American Philosophical Quarterly
Volume 36, Number 2, April 1999

ON THE DISTINCTION BETWEEN CONSCIOUS AND UNCONSCIOUS STATES OF MIND

David H. Finkelstein

> Yes, I preferred the elderly and discontented doctor, surrounded by friends and cherishing honest hopes; and bade a farewell to the liberty, the comparative youth, the light step, leaping impulses and secret pleasures, that I had enjoyed in the disguise of Hyde. I made this choice perhaps with some unconscious reservation, for I neither gave up the house in Soho, nor destroyed the clothes of Edward Hyde, which still lay ready in my cabinet.
>
> —Robert Louis Stevenson, "The Strange Case of Dr. Jekyll and Mr. Hyde"

In this passage from Stevenson's famous story, Dr. Jekyll recalls a choice that he made but failed to live by—a resolution to never again transform himself into Edward Hyde. Jekyll remarks that he "made this choice perhaps with some unconscious reservation." What work does the word "unconscious" do in this sentence? When is a reservation (or an intention or a fear) rightly said to be unconscious? My aim in what follows is to explain what it is that distinguishes an unconscious state of mind from a conscious one. My procedure will be as follows. I am going to present several tempting, but ultimately unsatisfactory, views concerning that by virtue of which a state of mind should be characterized as either unconscious or conscious. My criticisms of these views will yield a set of conditions that an adequate account of the distinction between conscious and unconscious mentality ought to meet. I'll then offer what I take to be such an account, and I'll show that it both meets the conditions of adequacy and helps us to answer a number of questions that would appear puzzling without it.

Perhaps I should add that my aim is not to try to convince a reader who is uncomfortable with talk about unconscious states of mind that it's *all right* to describe a person as, for example, unconsciously jealous of his brother. We do characterize people as subject to a wide range of unconscious mental states, and I'm going to take it for granted that we are often justified in doing so. I aim to elucidate what we mean by such characterizations. I will not, however, be

assuming that dreams are the royal road to the unconscious or that little boys are unconsciously afraid of castration or anything else that is distinctively Freudian or psychoanalytic. We should remember that Freud was not the first to speak of unconscious states of mind (as is evidenced not only by such stories as "The Strange Case of Dr Jekyll and Mr. Hyde"—which was first published in 1886—but also by 19th century theoretical writings about the nature of the human mind[1]).

I. The Very Simple and Not-so-simple Views

Let's begin by considering a very simple way of understanding what it means to say that someone's mental state is either conscious or unconscious. The view might be put as follows: "Your mental state is conscious if you know that you are in it. Your mental state is unconscious if you don't know that you're in it. To say that you, for example, unconsciously believe that no one could ever fall in love with you is to say: (1) that you believe that no one could ever fall in love with you and (2) that you don't know—you're unaware of the fact—that you believe this. Call this the very simple view.[2]

A bit of reflection reveals that the very simple view is unsatisfactory. Imagine someone—call him Harry—who says: "My therapist tells me that I unconsciously believe no one could ever fall in love with me, and she's generally right about such things, so I suppose I must have this belief." Let's imagine that Harry's therapist *is* right about him, and that Harry is justified in believing that she's right about him. Harry is, then, aware of his belief that no one could ever fall in love with him. According to the very simple view, we should say that Harry's belief that no one could ever fall in love with him went from being unconscious to being conscious as a result of his conversation with his therapist. But we can imagine that Harry holds no such conscious belief. When asked about his future, Harry says, "Oh, I'm sure that eventually someone will fall in love with me, even though my therapist has convinced me that unconsciously I believe it's impossible that anyone should." This is a perfectly intelligible remark. What Harry's therapist has made him aware of is not that he is in some way unlovable, but only that he unconsciously believes this to be so. The very simple view, however, cannot allow for the intelligibility of what Harry wants to say about himself. According to the very simple view, if someone is aware that he believes such-and-such, then his belief is conscious. But, in our example, Harry is aware of his unconscious belief that no one could fall in love with him.

The very simple view is *too* simple; it's not faithful to the way we use the words "conscious" and "unconscious." There is, however, a use of these words to which something like the very simple view *is* faithful. The case of Harry demonstrates that there is a distinction to be drawn between two uses of the word "conscious" or "unconscious." First, there is a relatively unpuzzling use that is generally in place when the word "conscious" or "unconscious" is followed by the word "of" or "that." I might say, "Until the lights came on, I had been unconscious *of* the person in the seat next to me." Or: "Lois suddenly became conscious *that* she was the only patent attorney in the room." In such contexts, the word "conscious" means, roughly, *aware*. The corresponding use of "unconscious" means *unaware*. Among the things I might become aware, or conscious, *of* are my own states of mind. But to say that I am conscious of, for example, my fear of abandonment is not to say either that I consciously fear abandonment or—what amounts to the same thing—that my

fear of abandonment is conscious. Describing Harry as "conscious of unconsciously believing that no one could ever fall in love with him" (or "conscious of his unconscious belief that no one could ever fall in love with him") seems to involve a contradiction only if we confuse two senses of the word "conscious." It's one thing to be consciously angry or jealous or believing such-and-such and quite another to be conscious of one's own anger or jealousy or belief. We can think of this fact as providing us with our first constraint on an adequate account of the distinction between conscious and unconscious mentality.

> Constraint 1: An account of the distinction between conscious and unconscious mentality should respect the difference between:
>
> (a) someone's consciously believing or being consciously afraid, i.e., a belief or a fear's being conscious, rather than unconscious;
>
> AND
>
> (b) someone's being conscious *of* her own belief or fear, i.e., conscious *that* she believes or fears such-and-such.

A variation on the very simple view might appear to meet Constraint 1. The view that I have in mind could be expressed as follows: "The problem with the very simple view is that it doesn't take into account the fact that there are various *kinds* of knowledge. If my knowledge that I believe *p* is based only upon the testimony of my therapist, then—while I may be said to be conscious *of* my belief that *p*—I cannot be said to consciously believe *p*. The kind of self-knowledge that the word 'consciously' picks out is not knowledge by *testimony*. What it means for me to consciously believe that *p* (or to be consciously hopeful or afraid, etc.) is that I know my mental state via a *particular* cognitive mechanism—the mechanism by which I ordinarily find out about my own states of mind. It is tempting to refer to this mechanism as 'inner sense,' but perhaps we should just call it 'mechanism M.' What it means for one of my mental states to be conscious (i.e., for me to be consciously angry, sad, or whatever) is that I'm aware of it *via mechanism M*. What it means for one of my mental states to be unconscious is that—although I may be aware of it—I am not aware of it via mechanism M." I'll call this the not-so-simple view.

It is a good deal more difficult to state a decisive objection to the not-so-simple view than to its very simple progenitor. I present what, I think, amounts to such an objection elsewhere,[3] but the argument is too lengthy to rehearse here. In the next couple of paragraphs, I'll merely offer a reason to be *suspicious* of the not-so-simple view—a reason that suggests the direction that I'll be taking later in the paper. Let us return to the case of Harry who consciously believes that someone will eventually fall in love with him, even though he is aware of his unconscious belief that no one could. The fact that, in this example, Harry's conscious and unconscious beliefs contradict each other helps us to see the inadequacy of the very simple view. But the need to distinguish between what someone is aware of believing and what he consciously believes does not depend on there being a flat-out contradiction between his conscious and unconscious beliefs. To see this, consider a variation on the example. Imagine that Harry says: "I unconsciously believe that no one could ever fall in love with me," whereupon he's asked whether anyone *could* fall in love with him. He answers, "Maybe; I'm not sure." Here too, Harry's belief that no one could fall in love with him is an unconscious one *of* which he is aware.

So, Harry's conscious opinions need not quite *contradict* his unconscious belief that no one could ever fall in love with him. Nevertheless, there *would* seem to be something wrong with Harry's saying, "I unconsciously believe that no one could ever fall in love with me; moreover, no one could ever fall in love with me." If Harry is willing to assert that no one could ever fall in love with him, then it's not right for him to say that he believes this unconsciously. We might think of this point in connection with Moore's paradox. Moore pointed out that it would be absurd for someone to utter a sentence of the form, "I believe that *p*, and it is not the case that *p*." A number of writers have since noted that Moore's point does not hold for self-ascriptions of unconscious belief. In other words, there's nothing wrong with saying, "I *unconsciously* believe that *p*, and it is not the case that *p*." I'm calling something further to your attention: that, *prima facie*, there does seem to be something wrong with saying, "I unconsciously believe that *p*, and it *is* the case that *p*" (even though, as with Moore's paradox, both conjuncts might be true).[4] When we consider unconscious mental states, we find not only the failure of Moore's paradox, but, as it were, the inversion of it. We might call this Eroom's paradox.

