STEVEN D. HALES

SELF-DECEPTION AND BELIEF ATTRIBUTION

ABSTRACT. One of the most common views about self-deception ascribes contradictory beliefs to the self-deceiver. In this paper it is argued that this view (the contradiction strategy) is inconsistent with plausible common-sense principles of belief attribution. Other dubious assumptions made by contradiction strategists are also examined. It is concluded that the contradiction strategy is an inadequate account of self-deception. Two other well-known views – those of Robert Audi and Alfred Mele – are investigated and found wanting. A new theory of self-deception relying on an extension of Mark Johnston's subintentional mental tropisms is proposed and defended.

A prominent view about self-deception, probably the predominant one, is that self- deceivers both believe p and believe not-p.[1] The self-deceiver states, avows, asserts, or what have you, that p, and yet the self-deceiver's non-verbal actions belie a belief in p. Even more, it is assumed that these non-verbal actions positively provide strong evidence of a belief that not-p. Many have thought that these seeming facts are a contradiction in the making, or some kind of paradox. The literature is replete with attempts either to dissolve the paradox (by appeal to the subconscious, homuncularism, etc.), or arguments that there really is no such thing as self-deception, and that the evidence in its favor can be plausibly explained some other way.[2]

The central project of this paper is to undercut this entire approach. What I will argue is that modeling an account of self-deception on believing both sides of a contradiction (what I will call the contradiction model or strategy) is a seriously flawed enterprise, and that it will be far more fruitful to try to understand self-deception in some other way. The main difficulty with the contradiction strategy lies with its failure to cohere with our commonsense methods of belief attribution. More to the point, one cannot consistently accept both the contradiction strategy and our ordinary methods of ascribing beliefs. Other models of self-deception besides the contradiction strategy also suffer as the result of insufficient attention to belief ascription procedures, and I will review two well-known competitors. At the end of the paper I will outline a more promising account.

Synthese **101**: 273–289, 1994.

We ordinarily employ numerous methods, principles, and rules of thumb in ascribing beliefs to others. Some are good, others not so good. Let us consider a few of the more promising ones. One common sense principle of attribution is this:

> **BA**: If S's overall behavior is best explained by attributing to S the belief that p, then attribute to S the belief that p.

BA is of course too simple to be much theoretical use, and we may safely set it aside in favor of principles that exploit a division between two kinds of behavior: verbal and non-verbal. The usefulness of this taxonomy will become apparent shortly, as contradiction strategists are largely committed to this kind of division. They adopt, if not explicitly, belief attribution principles like these two:

> **BA1**: If S sincerely verbally avows p, then attribute to S the belief that p.
> **BA2**: If the best explanation of S's non-verbal behavior includes attributing to S the belief that p, then attribute to S the belief that p.[3]

A note on BA1: the requirement of sincerity is not supposed to logically entail belief. One natural way of interpreting sincerity is that sincerity consists in believing what one says. Another, non-coextensive, way is that sincerity is a matter of not intentionally deceiving. This latter is how sincerity is meant in BA1. If Mary sincerely claims that p, then she is not intentionally deceiving her audience about what she believes. I wish to leave open the possibility that Mary does not in fact believe p, and so may deceive her audience by asserting p. In such a case, Mary would have a false second-order belief – she takes herself to believe that p, but is mistaken. So by sincerely avowing p, Mary is attempting to report accurately a first-order belief, but might fail in the attempt.

Now, to see just how principles BA1 and BA2 are employed by contradiction strategists. Let us consider a benchmark case of self-deception. This example has appeared previously in the literature, and I take it to share all the relevant features with any case proposed as an example of the received paradigm of self-deception.

> If anyone is ever self-deceived, Dr. Laetitia Androvna is that person. A specialist in the diagnosis of cancer, whose fascination for the obscure does not usually blind her to the obvious, she has begun to misdescribe and ignore symptoms that the most junior premed-

ical student would recognize as the unmistakable symptoms of the late stages of a currently incurable form of cancer. Normally introspective, given to consulting friends on important matters, she now uncharacteristically deflects their questions and attempts to discuss her condition. Nevertheless, also uncharacteristically, she is bringing her practical and financial affairs into order: 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 complex but nevertheless straightforward explanations. But let's suppose Laetitia Androvna's case is not like that. The best explanations of the specific changes in her behavior require supposing that she has, on some level and in some sense, recognized her condition. (Rorty, 1988, p. 11).

