the presumption of their integration. From the point of view of the picture of the self as a complex naturalized survivor, the first picture suffers from delusions of grandeur. Rational subsystems and their modes of integration may well claim centrality. That is their function and that is their business. Such regulative principles are meant to (attempt to) establish the dominance of rational strategies. But a subsystem, even a centrally important subsystem, claiming dominance by no means thereby establishes the validity of such a claim. From the point of view of the second picture, the first picture is unnecessary: whatever benefits it can genuinely bring can be captured within the second picture: and any other benefits it might claim, are illusory.

But though each of the two pictures claims to represent the other - claims to give an account of what seems right about the other -, they remain stubbornly and (it seems) irreducibly opposed. The apparent intractibility of self-deception comes from the superimposition of these two pictures. In the very nature of the case, we cannot let the two pictures stand side by side: we not only compare them, we superimpose them. And when we do, we see why we persist in taking self-deception seriously, without re-describing or reclassifying it. On first picture, the notion of a self committed to its own integration is not eliminable or compounded. Sub-

systems may be temporarily disassociated; but if they systematically and in principle resist integration, the self is deceived by itself. On this view the self is a reflexive entity designed to scan its subsystems for the sake of correcting, or at least avoiding error. When one part has misled another, the self has been deceived by itself; any failure of integration among parts is a presumption, though not a proof, of self-deception, because the system as a whole is set to integrating its subsystems. On the second picture of the self, absence of integration is not necessarily failure of integration; it requires no special explanation. When we superimpose the two pictures, what (on the second picture) counts as non-integration between distinctive subsystems is taken to heart by the first picture as a failure, a piece of inexplicable irrationality.

Why aren't the phenomena of self-deception saved by the second picture? The second picture might simply allow the modification to which the first picture is brought: responsibility and rationality are not all-or-none matters. Because we have trouble determining to what degree a person is epistemically responsible in a particular region, we have difficulty determining whether self-deception exists. The subsystems that represent the capacities of critical rationality will either succeed in establishing their presumptive claims to centrality and dominance or they won't. When they succeed, there was no self-deception. When they fail, there could not have been any self-deception because there was no possibility of integration. In any case, the survival picture of independent subsystems that attempts to give an account of responsible agents and believers must give an account of the strong dominance of the subsystem of critical rationality as scanning and integrating all other subsystems. But this efficiency redescribes the second picture with a superimposed the first picture.

So, then, it is only when the two systems are superimposed that the phenomena of self-deception appear intractible and irreducible. But this means that only certain sorts of people can be charged with self-deception. Like akrasia, self-deception is a disease only the presumptively strong can suffer: only those who take themselves to be responsible agents and believers, identify with, and by, their capacities for critical self-reflection, only those who take themselves to be defined by their achieving the first picture of the self can be charged with having failed in their commitment. From the point of view of the first picture, when the conditions for integration and responsibility are satisfied, there is no self-deception. When they are not, there cannot be self-deception. Those who assume epistemic responsibility, who assume the tasks of integration as an effective ideal, have committed themselves to the first picture of the self. For them, the ideal of integration is not one among other ideals: it is directly and centrally effective. One can only fail to fulfill ideals which are presumptively actually regulative: the rest is wishful thinking. Yet only those who, despite their effective commitment to the first picture, are

actually composed of relatively independent subsystems can fail to integrate what they believe.

But from the point of view of the second picture, the conditions for integration and responsibility are not global but regionally specific; and within each region, the satisfaction of these conditions is a matter of degrees. Since the self is not strongly unified, it does not as a whole, unified entity deceive itself: it can only fail to be integrated. If there is self-deception, in contrast to the failure of integration, it must occur regionally, within some regionally defined set of sub-systems. But when self-deception is regionalized and relativized, it is exactly correlative with the capacities for integration, region by region, degree by degree. From the point of view of the first picture, any localized failure, is a failure of the whole. Only if, and to the extent that she took integration as an effective ideal, was Laetitia Androvna self-deceived.

Notes

1) A first rough approximate characterisation of selfdeception, as it applies to that subclass of selfdeception covering propositionalizable beliefs:

   1. The person believes that p

   2. Either a. The person believes not-p. Standardly this involves the person believing g, which (given her beliefs and her strongly entrenched habits of inference), she ought to recognize as equivalent to not-p. Or b. The person denies that she believes p.

   3. If self-deception does not reduce the error, the person must on some level recognize that she has conflicting beliefs. Standardly, attributing such recognition is an inference to the best explanation of the person's behavior or patterns of inferences.

   4. If self-deception does not reduce to conflict, the person must on some level deny that her beliefs conflict. Sometimes this is achieved by an ad hoc strategy for reconciling the apparent conflict. The selfdeceiver usually makes no attempt to suspend judgement, or to determine which of her beliefs are defective.

   5. The attribution of self-deception presupposes an account of what the person would normally believe, perceive, notice, infer; it not only presupposes that she accepts and normally applied certain canons of rationality, but also that she is alert to the sort of evidence that weighs against her belief.

   There appears to be an interesting difference in the focus of discussion of the family of cases: philosophers in England and the United States have largely focused on cases of self-deception, cases where an individual adopts a complex strategy in the face of a conflict of specific belief of which she is presumptively aware. While such cases have not been neglected in France and Germany, philosophers there have focused

primarily on mauvaise foi as a general condition in which consciousness denies its condition as nothing-but-the-reflection-of-some-arbitrary-content-before-it. Or they have focused on false consciousness as a condition of a class of people whose beliefs and desires have been manipulated and directed in such a way as to violate their natural latent awareness of their real condition.

2) Straightway, then, self-deception seems problematic. For a start, its attribution seems to require a suspicious regression. Because I want to concentrate on other issues, I shall set aside at least some of the familiar puzzles about the attribution of self-deception. Certainly the general difficulties of identifying and attributing beliefs in opaque contexts make it difficult to demonstrate that there are bona fide cases of strict self-deception that not reduce to error or to conflict. But since these difficulties attend any attribution of belief - let alone the attribution of conflict of belief - they should not be themselves cast doubt on the existence of cases of strict bona fide self-deception.

3) E.g. the Platonic tri-partite soul in which each part not only assumes specialized function but also performs, at least at a minimal level, the functions assigned to the other parts; the Freudian account of a person as composed of internalized and introjected formative personae.

I am grateful to Owen Flanagan, Michael Martin, Richard Schmitt, Rüdiger Bittner, Jens Kulenkampff, Brian MacLaughlin, Martin Bunzl for comments and stimulating discussions. I benefitted from discussion at CUNY-Brooklyn and at SUNY-Albany. This article will also be published in: The Forms of Self-Deception, edited by Brian MacLaughlin and Amelie Rorty, Berkeley, forthcoming.

Bibliography

Dennett, D.C. (1985), The Self as a Center of Narrative Gravity, in: P.M. Cole/D.L. Johnson/F.S. Kessel (eds.), Self and Consciousness, New York

Fingarette, H. (1969), Self-Deception, New York

Pears, D. (1984), Motivated Irrationality, Oxford

Rorty, A. (1980), Self-deception, Akrasia and Irrationality, in: Social Science Information 19 (6), 905-922

- (1983), Akratic Believers, in: American Philosophical Quarterly 20, 175-183

Stich, St. (1984), From Folk-Psychology to Cognitive Science, Cambridge/Mass.

Van Gulick, R. (1982), Mental Representation - A Functionalist View, in: Pacific Philosophic Quarterly 63, 3-20