Although I won't call it a *constraint* on any satisfactory account of the difference between conscious and unconscious mentality, it does seem reasonable to expect that such an account would shed light on Eroom's paradox. In other words, it's reasonable to expect that if we come to understand what distinguishes unconscious states of mind from conscious ones, we'll also understand what would be wrong with saying, "I unconsciously believe that my brother has ruined my bid for reelection; moreover, he has ruined it." And a reason to be suspicious of the not-so-simple view is that it doesn't look like it will help us come to grips with what would be problematic in such an utterance. According to the not-so-simple view, a belief's being unconscious lies in the fact that the subject doesn't know about it via a particular cognitive mechanism. But now, while there *would* seem to be something wrong with most imaginable utterances of the form, "I unconsciously believe that *p*; moreover, *p*," it's not at all clear what, if anything, would be wrong with someone's saying, "My knowledge that I believe *p* is not based on cognitive mechanism M; moreover *p*." Thus, the not-so-simple view leaves Eroom's paradox looking like a mystery. Again, I don't claim that this constitutes anything like a decisive objection to the not-so-simple view. Nonetheless, I do think it provides us with a reason to seek another account of unconscious mentality—one that helps us to make sense of Eroom's paradox.

According to both the very simple and the not-so-simple views, *unconscious* and *conscious* are, as it were, epistemic notions: on either view, to say that a mental state is unconscious is to say that the subject lacks some sort of *knowledge* that he would enjoy were the mental state conscious. Perhaps the difficulties that I have raised for the two views suggest that this is not the best way to think about the difference between conscious and unconscious mental states. In what follows, I'll try to show that there is a better way to think about this difference—one according to which *conscious* and *unconscious* are not epistemic notions.

Moore's paradox indicates that our statements about the world and our self-ascriptions of conscious belief hang together in a particular way. Eroom's paradox indicates that our statements about the

world and our self-ascriptions of unconscious belief come apart in a particular way. What Eroom's paradox starts to bring out, I think, is that when we speak of someone's believing something unconsciously, we are characterizing, not an epistemic lack, but rather, a certain kind of rupture in the way that a person's beliefs (as expressed in his claims about the world) and his self-ascriptions of belief ordinarily hang together. This paper can be understood as an attempt to shed light on the kind of rupture that this is.

II. "Unconscious Mental Representations"

I have distinguished between two ways in which we use the words "conscious" and "unconscious": it's one thing to say that someone's fear or jealousy is conscious rather than unconscious (i.e., that he is consciously afraid or consciously jealous), and quite another thing to say that someone is conscious *of* (i.e., aware of) his own fear or jealousy. Now, there are other ways in which we use the words "conscious" and "unconscious" that I won't be concerned with. For example, we sometimes speak of someone's being "knocked unconscious." This usage is connected to the "conscious of" locution. A person who is knocked unconscious is not conscious of anything.[5]

I do want to devote a few paragraphs to saying a little bit about a potentially confusing use of the word "unconscious" that has become common in certain circles. Cognitive scientists sometimes characterize (what they call) mental representations as unconscious. A cognitive psychologist who was interested in visual object recognition might say, "When you read a book, while you are thinking about what the author is saying, you are also constructing an edge map—an unconscious mental representation of the edges in your visual field." It's a mistake to understand such representations as akin to unconscious attitudes and emotions. It will help to introduce some of the themes that will be important later in the paper if I say a little bit here about the character of this mistake.

Let's stay with the example of edge maps. According to David Marr's theory of visual cognition and many theories that have come in its wake, at an early stage in visual processing, a provisional map of the edges in the visual field is computed.[6] This is thought by some to occur in what are called the ocular dominance columns located in striate visual cortex.[7]

Given suitable circumstances, you might want to say, "A set of my ocular dominance columns *believes* that edges X and Y meet." But it would be courting confusion if you were to call such a "belief" one of *your* beliefs—conscious or unconscious. To see the sort of confusion that would be invited by this way of speaking, it helps to know that each eye feeds information to a separate set of ocular dominance columns, and each of the two sets computes its own edge map. To the extent that your two sets of ocular dominance columns (the one dominated by your left eye and the one dominated by your right eye) can be described as "thinking" or "believing" anything at all, their "beliefs" sometimes contradict each other. In such a case, there is no good reason to identify one of these "beliefs," rather than the other, as what *you* unconsciously believe.

But the real problem with characterizing a state of a set of ocular dominance columns as one of *your* beliefs does not lie in the fact that you have two sets of ocular dominance columns. It lies, rather, in the fact that a set of ocular dominance columns is not the sort of thing that can have a full-blooded belief. If we choose to speak of a set of ocular dominance columns as having beliefs at all, the content of such

So, Harry's conscious opinions need not quite *contradict* his unconscious belief that no one could ever fall in love with him. Nevertheless, there *would* seem to be something wrong with Harry's saying, "I unconsciously believe that no one could ever fall in love with me; moreover, no one could ever fall in love with me." If Harry is willing to assert that no one could ever fall in love with him, then it's not right for him to say that he believes this unconsciously. We might think of this point in connection with Moore's paradox. Moore pointed out that it would be absurd for someone to utter a sentence of the form, "I believe that *p*, and it is not the case that *p*." A number of writers have since noted that Moore's point does not hold for self-ascriptions of unconscious belief. In other words, there's nothing wrong with saying, "I *unconsciously* believe that *p*, and it is not the case that *p*." I'm calling something further to your attention: that, *prima facie*, there does seem to be something wrong with saying, "I unconsciously believe that *p*, and it *is* the case that *p*" (even though, as with Moore's paradox, both conjuncts might be true).[4] When we consider unconscious mental states, we find not only the failure of Moore's paradox, but, as it were, the inversion of it. We might call this Eroom's paradox.

Although I won't call it a *constraint* on any satisfactory account of the difference between conscious and unconscious mentality, it does seem reasonable to expect that such an account would shed light on Eroom's paradox. In other words, it's reasonable to expect that if we come to understand what distinguishes unconscious states of mind from conscious ones, we'll also understand what would be wrong with saying, "I unconsciously believe that my brother has ruined my bid for reelection; moreover, he has ruined it." And a reason to be suspicious of the not-so-simple view is that it doesn't look like it will help us come to grips with what would be problematic in such an utterance. According to the not-so-simple view, a belief's being unconscious lies in the fact that the subject doesn't know about it via a particular cognitive mechanism. But now, while there *would* seem to be something wrong with most imaginable utterances of the form, "I unconsciously believe that *p*; moreover, *p*," it's not at all clear what, if anything, would be wrong with someone's saying, "My knowledge that I believe *p* is not based on cognitive mechanism M; moreover *p*." Thus, the not-so-simple view leaves Eroom's paradox looking like a mystery. Again, I don't claim that this constitutes anything like a decisive objection to the not-so-simple view. Nonetheless, I do think it provides us with a reason to seek another account of unconscious mentality—one that helps us to make sense of Eroom's paradox.

According to both the very simple and the not-so-simple views, *unconscious* and *conscious* are, as it were, epistemic notions: on either view, to say that a mental state is unconscious is to say that the subject lacks some sort of *knowledge* that he would enjoy were the mental state conscious. Perhaps the difficulties that I have raised for the two views suggest that this is not the best way to think about the difference between conscious and unconscious mental states. In what follows, I'll try to show that there is a better way to think about this difference—one according to which *conscious* and *unconscious* are not epistemic notions.

Moore's paradox indicates that our statements about the world and our self-ascriptions of conscious belief hang together in a particular way. Eroom's paradox indicates that our statements about the

fear of abandonment is conscious. Describing Harry as "conscious of unconsciously believing that no one could ever fall in love with him" (or "conscious of his unconscious belief that no one could ever fall in love with him") seems to involve a contradiction only if we confuse two senses of the word "conscious." It's one thing to be consciously angry or jealous or believing such-and-such and quite another to be conscious of one's own anger or jealousy or belief. We can think of this fact as providing us with our first constraint on an adequate account of the distinction between conscious and unconscious mentality.

> Constraint 1: An account of the distinction between conscious and unconscious mentality should respect the difference between:
>
> (a) someone's consciously believing or being consciously afraid, i.e., a belief or a fear's being conscious, rather than unconscious;
>
> AND
>
> (b) someone's being conscious *of* her own belief or fear, i.e., conscious *that* she believes or fears such-and-such.

A variation on the very simple view might appear to meet Constraint 1. The view that I have in mind could be expressed as follows: "The problem with the very simple view is that it doesn't take into account the fact that there are various *kinds* of knowledge. If my knowledge that I believe *p* is based only upon the testimony of my therapist, then—while I may be said to be conscious *of* my belief that *p*—I cannot be said to consciously believe *p*. The kind of self-knowledge that the word 'consciously' picks out is not knowledge by *testimony*. What it means for me to consciously believe that *p* (or to be consciously hopeful or afraid, etc.) is that I know my mental state via a *particular* cognitive mechanism—the mechanism by which I ordinarily find out about my own states of mind. It is tempting to refer to this mechanism as 'inner sense,' but perhaps we should just call it 'mechanism M.' What it means for one of my mental states to be conscious (i.e., for me to be consciously angry, sad, or whatever) is that I'm aware of it *via mechanism M*. What it means for one of my mental states to be unconscious is that—although I may be aware of it—I am not aware of it via mechanism M." I'll call this the not-so-simple view.