There is much to say about this case, and I shall say very little of it. For the purposes of this discussion, the most salient thing about Laetitia Androvna is that she appears to believe both elements of a contradictory pair. On the one hand she sincerely denies that she has cancer, and provides esoteric etiologies for her symptoms. On the other she has made up a will, written farewells to her friends, and so on: behavior that – given her personality and mental faculties – seems best explained by a belief that she has terminal cancer. According to the contradiction model of self-deception, Laetitia Androvna believes p and believes not-p. Furthermore, she is disposed to deny that she believes p. She denies that "deep down", or anywhere else, she believes that she has cancer.

Now, it is on the basis of Laetitia Androvna's sincere verbal avowals that we attribute to her the belief that she does not have cancer. Likewise, it is because the best explanation of Androvna's non-verbal behavior includes attributing to her the belief (at some level, in some sense) that she does have cancer that we also attribute to her the belief that she has cancer. Both principles BA1 and BA2 are engaged in this scenario, and, I submit, they are almost always similarly engaged in other ostensible cases of self-deception.

BA1 and BA2 are not the only belief attribution principles relevant to self-deception; there are others that are just as plausible. One principle in particular is worthy of note. As we shall see, a purely general result of the assumption of self-deception is that adding this principle to the set of things to which the Contradiction Strategist is committed generates inconsistency. No special assumptions about whether it is

possible to believe explicit contradictions, or any of the other so-called paradoxes attaching to self-deception are needed. The ascription principle I have in mind is the flip side of BA1 and BA2. If BA1 and BA2 make sure we attribute enough beliefs, BA3 keeps us from attributing too many.

> **BA3**: If S sincerely verbally avows not-p, then do not attribute to S the belief that p.

This principle cannot be consistently endorsed by the Contradiction Strategist. We typically rely on BA3 all the time in distributing beliefs to subjects. If Smith says 'I'm not going to go to the store tonight after all', we are not inclined to think that Smith believes that he is going to go to the store tonight. Moreover, it is not out of simple habit that we do not assign the belief that p to Smith, rather, we *positively avoid* assigning him the belief that p on the basis of what he has said. If someone were to ask one of my friends, 'does Steve believe that Rhode Island has an incorruptible government?', my friend could justifiably reply 'definitely not – he has said the opposite many times'. Without BA3, no such inference is justified. If we reject BA3, we could not claim that S does not believe that God exists on the basis of such statements as 'it is false that God exists'. However, if S were to say 'it is false that God exists', BA1 allows us to attribute to S the belief that God does not exist. This is surely bizarre. In ordinary cases BA3 enjoys nearly as much employment as BA1; they seem to be on a par. Rejecting the former and keeping the latter is a radical modification of ordinary behavior. It would mean that we will have a much easier time figuring out what people believe than figuring out what they don't believe.[4] This is – minimally – a puzzling asymmetry. A philosophical theory that cripples or rejects BA3 in order to give precedence to BA1 needs considerable reasons to do so. Yet this is precisely what the contradiction strategist must do.

Here is why a contradiction model cannot include BA3. Consider: the standard case of self-deception is one in which these two things hold:

(1) S sincerely verbally avows that p.
(2) The best explanation of S's non-verbal behavior includes attributing to S the belief that not-p.

So far, no problem: the contradiction strategist just assigns a contradiction to S – believes *p* and also believes not-*p*. But genuine inconsistency lurks nearby. (1) and BA3 tell us that S does *not* believe *not-p*. However, (2) and BA2 tell us that S *does* believe not-*p*. Our principles and evidence entail that we both are and are not to attribute believing not-*p* to S. Thus taking BA2, BA3, and the empirical evidence about S together has inconsistent results. Something has to go. The contradiction strategist is committed to retaining BA2 and the evidence about S, since his stories about self-deception essentially rely on them to get off the ground. Hence the contradiction strategist must spurn BA3 in cases of self-deception.[5]

It is not too surprising that our attribution principles sometimes provide conflicting advice. A good theory of belief attribution would tell us what these cases are, and how to ameliorate the conflict. After all, prescriptive principles like these are generally defeasible given the right circumstances, and perhaps these are the right circumstances. However, it is a *prima facie* plausible assumption that all three principles independently plausible and are equally so. Hence in a case where the three taken together provide conflicting advice, we need an argument as to which one(s) should be overridden in favor of others. While this is a problem for anyone who accepts all three principles, it is particularly acute for contradiction strategists. They are committed to setting aside the advice of BA3 when conflict arises. But an argument is needed. This places the burden of explaining why BA1 and BA2 work but BA3 does not squarely upon the shoulders of those pushing a contradiction model of self-deception. Or, put another way, why do the data of self-deception serve to override BA3 instead of BA1 or BA2? This burden alone, even if answerable, serves to undermine the authority of a contradiction model as being the best explanatory account of self-deception.