It is a good deal more difficult to state a decisive objection to the not-so-simple view than to its very simple progenitor. I present what, I think, amounts to such an objection elsewhere,[3] but the argument is too lengthy to rehearse here. In the next couple of paragraphs, I'll merely offer a reason to be *suspicious* of the not-so-simple view—a reason that suggests the direction that I'll be taking later in the paper. Let us return to the case of Harry who consciously believes that someone will eventually fall in love with him, even though he is aware of his unconscious belief that no one could. The fact that, in this example, Harry's conscious and unconscious beliefs contradict each other helps us to see the inadequacy of the very simple view. But the need to distinguish between what someone is aware of believing and what he consciously believes does not depend on there being a flat-out contradiction between his conscious and unconscious beliefs. To see this, consider a variation on the example. Imagine that Harry says: "I unconsciously believe that no one could ever fall in love with me," whereupon he's asked whether anyone *could* fall in love with him. He answers, "Maybe; I'm not sure." Here too, Harry's belief that no one could fall in love with him is an unconscious one *of* which he is aware.

assuming that dreams are the royal road to the unconscious or that little boys are unconsciously afraid of castration or anything else that is distinctively Freudian or psychoanalytic. We should remember that Freud was not the first to speak of unconscious states of mind (as is evidenced not only by such stories as "The Strange Case of Dr Jekyll and Mr. Hyde"—which was first published in 1886—but also by 19th century theoretical writings about the nature of the human mind[1]).

I. The Very Simple and Not-so-simple Views

Let's begin by considering a very simple way of understanding what it means to say that someone's mental state is either conscious or unconscious. The view might be put as follows: "Your mental state is conscious if you know that you are in it. Your mental state is unconscious if you don't know that you're in it. To say that you, for example, unconsciously believe that no one could ever fall in love with you is to say: (1) that you believe that no one could ever fall in love with you and (2) that you don't know—you're unaware of the fact—that you believe this. Call this the very simple view.[2]

A bit of reflection reveals that the very simple view is unsatisfactory. Imagine someone—call him Harry—who says: "My therapist tells me that I unconsciously believe no one could ever fall in love with me, and she's generally right about such things, so I suppose I must have this belief." Let's imagine that Harry's therapist *is* right about him, and that Harry is justified in believing that she's right about him. Harry is, then, aware of his belief that no one could ever fall in love with him. According to the very simple view, we should say that Harry's belief that no one could ever fall in love with him went from being unconscious to being conscious as a result of his conversation with his therapist. But we can imagine that Harry holds no such conscious belief. When asked about his future, Harry says, "Oh, I'm sure that eventually someone will fall in love with me, even though my therapist has convinced me that unconsciously I believe it's impossible that anyone should." This is a perfectly intelligible remark. What Harry's therapist has made him aware of is not that he is in some way unlovable, but only that he unconsciously believes this to be so. The very simple view, however, cannot allow for the intelligibility of what Harry wants to say about himself. According to the very simple view, if someone is aware that he believes such-and-such, then his belief is conscious. But, in our example, Harry is aware of his unconscious belief that no one could fall in love with him.

The very simple view is *too* simple; it's not faithful to the way we use the words "conscious" and "unconscious." There is, however, a use of these words to which something like the very simple view *is* faithful. The case of Harry demonstrates that there is a distinction to be drawn between two uses of the word "conscious" or "unconscious." First, there is a relatively unpuzzling use that is generally in place when the word "conscious" or "unconscious" is followed by the word "of" or "that." I might say, "Until the lights came on, I had been unconscious *of* the person in the seat next to me." Or: "Lois suddenly became conscious *that* she was the only patent attorney in the room." In such contexts, the word "conscious" means, roughly, *aware*. The corresponding use of "unconscious" means *unaware*. Among the things I might become aware, or conscious, *of* are my own states of mind. But to say that I am conscious of, for example, my fear of abandonment is not to say either that I consciously fear abandonment or—what amounts to the same thing—that my

American Philosophical Quarterly
Volume 36, Number 2, April 1999

ON THE DISTINCTION BETWEEN CONSCIOUS AND UNCONSCIOUS STATES OF MIND

David H. Finkelstein

> Yes, I preferred the elderly and discontented doctor, surrounded by friends and cherishing honest hopes; and bade a farewell to the liberty, the comparative youth, the light step, leaping impulses and secret pleasures, that I had enjoyed in the disguise of Hyde. I made this choice perhaps with some unconscious reservation, for I neither gave up the house in Soho, nor destroyed the clothes of Edward Hyde, which still lay ready in my cabinet.
>
> —Robert Louis Stevenson, "The Strange Case of Dr. Jekyll and Mr. Hyde"

In this passage from Stevenson's famous story, Dr. Jekyll recalls a choice that he made but failed to live by—a resolution to never again transform himself into Edward Hyde. Jekyll remarks that he "made this choice perhaps with some unconscious reservation." What work does the word "unconscious" do in this sentence? When is a reservation (or an intention or a fear) rightly said to be unconscious? My aim in what follows is to explain what it is that distinguishes an unconscious state of mind from a conscious one. My procedure will be as follows. I am going to present several tempting, but ultimately unsatisfactory, views concerning that by virtue of which a state of mind should be characterized as either unconscious or conscious. My criticisms of these views will yield a set of conditions that an adequate account of the distinction between conscious and unconscious mentality ought to meet. I'll then offer what I take to be such an account, and I'll show that it both meets the conditions of adequacy and helps us to answer a number of questions that would appear puzzling without it.

Perhaps I should add that my aim is not to try to convince a reader who is uncomfortable with talk about unconscious states of mind that it's *all right* to describe a person as, for example, unconsciously jealous of his brother. We do characterize people as subject to a wide range of unconscious mental states, and I'm going to take it for granted that we are often justified in doing so. I aim to elucidate what we mean by such characterizations. I will not, however, be

"beliefs" is extremely thin. In striate visual cortex, there aren't even any object-representations. A set of ocular dominance columns has no idea, if you will, that what it's mapping are the edges of *objects*.[8] Your beliefs, both conscious and unconscious, about edges (or anything else) make sense in light of what you know about the world, which is a great deal. The "beliefs" of a set of ocular dominance columns are not your beliefs.

The "beliefs" of a set of ocular dominance columns could be compared to the states of an immune system. An immunologist might say that what causes juvenile diabetes is the immune system's mistakenly "thinking" that the insulin-producing cells in the pancreas are viruses. The "beliefs" of a set of ocular dominance columns are no more *yours* than are the "beliefs" of your immune system.

Indeed, the states of a set of ocular dominance columns, like those of an immune system, may be said to *be* beliefs—even beliefs with very thin content—only in the most metaphorical of senses. This is connected to the fact that we feel no inclination to ascribe fear, joy, or anger to a set of ocular dominance columns.[9] Genuine beliefs fit, in a way to be discussed shortly, into a coherent pattern of actions and other mental states of *various* types. Even a dog's beliefs make sense in light of his expressions of fear, anger, dissatisfaction, and joy. (Under the right circumstances, Fido's joyful barking might make manifest his belief that one of his friends is at the door.) The informational states that are ascribed to items pictured in cognitive psychological flow-charts do not have this character; they do not figure in a life in anything like the way that an organism's beliefs do.

The only lesson I would have you draw from this little discussion of cognitive neuropsychology is that we should take care to distinguish unconscious mental states from subpersonal informational states which (I want to say) are not conscious because they are not genuinely mental. But the themes of this discussion—that unconscious mental states are the states of a person and that they make sense only in light of a person's other attitudes and emotions—will figure prominently throughout the rest of the paper.

III. The Unconscious as a Quasi Person

If we accept that anything subject to full-blooded mental states must have the sort of complexity that a person has, we may be tempted to draw the conclusion that unconscious mental states should be thought of as belonging, not to a mere subpersonal information processor, but to an, as it were, inner person. A number of considerations reinforce this conclusion. Earlier, we saw that a person may be conscious *of* his belief that *p* without consciously believing *p*—indeed while consciously believing that not-*p*. How are we to make sense of the idea that I might be conscious of a mental state that I'm not consciously in? One familiar example of this phenomenon occurs when I attend to another person's state of mind. I may be conscious of *your* belief that juggling is a high art form without myself consciously believing this (or believing it at all). This suggests that we might make use of an interpersonal model in order to understand how it is that I may be conscious of my own unconscious belief that *p*. If we think of my unconscious mental states as, in some sense, the beliefs of a quasi-independent agent, then it might start to make sense that I may be conscious of my unconscious mental states.

This strategy for understanding unconscious mentality gains further appeal when

we notice that there are other ways in which my unconscious states of mind seem to be like the mental states of another person. First, I can think of my unconscious beliefs as false or crazy, just as I can think of your beliefs as false or crazy. And second, I don't speak with first-person authority about either your states of mind or about my unconscious states of mind.

Thus, for a number of reasons, it is tempting to think of my unconscious mental states as, in some sense, the mental states of another person. In a paper called "Freud and Moral Reflection," Richard Rorty credits Freud with discovering that unconscious mental states should be understood in this way—as the states of what Rorty calls a "quasi person."[10] I should say at the outset that I don't think the account of unconscious mentality that Rorty attributes to Freud is particularly faithful to Freud's writings. For this reason, I am going to refer to it as Rorty's account. The merit of Rorty's account is that he tries to take very seriously the idea that unconscious mental states are the states of something like a separate person. It will help us to see just what goes wrong when we try to see things this way.

According to Rorty, the difference between Freud and Hume is that Freud grasped the idea that the existence of an intentional state presupposes a more or less consistent network of other mental states. This led him to posit the existence of more than one coherent network of mental states within what we think of as a single person—different quasi persons who share a body. Rorty writes:

> The mechanization of the self that Hume suggested, and that associationist psychology developed, amounted to little more than a transposition into mentalistic terminology of a rather crude physiology of perception and memory. By contrast, Freud populated inner space not with analogues of Boylean corpuscles but with analogues of persons—internally coherent clusters of belief and desire. Each of these quasi persons is, in the Freudian picture, a part of a single unified *causal* network, but not of a single person (since the criterion for individuation of a person is a certain minimal coherence among its beliefs and desires). (F&MR, pp. 147–148)

Freud found that the propositional attitudes he wanted to attribute to people were wildly inconsistent with one another. He wanted to say that a boy may, for example, both believe and strongly disbelieve that his father is liable to castrate him. Freud accounted for this sort of inconsistency by distinguishing between conscious and unconscious attitudes. Rorty thinks that the way to understand what Freud was doing in so dividing up mental states is to see him as populating inner space with analogues of persons. According to this view, which Rorty himself endorses, a human being constitutes a single causal network, but two separate rational networks. Rorty writes, "The same human body can play host to two or more persons" (F&MR, p. 147).

Rorty credits Donald Davidson with inspiring this way of understanding what Freud had to teach us. In "Mental Events," Davidson writes:

> There is no assigning beliefs to a person one by one on the basis of his verbal behaviour, his choices, or other local signs no matter how plain and evident, for we make sense of particular beliefs only as they cohere with other beliefs, with preferences, with intentions, hopes, fears, expectations, and the rest. . . . The content of a propositional attitude derives from its place in the pattern.[11]

According to Davidson, propositional attitudes derive their contents from a *pattern* that is revealed when we interpret one another as agents.

What *kind* of a pattern is this? It's a *rational* pattern—a network of states and

events that owe their location in the network to their rational (rather than, e.g., spatial or causal) relations to one another. The items in such a pattern—the beliefs, preferences, hopes, fears, and actions of a person—hang together *rationally*. On Davidson's picture, a person's propositional attitudes and actions may be said to bear internal, rational relations to one another. A particular propositional attitude is what it is by virtue of its rational relations to other attitudes and actions.

Rorty takes this approach to understanding propositional attitudes for granted, only he understands a single human body to contain two independent networks of propositional attitudes—two separate quasi persons. Given Rorty's view, we might picture what we'd pretheoretically call "a human mind" as follows:

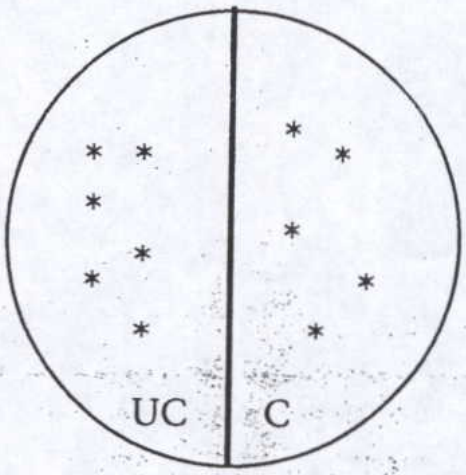

Figure 1

In this figure, C is the conscious mind, and UC is the unconscious mind. Stars represent mental states. C and UC are quasi persons, i.e., rationally coherent networks of mental states, separated in the diagram by a vertical partition line. We assign mental states to a quasi person in light of what we take to be *its* other mental states and *its* actions according to what Davidson calls "the constitutive ideal of rationality."[12] We interpret the actions and mental states of a quasi person holistically so as to render them rationally intelligible. The mental states of a single quasi person are thus understood to bear rational, internal relations to one another. What we *don't* do is treat the two quasi persons together—the human being as a whole—as a single rationally coherent agent. This is to say, we don't interpret the mental states on one side of the partition line so as to be rationally coherent with those on the other side. The relations between mental states on one side of the line and those on the other side are (like the relation between Smith's intention to hit Jones and Jones's subsequent feeling of pain) causal, but not internal. According to Rorty, the intentional states that we want to attribute to a human being don't exhibit the minimal coherence that is a criterion for the individuation of a single person.

Along with Rorty's picture of the unconscious as a quasi person, there is a corresponding picture of the sort of self-knowledge that can be gained through psychoanalysis. Rorty writes:

> Self-knowledge will be a matter of getting acquainted with one or more crazy quasi people, listening to their crazy accounts of how things are, seeing why they hold the crazy views they do, and learning something from them. (F&MR, p. 150)

Rorty refers to this kind of self-knowledge as "the aim of psychoanalytic treatment" (F&MR, p. 150). The aim is, in other words, to learn the views of a quasi person with whom one shares one's body. He goes on to say:

> The point of psychoanalysis . . . is to find new self-descriptions whose adoption will enable one to alter one's behavior. Finding out the views of one's unconscious about

> one's past is a way of getting some additional suggestions about how to describe (and change) oneself in the future. (F&MR, p. 153)

One needn't know much about psychoanalysis to understand that something has gone wrong in these passages. If the point of undergoing psychoanalysis were to gather useful *suggestions* about how one might describe and change oneself, then analysis, at least traditional analysis, would not take the form that it does—with the analyst saying so little. Here Rorty might reply that an analysand's unconscious mind is liable to provide *better* suggestions than his analyst could; hence it makes sense for the analyst to keep most of her suggestions to herself. But this would be an odd thing to say in light of the passage quoted above in which Rorty describes the unconscious as providing "crazy accounts of how things are." Moreover, even if Rorty represented the unconscious as a wise and sober quasi person, his position would miss the point of psychoanalysis. A distinction I discussed earlier, between being conscious *of* a mental state and being consciously in it, is here relevant. Rorty characterizes psychoanalysis as if its aim were to make the analysand conscious of what he unconsciously thinks and feels about things. In analysis, I (or anyway, my conscious mind) come to be aware of the thoughts and feelings of my unconscious mind via a funny sort of conversation. But becoming aware of, e.g., one's unconscious anger at one's mother represents a far more modest therapeutic goal than that of making the anger conscious. I might be made aware of the fact that I am unconsciously angry at my mother via the testimony of my therapist—or on Rorty's view, via the testimony of a quasi person inside my body. Either way, the result would be my becoming conscious of my unconscious anger. This is not the same as the anger's becoming conscious. Freud was concerned to give us a technique, not just for *finding out* about unconscious mental states, but for lifting repressions—a technique by which some of a person's unconscious mental states might be made conscious.[13]

Rorty's mischaracterization of the sort of self-knowledge that is the goal of psychoanalysis is not unrelated to his view of what unconscious mental states are. If we understand my unconscious mental states as the states of another person who shares this body with me, then it is natural to think that an unconscious mental state can come to consciousness only in the sense that I—that is, the conscious mind in this body—come to be aware of it.

Here, someone might reply that Rorty's basic position leaves open the possibility that my unconscious mind might do more than inform me of its views and feelings: "Let's say that you unconsciously believe that your father is insane. During the course of your analysis, your unconscious mind might *convince* you (i.e., convince your conscious mind) to believe this. In such a case, you would not merely come to be aware of what your unconscious mind thought about your father; you would come to believe the same thing consciously. Thus, given Rorty's basic framework, we *can* make sense of the fact that mental states go from being unconscious to being conscious."

But this attempt to save Rorty's view will not work. If my unconscious mind were to convince my conscious mind that my father was insane, then at the end of the day, both my unconscious mind *and* my conscious mind would believe that my father was insane. There would be two quasi people inside me who were in perfect agreement about my father. Given such a picture, it would make sense for me to say, "I both consciously and unconsciously believe

that my father is insane." But we don't talk this way about our beliefs. To say that someone unconsciously believes something is at least to say that she doesn't believe that very thing consciously.

When a mental state that was unconscious becomes conscious, it seems as if some sort of barrier is broken down or removed. A mental state that was in some sense *blocked* no longer is, and no mental state can be both blocked and unblocked. It will be a while before we understand the sense in which an unconscious mental state is blocked. But we can already see that Rorty's account fails to capture what might be called the *grammar* of our talk about an unconscious mental state's becoming conscious.

Here, let's set down a second constraint on an adequate account of the distinction between conscious and unconscious mentality.

> Constraint 2: An account of the distinction between conscious and unconscious mentality should help us to understand what it is for an unconscious mental state to become conscious.

While Rorty's account of unconscious mental states satisfies Constraint 1, it does not satisfy this second constraint.

I have been arguing that Rorty's account of unconscious mental states leaves no room for an adequate understanding of the sort of self-knowledge that psychoanalysis (or maturity) makes possible. But there is another sort of problem with his position. We sometimes *act* on our unconscious desires. Given Rorty's view, acting on an unconscious desire is tantamount to having one's body temporarily taken over by someone else. If the person writing these words is my conscious self, then my unconscious self is someone *else*. If my body acts on its (his?) intentions, *I* am not responsible and should not be held accountable. I should, rather, be pitied—as you would pity someone whose body had been inhabited by a demon.