One unsatisfactory rebuttal that the contradiction strategists might launch is to cite the existence of self-deception as *evidence* for the relative priority of BA1 and BA2. The contradiction strategists are partly concerned to demonstrate the existence of self-deception against the skeptics, and they do this by devising Androvna-type cases. But to make these cases work, and get the conclusion that a contradiction really is believed, they *rely* upon the priority of BA1 and BA2 over BA3. That is, they employ an attribution theory in which the dictates

of BA3 are set aside in favor of those of BA1 and BA2. Thus to defend the bias towards BA1 and BA2 by citing self-deception is to beg the question, as the contradiction model is premised upon this bias.

There are other problematic, perhaps deeper, assumptions made by contradiction strategists. For example, when the empirical evidence plus our belief attribution principles yields the assignment of contradictory beliefs to S, it is assumed that it is better to cling tight to our syllogism and attribute the contradiction than to say that we just don't know what S believes.[6] When Laetitia Androvna sincerely avows that she does not believe that she has cancer, while her non-verbal behavior suggests that she does, we need not conclude that she believes *p* and believes not-*p*. To be sure, this is the result of consistent application of our belief attribution principles. Yet this is not an inescapable result. It is just that we imagine that it is better to interpret Androvna as believing both elements of a contradictory pair and devise labyrinthine explanations of how this can be so than it is to accept any of these options: (i) holding that the conflicting evidence shows that we have not yet collected enough data on Androvna to merit belief attribution, or (ii) holding that the union of our belief attribution principles with the empirical evidence has produced a bastard child, thus showing that there is something wrong with our ascription theory, or (iii) holding that the conflicting evidence shows that Androvna is pathological in some way (she has a brain disorder, say) and may not be exhibiting intentional behavior.[7]

Why not take option (i)? Presumably in ordinary cases of attribution there is a minimal amount of empirical information needed before we are justified in assigning beliefs. No doubt we are to assume that in the Androvna case this level has been reached. But the information we have on Androvna is conflicting – suggesting that surely in this particular case we need more than the minimal data normally needed to justify belief attribution. Let us assume then that we have as much empirical observational data of Androvna as is possible to obtain. Our secret spies observe her every move and utterance, and still the behavioral data conflicts. We must assume that she believes both sides of a contradiction, or admit that our belief attribution theory is not complete. There is seemingly intentional behavior that it cannot in principle explain. If this were to happen, it would violate what seems like a desideratum for a theory of ascription – completeness. Without the requirement of completeness. we are faced with the prospect of behavior that

stems from beliefs and desires, but is such that we cannot in principle explain it by appeal to beliefs and desires. If this happens, then I think that we lose whatever motivation we had to think that their behavior was really intentional. Of course, I have no proof, or even any evidence, that our commonsense "theory" of attribution is complete. Perhaps it is because the folk theory is *in*complete that we get such a strange result in the case of Androvna. The thing to note is that contradiction strategists are committed to the assumption of completeness. Without this assumption there is an explanatorily powerful competitor to their account of Androvna's behavior – our ascription theory itself runs aground in cases such as Androvna's.[8]

Option (ii) claims that if our attribution theory tells us to assign contradictory beliefs to an agent, then this is reason to think that our theory is unsound – it can give us an explanation of all intentional behavior, but sometimes gives us the wrong answer.[9] Without the assumption that the theory is sound, it becomes very plausible to conclude that the real fault lies not so much with Androvna, but with the ascription theory itself. Contradiction strategists must assume that the theory does not tell us to assign beliefs where none exist.

Or maybe the seemingly intentional behavior is not intentional after all. Perhaps, as option (iii) suggests, Androvna's behavior is pathological, and akin to things like Tourette's Syndrome. Even the contradiction strategists are forced to admit that her thoughts are internally opaque in a certain way; else she would recognize the contradiction. She lacks the ability to reflect and transparently introspect her thoughts. Instead of struggling to explain this mental failing or deficiency (strong language, but consider that self-deceivers are typically held accountable for their deception), an alternative route would be to explain the deficiency in terms of a mental pathology. The exhibition of conflicting behavior is evidence of a non-intentional etiology for her behavior and clearly attribution principles are not to be employed when we have reason to believe that a mental pathology is responsible for the behavior under consideration. "Self-deception" may be a benign, or even beneficial pathological condition, but it is still a non-intentional mental failing.