But I do seem to be responsible, at least to some degree, when I act on my unconscious desires. Imagine that because I unconsciously want to harm my cousin Larry, I, as it were, "forget" to pick him up at the airport when he comes to town. If Larry were to find out that my stranding him at the airport was motivated by an unconscious desire to do him harm, he would not think: "Oh well, that's just my cousin's unconscious. My cousin isn't the slightest bit responsible. *He* means me no harm." No; Larry would blame *me* for leaving him at the airport. That is, he would blame the unitary person who unconsciously wanted to do him harm. (This is not to say that it would make no difference to Larry whether I was acting on a conscious or an unconscious desire to do him harm. The point is not that we take people to be *as* responsible for their behavior—or responsible in just the same way—when they act on unconscious desires as when they act on conscious ones. It is, rather, that we don't respond to someone whom we take to have acted on an unconscious desire as if he were split into two people, only one of whom had acted at all.)

On Rorty's view, my body is shared by two independent quasi people, and there is no, as it were, *overall* person of whom these two quasi people are parts. If I suffered from multiple personality disorder, this might be an appropriate way to describe me. A multiple's various selves really can seem to be so separate from one another that it doesn't make sense to blame one of them for the doings of another. But this is not how it is with me. When I act on my unconscious desires, the actions are mine in a way that Eve Black's actions were not Eve White's. This suggests a third constraint on an adequate account of the

distinction between conscious and unconscious mentality:

> Constraint 3: An account of the distinction between conscious and unconscious mentality should represent my unconscious mental states as *my* mental states.

Rorty doesn't provide such an account; he represents my unconscious mental states as belonging to someone else who happens to share a body with me.

Let us consider one last difficulty for Rorty's view, a difficulty that will help steer us toward a better view. Imagine someone—call her Jill—who consciously believes that her father is rather fond of her husband Jack, while unconsciously believing the truth: that her father dislikes Jack intensely. Now, it makes sense to attribute either one of these beliefs—indeed *any* belief—to Jill only against the backdrop of *other* beliefs that rationally cohere with it. Rorty is well aware of this.[14] But something that Rorty doesn't seem to notice is that even where a conscious and an unconscious belief are in direct contradiction—as they are in Jill's case—the two beliefs typically presuppose many of the *same* background beliefs.[15] In order for Jill to think either that her father is fond of Jack or that her father dislikes Jack, she must believe that her father is alive, that he knows of Jack's existence, etc. Rorty would have to say, "Jill's conscious mind and her unconscious mind *both* believe that her father is alive. Jill believes that he's alive both consciously and unconsciously." We've already seen that such a statement would be inconsistent with the ways in which we actually speak about conscious and unconscious beliefs. But there is a further point to be made here. Imagine that, in order to correct someone who is under the misapprehension that her father has recently died, Jill says, "My father is still alive." On Rorty's view, there should be a real question about *who* is speaking, i.e. about which quasi person in Jill's body is expressing its belief. In such a circumstance, however, there simply would be no question but that Jill's remark was an expression of what she consciously believed

IV. A Variation on Rorty's View

We can use the case of Jill in order to introduce a variation on Rorty's view of unconscious mental states. On Rorty's view, Jill's mind could be represented as shown in Figure 2 on the following page. Both Jill's unconscious mind and her conscious mind hold the belief that Jill's father is alive. Given Rorty's conception of unconscious mentality, this belief needs to appear on both sides of the partition line.

Now, consider a different picture of Jill's mind: Figure 3 on the following page. Here, the partition line is shortened, and there are stars immediately on either side of it. There are other stars around the periphery of the circle. The stars on the immediate left of the partition line represent unconscious mental states. All other stars in the figure represent conscious mental states. Between a mental state pictured on the immediate left of the partition line and a mental state on the immediate right of the line there are no internal, rational relations. We don't make sense of such mental states in light of each other as we normally make sense of a person's mental states in light of his other mental states. We don't try to make mental states out to be consistent across the partition line. Relations across the partition line are, as it were, merely causal.

So far, all that's been said about the partition line in this figure could be said about the line that separates Rorty's two quasi persons in Figure 2. The partition line in Figure 3 is different from the one in Figure 2 in that the items immediately on

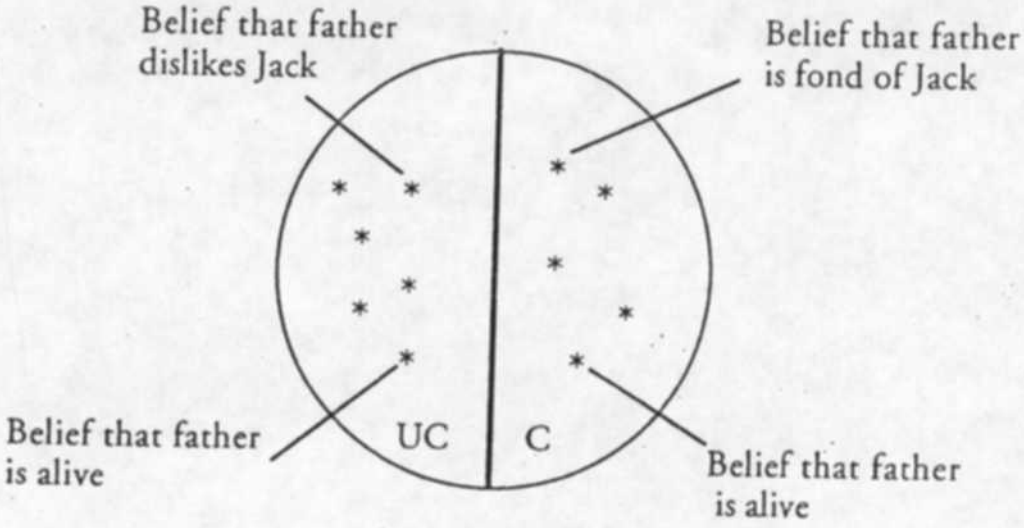

Figure 2

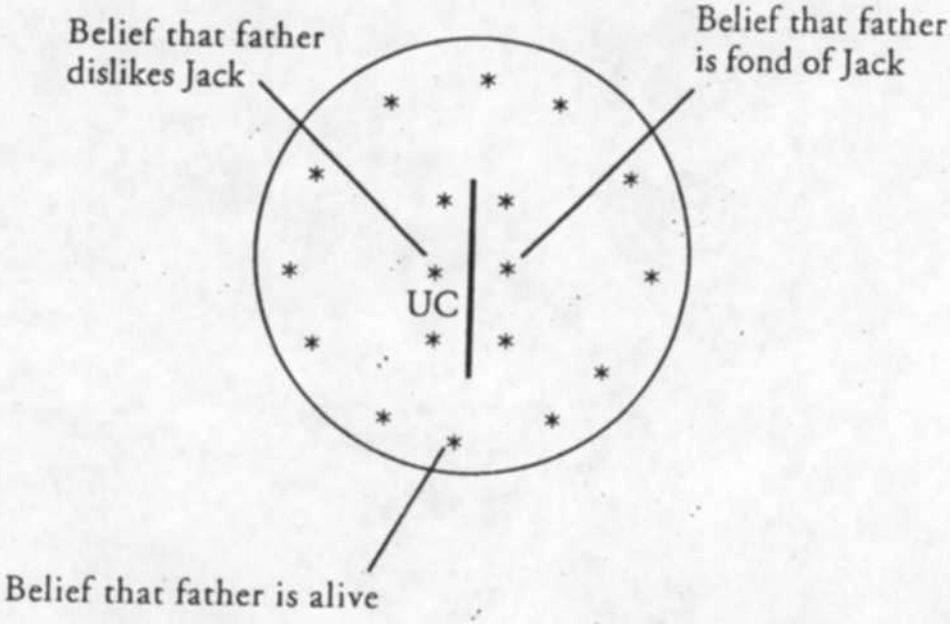

Figure 3

either side of it share a background of mental states. The stars around the periphery of the circle represent conscious mental states that bear internal, rational relations to mental states on *both* sides of the partition line. Thus, Jill's belief that her father is alive need not be represented twice in Figure 3. This belief is part of the background against which we make sense both of her conscious belief that her father is fond of Jack and her more accurate unconscious belief that her father dislikes Jack. Jill's belief that her father is alive need only be represented once because Figure 3 doesn't represent Jill as divided into two completely separate persons. Overall, Jill is represented as exhibiting a fair degree of rational coherence, even though there is, in her mind, a local region of incoherence. Figure 3 represents a unitary person, a person with certain mental states that are rationally cut off from certain other mental states, but a unitary person nonetheless.

Obviously, in interpreting Figure 3, I have been, in effect, suggesting a way of refining Rorty's story about unconscious mental states. None of the objections that I raised earlier against Rorty's view cut against this refinement of it. Nonetheless, it would be a mistake to claim that Figure 3 satisfactorily elucidates the distinction between conscious and unconscious mentality. The figure fails us at a critical point: nothing in it makes clear by virtue of what the stars on the immediate left of the partition line—rather than, say, those on the immediate right—represent *unconscious* mental states. Figure 3 illustrates how a single person can be understood to believe both that her father is fond of her husband and that her father dislikes her husband. The figure does not show why we should think of either of these beliefs as unconscious rather than conscious. Useful as Figure 3 might be for representing irrationality, it provides no explanation of what distinguishes unconscious mental states from conscious ones. Figure 3 will prove, however, to be a step in the direction of such an explanation.

V. Expression and Self-ascription

Rorty follows Davidson in being sensitive to the way our talk about a person's psychology makes sense only as the propositional attitudes of a person are understood to hang together rationally. Now according to Davidson, it's not only the *propositional attitudes* of a person that must be seen as hanging together rationally if we are to understand her as having a mind at all. Our *actions* must hang together rationally with our beliefs and desires. In order to understand someone as acting at all, or as having propositional attitudes at all, we need to see her actions as making sense in light of her beliefs and desires. Our beliefs and desires bear internal relations, not only to each other, but to the actions that they rationalize.

Thus, according to Davidson and Rorty, if we are to see a thing as a person, we must view it as having beliefs, desires, and actions that hang together rationally. But what makes the story of a thing intelligible as the story of a person is not merely that the thing has a set of propositional attitudes and actions that can be understood to be rationally coherent. Davidson tends to focus on *one* way in which mental states and behavior—the inner and the outer—must hang together if we're to have mental states in view at all: actions must be rationalized by belief-desire pairs. But mental states and behavior need to hang together in ways that are not captured by the thought that our attitudes rationalize our actions.

Here, we should consider the notion of expression. Our actions are said to express the beliefs and desires that rationalize

them. If a man runs after a bus because he doesn't want to be late for work, then his running expresses his desire to be on time. But now notice that, often, a mental state is said to be expressed by a bit of behavior that it does not rationalize. There are expressive relations between the inner and the outer that are not reason-relations. To take a simple example, when someone expresses his joy by smiling, his smile is not rationalized by his joy. Given the way that Davidson understands reasons, smiling isn't something that one typically does for any reason. Still, someone's joy and his expressions of joy—like his reasons and the actions that express them—make sense together, in light of one another. The rational relation that, as Davidson points out, obtains between reasons and actions is a special case of what we might think of as a more generic internal relation—a hanging-together relation—that obtains between mental states and behavior. The word "expression" picks out this more generic internal relation.

Consider another example of a bit of behavior that expresses a state of mind but is not rationalized by it. A writer gets frustrated with her work and decides to get away from her computer for a while. She leaves her apartment and, in leaving, slams the door. Her door closing is an intentional action. It expresses her desire that her apartment not be robbed while she is gone—a desire which, along with certain beliefs about her neighborhood, rationalizes the action. But the *way* she closes the door, her *slamming* it, expresses something else: her frustration at her work. Her door slamming isn't *rationalized* by her frustration, but it expresses it nonetheless. When we express our states of mind, we make them manifest in behavior to which they are internally, though not always rationally, related.

I have been talking about expressive relations that are not reason-relations. The examples considered thus far have involved mental states that are expressed non-linguistically. Of course, we often express what we think and feel in what we *say*. I might express my anger at someone by insulting him, or my gratitude by thanking him. Now, one kind of linguistic expression takes the form of self-ascription or avowal. I say, "I'm angry at you," and in so doing, I express my anger. It is a distinctive feature of mental state avowals that they allow a person both to say that he is in a certain mental state and to express that mental state. When I say that someone else is happy, I don't express the happiness. When I smile, I express my happiness, but I don't say anything about it. When I avow that I'm happy, however, I both say that I'm happy and, thereby, express my happiness.[16]

Although mental state self-ascriptions typically express that which they are ascriptions of, they don't always. If Harry says, "My therapist has convinced me; I'm unconsciously angry at my mother," he does not, thereby, express his anger at his mother. He expresses his *belief* that he's angry at his mother, but he doesn't express his anger. This is not to say that Harry's unconscious anger at his mother goes completely unexpressed; his therapist probably would not have come to the conclusion that he was angry at his mother unless he occasionally expressed his anger in one way or another, e.g., by refusing to visit her. But while Harry is able to express his unconscious anger, he is unable to express it simply by ascribing it to himself.

I want to claim that it's a defining characteristic of our unconscious mental states that we lack the ability to express them simply by self-ascribing them. Like all mental states, the unconscious ones may

be expressed in our behavior. But what is *distinctive* about unconscious mental states is that we're unable to express them simply by self-ascribing them. If Jill unconsciously believes that her father doesn't like her husband, she might express this belief in any number of ways. But not by saying, "I believe that my father doesn't like my husband." Jill *might* utter these words. She might say: "Well you've convinced me. The only way to make sense of my behavior is by taking me to have this crazy unconscious belief. Unconsciously, I believe that my father doesn't like my husband." Here, Jill expresses her opinion that she has a particular unconscious belief, but she doesn't express the unconscious belief; she doesn't express the belief that her father dislikes her husband. (Indeed, she expresses the opposite opinion.) The point may be put as follows: *Someone's mental state is conscious if she has an ability to express it simply by self-ascribing it. If she lacks such an ability with respect to one of her mental states, it is unconscious.*

Of course, what this means for our understanding of the distinction between conscious and unconscious mentality depends on how we think about expression and expressive abilities. Part of what I've been trying to do in this paper (especially in this section, but really since I began talking about rational, internal relations between attitudes and actions in §III) is to provide an elucidation—or, at least, the beginning of an elucidation—of the notion of expression.[17,18] Such an elucidation together with the point that I put in italics at the end of the last paragraph may be said to constitute an account of the distinction between conscious and unconscious mental states.

We are now in a position to address a number of issues that arose earlier. First, what accounts for Eroom's paradox? What is the problem with an utterance of the form, "*P*; moreover, I unconsciously believe that *p*." On my view, a mental state's being unconscious lies in a subject's lacking the ability to express it simply by self-ascribing it. Thus, if Jill were to say, "My father is alive; moreover, I unconsciously believe that he's alive," her utterance would be true just in case her father were alive *and* she lacked the ability to express her belief that he's alive by self-ascribing it. What would be strange about such an utterance is that anyone who is in a position to sincerely assert that her father is alive should be able to express her belief that he was alive by self-ascribing it. If a person lacks the ability to express her belief that *p* by self-ascribing it, then she cannot sincerely assert that *p* either.

Second, in what sense is your relation to one of your own unconscious mental states like your relation to the mental state of another? According to the view recommended here, the answer lies in this: Whether you say that *another person* is angry or that you are *unconsciously* angry, you are engaged in, as it were, mere description; you do not thereby express the anger about which you are talking.

Earlier, I noted that we don't speak with first-person authority about our own unconscious mental states. I should point out that what I've offered here is the beginning of an explanation of first-person authority. A central feature of the phenomenon of first-person authority is that we seem able to responsibly ascribe conscious mental states to ourselves without needing any evidence in support of the ascriptions. The explanation of this lies in the fact that there is an important respect in which a typical self-ascription of, for example, happiness is like a smile. Just as you might smile and thereby express your happiness

without needing any evidence in support of the claim that you are happy, you can *say*, "I'm so happy," and thereby express your happiness without needing any such evidence. By contrast, when you report someone else's mental state, or your own unconscious mental state, you don't express the mental state in question, so you require evidence supporting your claim about it; you don't speak with first-person authority.[19]

Finally, the above considerations put us in a position to address an issue that did *not* come up earlier, namely, why it is that the conscious/unconscious distinction doesn't seem to get a foothold when we are talking about the mental states of nonlinguistic animals. We often speak about, e.g., what our dogs want and believe, but we don't characterize these attitudes as either conscious or unconscious. We don't say things like, "Fido unconsciously wanted to go outside," or, "Fifi consciously believed that there was a squirrel in that tree." This, I think, is because expression in the form of mental state self-ascriptions is never an option for dogs. It is only in the context of a linguistic life that it makes sense to distinguish attitudes and emotions according to whether or not they can be expressed in self-ascriptions of them.

Let us now consider a couple of objections. Responding to these will help to clarify the main point.

Objection 1: "Imagine that I occasionally express my unconscious anger at my father by speaking in a peculiar, clipped tone of voice. One day, while speaking in this tone of voice, I say, 'My therapist tells me that I'm unconsciously angry at my father, and I suppose she must be right.' Through my tone of voice, I express my anger at my father in a self-ascription of it, even though the anger is unconscious. What this case shows is that, *pace* the view set out here, having an ability to express one's state of mind by self-ascribing it is not a sufficient condition for the state's being conscious. And lacking such an ability is not a necessary condition for the state's being unconscious."

Reply: I said that someone's state of mind is conscious if and only if she has an ability to express it simply by ascribing it to herself. The sort of ability that's at issue is one that enables a person to express her state of mind in a self-ascription of it, where what matters—what carries the expressive force—isn't her tone of voice (or whether she is tapping her foot, or what she is wearing, or to whom she happens to be speaking), but simply the fact that she is giving voice to her sincere judgment about her own state of mind. That someone might manage to express her anger in a self-ascription of it via a clipped tone of voice doesn't show that she has the relevant sort of expressive ability. When I am consciously angry, I can say in a neutral tone of voice, "I'm furious," and thereby express my state of mind.