Thus those endorsing the contradiction model of self-deception must assume the soundness and completeness of our ordinary methods of belief attribution, and they must assume that Androvna's behavior is intentional. In addition, they must be prejudiced against attribution

principle BA3 in a way that they cannot be with respect to BA1 and BA2. No contradiction strategist to my knowledge has defended, or even acknowledged these assumptions. Nor has anyone developed a theory explaining the interactions between the three ascription principles discussed or tried to resolve the apparent conflict among them. If I am wrong and it is not necessary for contradiction theorists to assume these things, it is at least quite hard to see how they plan to make their theory work.

The contradiction strategy is not the only model of self-deception, merely the most popular one. There are other ways of understanding self-deception. However, examining some of these theories from the standpoint of belief attribution once again shows the need for increased sensitivity to the ways in which we actually ascribe beliefs, and the ways in which theories of self-deception require that we ascribe beliefs. One current view is Robert Audi's. According to Audi, S is self-deceived only if S unconsciously knows that not-p, and S sincerely avows p.[10] Audi further insists that the attribution of unconscious knowledge is largely (I am unsure if he would agree to 'exclusively') justified by the observation of non-verbal behavior. So Audi pretty clearly adopts, and relies on, BA2. His treatment of BA1 is less clear. He writes, "as important as sincere avowal normally is in indicating belief, it does not entail belief' (Audi, 1989, p. 250). That is, Audi rejects this principle:

> **B**: If S sincerely verbally avows that p, then S believes that p.

B appears to be much stronger than BA1. B is an *a priori* conceptual claim, whereas BA1 is merely a prescriptive rule. B is not without its supporters – those philosophers committed to the primacy of the linguistic over the intentional are probably necessarily committed to it. The key thing is that rejection of B does not provide for rejection of BA1. Even those (Davidson, e.g.) who hold language and thinking to be in the same epistemic boat cannot get along without the weaker claim of BA1. Yet Audi seems to reject BA1, or, at least, he refuses to apply it under warranted conditions. After drawing up an apparent case of self-deception, Audi writes,

> Yet her sincere avowal of the proposition is like an expression of belief: normally, in fact, sincerely avowing that p implies believing it. We have, then. both knowledge that not-p and the satisfaction of a major criterion for believing p. The criterion is not a

logically sufficient condition, but it is strong enough to make its satisfaction in avowing a false proposition seem like being deceived in so speaking. (Audi, 1989, pp. 251–2).

Audi does conclude (from the non-verbal behavioral evidence and BA2) that the subject believes not-*p*, yet he stops short of saying that *p* is believed as well. Why? Aristotle claimed that 'it is impossible for anyone to believe the same thing to be and not to be, as some think Heraclitus says" (*Metaphysics* 1005b23). Many have thought that to reject Aristotle's dictum is to face the risk of paradox.[11] Perhaps Audi is in this camp as well; the position endorsed above conveniently prevents him from running afoul of Aristotle.

Yet by not drawing the conclusion warranted by BA1 (that the self-deceiver believes *p* in addition to not-*p*), Audi suggests that the non-verbal evidence about the subject along with BA2 yields epistemically superior results to those produced by the sincere verbal avowals of the subject coupled with BA1. Elsewhere he writes, "the sincere avowals of the self-deceiver . . . do not express beliefs . . . in the attribution of beliefs, actions speak louder than words" (Audi, 1985, p. 173). This is surely a surprising position. To think that we are more justified in reading beliefs off of non-verbal behavior than we are off of sincere verbal avowals brings to mind yet another version of the old behaviorist joke: you believe *p*, do I?

Audi admits, more or less directly, that BA1 has a strong claim on our allegiance. Yet to skirt paradox he buys, perhaps unwittingly, a rather odd sort of attribution theory. Either BA1 has no place in Audi's theory, in which case Audi's ascription procedures are both radically different than our ordinary commonsense ones and strangely behavioristic, or the dictates of BA1 are overridden in Androvna-type cases. If it is the latter (a more plausible stance) we are left wondering why BA1 should be the principle overridden in this case and not, say, BA2. An argument is needed.