Objection 2: "Imagine someone—call him Harpo—who is consciously angry at his brother but has a phobia about avowing anger; he cannot bring himself to say aloud that he's angry, even when he is alone. He won't *write* that he's angry either. If he is asked whether he's angry, he'll deny it. Although Harpo is consciously angry, he is unable to express his anger in a self-ascription of it. That we can imagine such a person demonstrates that having the ability to express a mental state simply by self-ascribing it is not a necessary condition for the state's being conscious. And the absence of this sort of ability is not a sufficient condition for the state's being unconscious."

Reply: A person may be said to have an ability, even though she is, for some reason, prevented from exercising it at a particular moment. A major league pitcher

might have the ability to throw a 90 m.p.h. pitch even though he is, at the moment, handcuffed, and so not in a position to do so. A tenor who has the ability to hit a high C might not do so for a period of time because of a neurotic fear that should he hit a note above A, he'll die on the spot. Such a person would have an ability to hit a C even while he was afraid to exercise it.

One way to answer the Harpo objection would be to say that, as Harpo has been described, he does have the ability to express his anger merely by self-ascribing it, only he is afraid to exercise it. To thus imagine Harpo is different from imagining him without this ability. To see this, let us imagine that in addition to his conscious anger at his brother, Harpo is unconsciously angry at his mother. We can understand the difference between the character of Harpo's anger toward his brother and that of his anger toward his mother to lie in this: If Harpo didn't have a phobia about admitting that he was angry—if he were happy to engage in sincere discussions concerning the whole range of his feelings—he could express his anger at his brother merely by self-ascribing it, but he could not express his anger at his mother in the same way.

VI. Conclusion

Let us return to Figure 3 which, as I indicated earlier, might be used to illustrate certain forms of irrationality, but fails as an explanation of the distinction between conscious and unconscious mentality. We can now produce a more satisfactory figure. First, we need to add marks outside the circle to represent various forms of expressive behavior—behavior that expresses mental states which are represented by the stars inside the circle. Second, we need to shift the partition line to the periphery of the circle. In Figure 3, the partition line marks the absence of internal, rational relations between certain mental states and certain other mental states. In Figure 4, the partition line marks the absence of expressive relations between mental states and their self-ascriptions.

The fact that a mental state is partitioned off from expressive self-ascription need not mean that it cannot be self-ascribed at all. What it means for a star to lie behind the partition line in Figure 4 is that if the mental state it represents *is* self-ascribed, the self-ascription will not be an expression of that mental state. Figure 4 represents a person who lacks the ability to express some of his mental states in self-ascriptions of them.

In Figure 5, I have combined Figures 3 and 4 to produce a picture of Jill (the woman who consciously believes that her father is fond of her husband, Jack, while unconsciously believing that her father dislikes Jack). Figure 5 is meant to retain what is right in Figure 4, while also illustrating the fact that a person's unconscious mental states fail to hang together rationally with some of her conscious mental states. As they are represented in this figure, Jill's unconscious mental states are rationally related to one another. For this reason, it makes sense to think of the area between the two partition lines as constituting a kind of subdivision of mind—something we might call "*the* unconscious." Even so, the unconscious is not a person.

Let's now look again at our three constraints on an account of the distinction between conscious and unconscious mentality. First, such an account should respect the difference between someone's being consciously afraid (hopeful, expectant, angry, . . .) and someone's being conscious *of* her fear (hope, expectation, anger, . . .). On the account that I've suggested, when I am consciously afraid, I'm able to express my fear by self-ascribing it. I may be con-

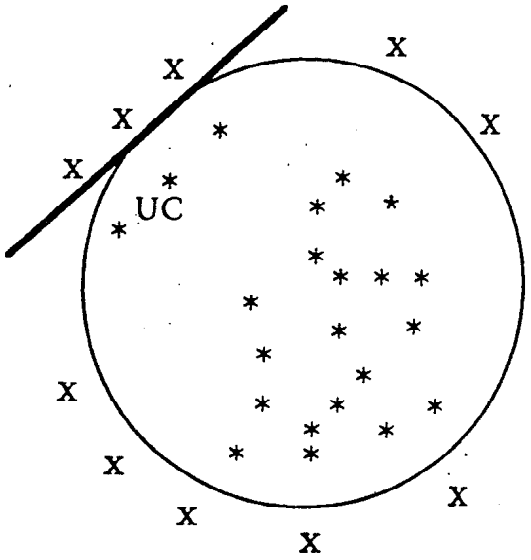

Figure 4

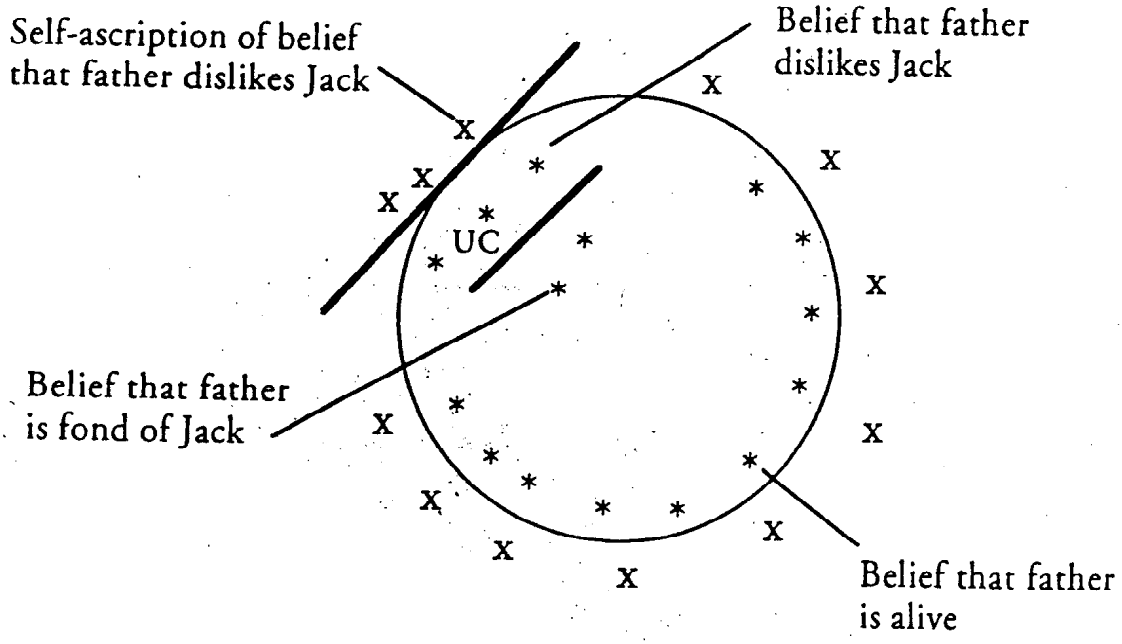

Figure 5

having this ability.

Second, an account of the distinction between conscious and unconscious mentality should help us to understand what it is for an unconscious mental state to become conscious. On my account, an unconscious mental state becomes conscious when a person who was once unable to express her mental state by self-ascribing it gains the ability to do so.

Finally, an account of the distinction between conscious and unconscious mentality should represent my unconscious mental states as *my* mental states. On the account I have suggested, this constraint is satisfied. My unconscious mental states are not represented either as the states of another person with whom I share a body, or as the states of a subpersonal component in my cognitive architecture. According to the view put forward here, unconscious attitudes and emotions are states of the unitary person who expresses them. *I* express my unconscious mental states in *my* actions—actions for which I am to a certain extent responsible. I lack the ability to express my unconscious mental states in self-ascriptions of them, but I express them nonetheless.

Freud writes:

> From what I have so far said a neurosis would seem to be the result of a kind of ignorance—a not knowing about mental events that one ought to know of. . . . Now it would as a rule be very easy for a doctor experienced in analysis to guess what mental impulses had remained unconscious in a particular patient. So it ought not to be very difficult, either, for him to restore the patient by communicating his knowledge to him and so remedying his ignorance. . . .
>
> If only that was how things happened! We came upon discoveries in this connection for which we were at first unprepared. *Knowledge is not always the same as knowledge*: are far from equivalent psychologically. . . . If the doctor transfers his knowledge to the patient as a piece of information, it has no result. . . . The patient knows after this what he did not know before—the sense of his symptom; yet he knows it just as little as, or he knows it just as little as he did. Thus we learn that there is more than one kind of ignorance.[20]

Freud expresses an important insight in this passage. It's one thing for me to know that I, for example, unconsciously fear my father and quite another thing for my fear to become conscious. This insight goes wanting in both the very simple view and in Rorty's view. What I've tried to do in this paper is to say how it is that "Knowledge is not always the same as knowledge." The sort of self-knowledge that is a goal of psychoanalytic treatment is, on my view, not really a kind of knowledge at all, but rather a certain sort of expressive ability. The important distinction to which Freud is calling our attention in this passage turns out to be a distinction between merely knowing that one is in a particular mental state and having the ability to express one's mental state simply by self-ascribing it.

Philosophers of mind have, on the whole, tended to view the distinction between conscious and unconscious states of mind as an epistemic matter—a matter of whether, or how, a subject knows something. Such views cannot do justice to the kind of rupture in a human life that is signaled by the word "unconscious." To have an unconscious mental state is not essentially a matter of being ignorant of something; it's a matter of being unable to *do* something—of being unable to express one's state of mind in a particular way.[21]

Indiana University

NOTES

1. See, e.g., Sir William Hamilton's *Lectures on Metaphysics and Logic*, vol. I (New York: Sheldon and Co., 1858), p. 242:

> [T]he supposition of an unconscious action or passion of the mind . . . has been gravely established as a conclusion which the phœnomena not only warrant, but enforce.

2. Colin McGinn endorses the very simple view when he writes:

> This . . . raises the interesting problem of what makes a propositional attitude unconscious. . . . For a desire (say) to be unconscious is for its possessor not to know or believe that he has that desire. ("Action and its Explanation," in *Philosophical Problems in Psychology*, ed. Neil Bolton [New York: Methuen & Co., 1979], p. 37)

3. In a paper called "Detection, Expression, and First-Person Authority," which I'm currently preparing for publication.

4. This point should not be overstated. It *is* possible to imagine a situation in which a sentence of the form, "I unconsciously believe that *p*; moreover *p*," would be intelligible. Imagine that Harry's mother, who is a terrible driver, drives Harry to work every morning. Harry wishes that he were able to drive himself because he thinks it's dangerous to be a passenger in his mother's car. He believes that his mother's inept driving is liable to result in his being injured or killed in a car accident. One day, Harry's psychoanalyst convinces him that he unconsciously believes that his mother means to murder him by poisoning his well water. In this situation, Harry could say, "I unconsciously believe that my mother is liable to harm me; moreover my mother *is* liable to harm me. I believe that she's liable to harm me both unconsciously and consciously."

In the face of this example, it is natural to respond that, even so, Harry doesn't really believe the same thing about his mother both consciously and unconsciously. His conscious belief is that she's liable to harm him *by involving him in a car accident* while his unconscious belief is that she's liable to harm him *by deliberately poisoning his well water*. Where it makes sense for someone to utter a sentence of the form, "I unconsciously believe that *p*; moreover *p*" (or "I believe that *p* both consciously *and* unconsciously"), we find that the contents of the conscious and unconscious beliefs in question can be further specified so as to display a difference between them. Notice that this is not true for sentences of the form, "My friend believes that *p*; moreover *p*," (or "My friend and I both believe that *p*"). My friend and I can be in, as it were, perfect agreement about *p*.

5. In his "Two Concepts of Consciousness," *Philosophical Studies* 49 (1986): 329–359, David M. Rosenthal writes: "Intuitively, a mental state's being conscious means just that it occurs in our stream of consciousness" (336). If this sentence is faithful to any non-philosophical use of the word "conscious," it's not a use with which I'm concerned in this paper. *Pace* Rosenthal, to characterize someone's mental state as conscious is not ordinarily to commit oneself to the claim that it somehow figures in her current stream of consciousness. To see this, consider a case in which a question arises concerning whether or not a particular mental state is conscious. Imagine that your cousin, Helen, says to you: "My son wants to kill me. Today, he tried to run me over with his car. I need to hide out at your house for a while." Later, speaking on the phone with a friend, you say, "Do you remember my cousin, Helen—the one with the devoted son, Roger? Now she thinks he's trying to kill her." Your friend replies, "Do you mean that now she *consciously* thinks Roger wants to do her in, or are you talking about some unconscious belief of hers?" In answering this question, you would not need to concern yourself with the question of what, if anything, was in Helen's stream of consciousness while you were speaking on the phone

with your friend. With Helen asleep on your couch, you could answer truthfully, "I'm talking about what she consciously believes now about Roger. Her conscious belief is that he wants her dead."

A related point: I won't be concerned in this paper with questions about what it's like to be this or that, or to be in this or that state. (I won't be worrying about what Ned Block calls "phenomenal consciousness." [See his "A Confusion about a Function of Consciousness," *Behavioral and Brain Sciences* 18 (1995): 227–287.])

6. See David Marr, *Vision* (New York: Freeman Press, 1982).

7. See David H. Hubel, *Eye, Brain, and Vision* (New York: W. H. Freeman and Co., 1988).

8. The drawings of edge maps that appear in cognitive psychology texts can, for this reason, be misleading. A drawing of an edge map looks like an object. In Marr's *Vision*, there's one that looks like a teddy bear. But as far as a set of ocular dominance columns is concerned, there aren't any teddy bears; there aren't even any objects with edges. (Thanks to Jacob Feldman for pointing this out to me.)

9. For an illuminating discussion of the way in which being subject to emotions distinguishes genuine subjects of belief from mere information processors, see Bennett Helm, "The Significance of Emotions," *American Philosophical Quarterly* 31, no. 4 (October 1994): 319–331.

10. Richard Rorty, "Freud and Moral Reflection" in *Essays on Heidegger and Others: Philosophical Papers*, vol. 2 (Cambridge: Cambridge University Press, 1991). Hereafter this will be cited as F&MR.

11. *Essays on Actions & Events* (Oxford: Oxford University Press, 1980), p. 221.

12. *Essays on Actions & Events*, p. 223.

13. This is not to say that *whenever* an analysand's suffering is due to an unconscious state of mind, the aim (or even, *an* aim) of psychoanalytic treatment is to make the state of mind conscious. If a patient suffers because he unconsciously believes some crazy claim (e.g., that if he were to succeed in business, he would thereby murder his father) the goal of analysis is not to bring him to consciously believe this claim (even for a little while). Lifting a repression is not *always* a matter of making an unconscious state of mind conscious. If an analysand unconsciously believes that *p*—where *p* is inconsistent with much of what he consciously believes—then lifting the repression associated with the belief that *p* will probably not result in a conscious belief that *p*.

14. He writes:

> One can only attribute a belief to something if one simultaneously attributes lots of other mostly true and mostly consistent beliefs. (F&MR, p. 147)

15. John Heil makes a similar point in his "Minds Divided," *Mind* 98 (October 1989): 571–583.

16. There is a tendency amongst analytic philosophers to assume that assertion and expression are, in a way, mutually exclusive, i.e., to assume that an assertion to the effect that the speaker is in a particular mental state cannot express that very mental state. This tendency (whose prevalence may be due, in part, to the influence of emotivism in ethics) can be seen in, e.g., the following from David M. Rosenthal's "Thinking that One Thinks" (in *Consciousness*, ed. M. Davies and G. W. Humphreys [Oxford: Basil Blackwell], 1993):

> I can communicate my suspicion that the door is open either by expressing my suspicion or by explicitly telling you about it. . . . In saying I suspect something, I report, rather than express, my suspicion. (200)

In his paper, "Expressing" (in *Philosophy in America*, ed. M. Black [Ithaca, N.Y.: Cornell University Press, 1967]), William P. Alston rightly rejects this assumption:

> I can express my enthusiasm for your plan just as well by saying "I'm very enthusiastic about your plan," as I can by saying "What a tremendous plan!" "Wonderful," or "Great!" I can express disgust at X just as well by saying "I'm disgusted," as by saying "How revolting!" or "Ugh." I can express approval as well by saying "I completely approve of what you are doing" as I can by saying "Swell," or "Good show." And I can express annoyance as well by saying "That annoys me no end" as by saying "Damn."
>
> This shows that expressing and asserting are not mutually exclusive in the way commonly supposed. (16)

17. I offer a more extended elucidation in "Wittgenstein's 'Plan for the treatment of psychological concepts'," which I'm currently preparing for publication.

18. I have discovered three other writers who point out that while it is possible for someone to ascribe an unconscious belief to himself, he does not thereby express the belief. These are Arthur Collins ("Unconscious Belief," *The Journal of Philosophy* 66, no. 20 [October 16, 1969]: 667–680), Georges Rey ("Toward a Computational Account of *Akrasia* and Self-Deception," in *Perspectives on Self-Deception*, ed. B. P. McLaughlin and A. O. Rorty [Berkelely: The University of California Press, 1988]), and Richard Moran ("Self-Knowledge: Discovery, Resolution, and Undoing," *European Journal of Philosophy* 5, no. 2 [August 1997]: 141–161). Although a discussion of the views set out in these papers will have to await another occasion, I'll say this for now: While I'm sympathetic with a good deal of what these writers have to say, it seems to me that all of them understand the notion of expression too much in terms of a distinctive feature of belief self-ascription, viz., that when someone expresses his belief by saying, "I believe that *p*," he thereby commits himself to the truth of the claim that *p*. While this is a noteworthy feature of *one* kind of expression of *one* kind of mental state, if we understand the very notion of expression too much in terms of it, we lose our grip on what unconscious beliefs have in common with unconscious fears, wishes, and revulsions.

19. In both "Wittgenstein's 'Plan'" and "Detection," I argue that first-person authority is not a kind of epistemic authority at all. It should not be understood as anything like the sort of authority with which an eye-witness speaks about what she has seen. We speak with authority about our conscious mental states, not because we *know* them so well, but because our self-ascriptions of them are expressions of them.

20. *Introductory Lectures on Psychoanalysis* (New York, W. W. Norton and Co: 1966), pp. 280–281, my emphasis.

21. I am very grateful to James Conant, Gary Ebbs, Martha Farah, Samantha Fenno, Kimberly Keller, Irad Kimhi, John McDowell, and Thomas Ricketts for helpful conversations about the material presented in this paper. In addition, I am indebted to Robert Almeder as well as to this paper's two anonymous referees for incisive comments on an earlier draft.