Moreover, it is hard to see why Audi fears the threat of paradox so strongly that he rushes into the arms of the behaviorists to escape it. Audi already accepts the idea of an unconscious, and unconscious beliefs. The conscious/unconscious split alone has often been considered enough to explain the failure of Aristotle's dictum, and avert the risk of paradox.[12] Why does Audi think that he needs additional firepower? Perhaps he is so enamored of Aristotle's dictum that in order to preserve it he prefers to reject (or at least severely weaken)

one of the most powerful and plausible belief attribution principles we have (BA1) and reject the widespread view that the existence of unconscious beliefs shows Aristotle to be mistaken. Yet at least in one place Audi grudgingly admits that both believing *p* and believing not-*p* is 'a bare possibility' (Audi, 1988, p. 96). If Audi's considered view is indeed that it is possible to believe *p* and believe not-*p*, then his motivations for rejecting BA1 become obscure.

Another recent approach is offered by Alfred Mele. He argues that self-deceivers do not believe both sides of a contradiction, but rather that self-deception consists in having a false belief that *p* because of a desire that *p* be true, and that this desiring causes the deceiver to manipulate and misconstrue the evidence that bears upon the truth-value of *p* (Mele, 1987b, p. 127). Applying his treatment to the Androvna case we have been considering, Mele would say that Androvna believes that she does not have cancer, and that's it. What she has done is to distort systematically the available evidence about her condition because of her desire not to have cancer. She has searched for esoteric etiologies to account for her symptoms, and so on. Her belief that she does not have cancer rests upon her successful mistreatment of the evidence bearing upon her self-diagnosis.

So far, so good. I think that Mele is right in explaining how an Androvna-type self-deceiver distorts the data relevant to her (self-deceived) belief that *p*. Mele is also right that this distortion is what is causally responsible for the belief that *p*. However, his account stumbles when accounting for Androvna's non-verbal behavior. Mele offers two possible explanations of this behavior. The first amounts to suggesting not that Androvna also believes that she has cancer, but instead she believes the chance that she is mistaken (about not having cancer) is high enough that it warrants getting her affairs in order, drawing up a will, etc. It's not the case that she also has the contradicting belief that she *does* have cancer, only that she has a belief that she *might* have cancer, and this belief is sufficient to motivate action.[13]

It is difficult to argue against Mele's explanation here, since he provides no real argument for it and is content to offer it as a logically possible alternative. However, the Androvna case is set out in such a way that it should trigger application of belief attribution principle BA2: if the best explanation of Androvna's non-verbal behavior includes attributing to her the belief that *p*, then attribute to her the belief that *p*. Mele apparently thinks that the best explanation of Androvna's non-

verbal behavior includes attributing to her not the belief that she has cancer, but the belief that she *might* have cancer. But it is not clear why attribution of only the weaker belief is warranted. Doing so avoids relapse to the contradiction model, but without further argument it appears incongruous with our ordinary methods of ascribing beliefs, and seems merely *ad hoc*.

Mele's second suggestion is that Androvna's non-verbal behavior, if it is really systematically at odds with her verbal avowals, provides 'an excellent reason to deny that [s]he believes what [s]he asserts' (Mele, 1987b, p. 131). Mele is right that we might well conclude that Androvna is lying to us (she doesn't want us to worry about her, say) about what she really believes. But suppose that she is *not* intentionally deceiving us. Then Mele's suggestion looks much like Audi's – ignore BA1 in favor of BA2 if you need to – and suffers from the same problems that plague Audi's view and were discussed above.

So it seems that Mele's approach carries the ball only halfway; while a plausible theory with respect to the verbal avowals of the self-deceiver, it does not adequately account for non-verbal behavior.

I want to outline briefly a more promising approach, one that is an adaptation, and in some respects a natural extension, of Mark Johnston's theory of mental tropisms.[14] Johnston argues that self-deception is made possible by mental tropisms. Tropisms are nonaccidental, purpose-serving, adaptive, subintentional mental processes. They are something like unreflective habits of mind. We might think of the mind here as a tidal pool, with many different forces at work, and many different bits of flotsam within. At times some of the flotsam might clump together enough that a largish piece floats to the surface and becomes a full-fledged belief, or desire, or fear. And then it might break apart and fall from view. But it still remains within the pool. The fact that a piece of flotsam is not well-organized or formed enough to float to the top and engage in nice macro-level causal relations in no way means that it has no causal powers in the pool. The small bits of flotsam beneath the surface – barely noticable emotions, little hunches, suspicions, attitudes that seem to vanish if you stare directly at them like faint stars – may well be causally involved in what is often called denial, repression, wishful thinking, and the like. These are currents mostly below the surface of the water, which still may have visible effects above. Moreover, just as the activity of a tidal pool is complex but not random, so it is with the mind. It is not improbable to think of this

activity beneath the surface as adaptive, and serving the purposes of evolution.

What these tropisms provide, according to Johnston, is the mechanism by which a self-deceiver can come to have the belief that p. Aside from the paradox involved in believing a contradiction, another "paradox" of self-deception is how a self-deceiver can successfully lie to herself anyway. Sartre succinctly puts it this way:

> [I]f I deliberately and cynically attempt to lie to myself, I fail completely in this undertaking; the lie falls back and collapses beneath my look; it is ruined *from behind* by the very consciousness of lying to myself . . . (Sartre, 1956, p. 49).

It is this problem that Johnston's theory of subintentional processes is primarily meant to solve. Self-deceivers do not *intentionally* deceive themselves; rather their successful deception is the result of a subintentional mental regularity, or tropism. In our paradigm case of Androvna, for example, the appeal to tropisms explains how it is possible for her to engage in the perversion of the evidence pertaining to her condition, and believe that she does not have cancer in the face of powerful evidence to the contrary. The operation of the tropism causally antecedent to her belief that she does not have cancer serves some adaptive end, for example, reducing her anxiety over having cancer. Tropistic processes persist for a variety of adaptive purposes (Johnston focuses on anxiety reduction, but as Pears (1991) points out, this is unnecessarily restrictive), but they are not intentionally employed for these purposes. They are not subject to intentional employment at all – they are in this way subdoxastic.

So, Johnston's tropisms play a causal role for subjects, and this role allows the development of the belief that p, the belief that the self-deceptive subject verbally avows. Johnston is silent on the issue of non-verbal behavior and whether we are justified in assigning the belief that not-p on the basis of this behavior. There is nothing in his account that entails rejecting the contradiction strategy. However, I want to suggest that the role of tropisms may be broader than Johnston assigns. For him mental tropisms are 'connections between mental state types' (Johnston, 1988, p. 88). But this is too narrow. It seems that not only do tropisms provide a way out of the paradox presented by Sartre and allow for explanation of how the belief that p is formed, but that much is gained by extension of their causal powers to the domain of behavior as well. More specifically, the domain of non-verbal behavior.

Androvna's anxiety over the prospect of having cancer is alleviated or reduced by the belief that she does not have cancer, and this belief is made possible by a mental tropism. But what of her non-verbal behavior? Her behavior does seem *purposeful*, in some way. This after all is why contradiction theorists are motivated to say that it is due to a belief. Surely Androvna does not simply find herself writing a will one day without any beliefs (or at least fully doxastic hypotheses) about why she is doing so. No doubt she takes herself to have actual reasons for writing a will. Why is she writing a will? What are her reasons? Well, say the contradiction theorists, she is writing up a will because she believes that she is dying of cancer, and so she has contradictory beliefs after all.[15] It does seem right that Androvna must take herself to have reasons for writing her will (and related activities). And yet I claim that despite this, it is not a belief that is at the root of her behavior; it is instead a tropism.

Not only is Androvna anxious over the prospect of having cancer, a fact addressed by a tropism that causes her to form the belief that she does not have cancer, but it is also reasonable to suppose that she has other related stresses and anxieties as well. Besides the prospect of cancer, there are also the prospects of leaving her affairs unsettled, and the grief of her friends. These possibilities are liable to lead to a kind of anxiety that would be reduced by increased attention to her practical and financial affairs, drawing up a will etc. Insofar as we are prepared to grant, along with Johnston, that a subintentional tropism allows for the anxiety-reducing belief that she does not have cancer, so we might well think that a similar tropism is at the root of the anxiety-reducing action that Androvna exhibits in her relevant non-verbal behavior.

Consider by comparison the commonplace act of looking in the refrigerator. People often open the door and stare without any clear plan to retrieve food, or even to learn about the contents. There is just something satisfying about the looking. People are, of course, unlikely to look in the refrigerator without fully doxastic hypotheses as to why they are doing so. That is, if asked, they could probably supply a rationale. So it is in the case of Androvna. Suppose we were to ask Androvna why she is writing her will, if she really doesn't think that she is dying of cancer. She would probably reply that she just figured that it was high time, that it has been on her list of projects for awhile and she has finally gotten around to it, that she recently read something in the paper about the importance of having a will, etc. One reason

she cannot cite is a belief that she has terminal cancer. If she did cite this as a reason, then she is not a self-deceiver after all – on any model. Thus, if a belief that she has terminal cancer is the reason (in the sense of a cause) that she is writing her will, it is one that she cannot consciously take to be a reason for will-writing; i.e. it is a non-conscious cause. But a subintentional tropism can account for this data as well as the contradiction model: it can be a non-conscious cause of action that a self-deceiver cannot cite as a reason for, or a cause of, her action. So the fact the Androvna takes herself to have reasons for writing her will is consistent with the actual causes for her will-writing residing in a subintentional tropism.

The solution I am recommending is in one sense a modification of our ordinary belief attribution procedures. I am claiming that BA2 may not be applicable in the case of Androvna – not because some other belief besides the one that she has cancer is held instead (as Mele suggests), but because there is no belief behind her behavior. The causal genesis of her behavior lies not in a belief but in a mental tropism. I am not proposing that BA2 is false, or overridden, but that it is not appropriately triggered in the case of self-deception. BA2 has to do with the assignment of belief, and in Androvna-type cases there is no fully doxastic attitude to be assigned. Note that I am not *ad hoc*-ly rejecting a belief attribution principle because of conflict with a pet theory of self-deception, as does Audi. Rather, this is a call for a more sophisticated account of mental state attributions, one which is sensitive to mental states less robust than beliefs, and one that does not over-rationalize what really happens in cases like self-deception. The mental habits that cause the behavior characteristic of self-deception may be good ones, or they may be bad ones, but they are something less than beliefs. This, I believe, serves to explain why attempts to explain self-deception wholly in terms of beliefs have been less than completely successful. The problems with the theory of belief attribution faced by the contradiction strategy are subverted by the model I am sketching. With an extended version of Johnston's tropisms, we can attribute to Androvna the belief that she does not have cancer, refrain from assigning her the belief that she does have cancer, and still explain her non-verbal behavior in a satisfactory way.

Questions remain, of course. When should we attribute a belief instead of a tropism? When should we do the reverse? These are not

easy to answer, but I think they are answerable, perhaps by psychology. In any case they are properly the subject of a different essay.[16]

NOTES

[1] The first person I know of to make this explicitly a requirement of self-deception is Demos (1960). Also see Siegler (1968), Miri (1974), McLaughlin (1988), and da Costa and French (1990).

[2] Skeptics include Haight (1980) and (1985) and Paluch (1967). For a survey of attempts to solve this paradox, see Mele (1987a).

[3] Compare the belief attribution principles proposed by Kripke (1979, pp. 248–49).

[4] It is worth noting that people who accept the direct reference of singular terms sometimes claim this. They claim, for exmple, that if S assents to the sentence 'Cicero was bald' we can conclude that S believes that Cicero was bald, but we cannot conclude that S does not believe that Cicero was bald if S assents to the sentence 'Tully was not bald'. So here we have a failure of BA3 in favor of BA1. However, even if the direct reference theorists are right about this, we have no reason to think that BA1 will *always* override BA3 in case of conflict. Maybe in other areas the reverse will happen. So to say that BA3 overrides BA1 in the case of self-deception requires an independent argument.

[5] Other attribution principles are problematic for Contradiction Theorists for similar reasons. This one for example:

> **BA4**: If the best explanation of S's non-verbal behavior includes attributing to S the belief that not-p, then do not attribute to S the belief that p.

This principle is more controversial than BA3, since it may imply that we have beliefs in propositions that we have never considered, and so the difficulties for the contradiction strategy are best made with BA3 alone.

[6] It might be profitable to compare the apparent conflict with the principle of charity, a principle often claimed to be needed for behavior interpretation and belief attribution. See especially Davidson (1984b) and (1984c).

[7] By 'intentional behavior', I mean behavior that stems from beliefs and desires. I do not mean *intended* behavior, which is something different.

[8] By comparison, consider the view of van Fraassen that the self-deception literature maintains its momentum on the basis of austere philosophical stories with thinly veiled ulterior motives. Examining tales offered in good faith – say, from literature proper – reveals deep complexity. In fact, the empirical facts in such stories are so complex that it becomes pragmatically impossible to apply our belief attribution principles, and we are unable even to sketch paradigm examples of self-deception. This radical failure of our attribution procedures results in our not knowing what to think. The relevant beliefs of the characters in the story become inscrutable. See van Fraassen (1988). Also cf. Nussbaum (1988).

[9] This position gains additional momentum if one accepts the position of daCosta and French that $(Bp \ \& \ B \sim p) \rightarrow (Bp \ \& \ \sim Bp)$. This principle seems false to me, but I will not pursue it here. See their (1990, p. 179): "At the 'factual level . . . $[(Bp \ \& \ B \sim p) \rightarrow (Bp$

& ~Bp)] holds good . . . . This, we claim, is the level on which the phenomenon of self-deception occurs". See also p. 189: "The conclusion we reach is that believing that not-p and not believing that p should not be regarded as distinct where self-deceivers are concerned".

[10] He defends this account in several places. See Audi (1985, p. 173), (1988, p. 94), (1989, p. 249).

[11] And some have thought even stranger things – cf. the da Costa and French entailment relation mentioned in footnote 9.

[12] See, e.g. McLaughlin (1988).

[13] See Mele (1987b, pp. 130–1).

[14] Johnston (1988). For some discussion of Johnston's view, see Pears (1991).

[15] Al Mele stressed this point in personal correspondence.

[16] A version of this paper was read at the 1992 Pacific Meetings of the American Philosophical Association. My commentator on that occasion was Taylor Carman. Conversations with James Dreier first got me started on this topic, and Al Mele and Steven Rieber generously provided me with detailed criticisms on earlier drafts, which resulted in (I hope) considerable improvements. I have also benefited from the comments of three anonymous reviewers for *Synthese*.

REFERENCES

Audi, R.: 1985, 'Self-Deception and Rationality', in Martin (ed.), 1985, pp. 169–94.

Audi, R.: 1988, 'Self-Deception, Rationalization, and Reasons for Acting', in McLaughlin and Rorty (eds.), 1988, pp. 92–120.

Audi, R.: 1989, 'Self-Deception and Practical Reasoning', *Canadian Journal of Philosophy* **19**, 247–66.

DaCosta, N. C. A. and French, S.: 1990, 'Belief, Contradiction, and the Logic of Self-Deception', *American Philosophical Quarterly* **27**, 179–97.

Davidson, D.: 1984a, *Inquiries into Truth and Interpretation*, Oxford University Press, Oxford.

Davidson, D.: 1984b, 'Thought and Talk', in Davidson (ed.), 1984a, pp. 155–70.

Davidson, D.: 1984c, 'Truth and Meaning', in Davidson (ed.), 1984a, pp. 17–36.

Demos, R.: 1960, 'Lying to Oneself', *The Journal of Philosophy* **57**, 588–95.

Haight, M. R.: 1980, *A Study of Self-Deception*, Harvester Press, Brighton, Sussex.

Haight, M. R.: 1985, 'Tales From a Black Box', in Martin (ed.), 1985, pp. 244–60.

Johnston, M.: 1988, 'Self-Deception and the Nature of Mind', in McLaughlin and Rorty (eds.), 1988, pp. 63–91.

Kripke, S.: 1979, 'A Puzzle About Belief', in Margalit (ed.), 1979, pp. 239–83.

Margalit, A., ed.: 1979, *Meaning and Use*, D. Reidel, Dordrecht, Holland.

Martin, M. W., ed.: 1985, *Self-Deception and Self-Understanding*, University Press of Kansas, Lawrence, Kansas.

McLaughlin, B.: 1988, 'Exploring the Possibility of Self-Deception in Belief', in McLaughlin and Rorty (eds.), 1988, pp. 29–62.

McLaughlin, B. and Rorty, A. O., eds.: 1988, *Perspectives on Self-Deception*, University of Califomia Press, Berkeley.

Mele, A.: 1987a, 'Recent Work on Self-Deception', *American Philosophical Quarterly* **24**, 1–17.
Mele, A.: 1987b, *Irrationality: An Essay on Akrasia, Self-Deception, and Self Control*, Oxford University Press, Oxford.
Miri, M.: 'Self-Deception', *Philosophy and Phenomenological Research* **34**, 576–85.
Nussbaum, M.: 1988, 'Love's Knowledge', in McLaughlin and Rorty 1988, pp. 487–514.
Paluch, S.: 1967, 'Self-Deception', *Inquiry* **10**, 268–78.
Pears, D.: 1991, 'Self-Deceptive Belief Formation', *Synthese* **89**, 393–405.
Rorty, A. O.: 1988, 'The Deceptive Self: Liars, Layers, and Lairs', in McLaughlin and Rorty (eds.), 1988, pp. 11–28.
Sartre, J.-P.: 1956, *Being and Nothingness*, trans. Hazel E. Barnes, Philosophical Library, New York; originally published as *L'Être et le Néant*, 1943.
Siegler, F. A.: 1968, 'An Analysis of Self-Deception', *Nous* **2**, 147–64.
van Fraassen, B.: 1988, 'The Peculiar Effects of Love and Desire', in McLaughlin and Rorty (eds.), 1988, 123–56.

Department of Philosophy
Bakeless Center for the Humanities
Bloomsburg University
Bloomsburg, PA 17815
USA