cation of the brain damage causing the epilepsy in the split brain patient (as well as the damage incurred during the operation), one might be able to study some functions located in a different part of the brain. To do that, one would have to show two things, first that the function being studied was indeed localized to the part(s) of the brain which are not damaged, and second that the brain damage had no effect on non-brain damaged tissue (the diaschisis phenomenon). These are challenging problems, but not insurmountable.

FIRST-PERSON ASPECTS OF AGENCY

Lynne Rudder Baker
Philosophy Department
Middlebury College
Middlebury, Vermont 05753

To understand human action, two paths need pursuit. One has had ample traffic; the other has been relatively neglected. The well-worn path deals with questions of the causal history of an action: Did it have its source, as Aristotle would say, in the agent? At the time of the action, was the agent constrained, coerced, under duress? In what sense could the agent have done otherwise? Principles based on answers to these questions would constitute a set of conditions which must obtain at t if a given change at t is to count as an action.

These questions presuppose familiarity with the less traveled path, the one which leads to determining the very conditions of agency. It is often assumed that agency can be understood in terms of agent causation.[1] My concern here is not with the causation of actions but with what underlies the ability to cause actions: What is required for an entity to be an agent at all? The claim is that the crucial condition for agency is self-consciousness. Increasing plausibility accrues to this claim as the role of self-consciousness in agency is explored in detail. Any satisfactory account of the causation or initiation of actions will have to include the first-person aspects of agency developed here.

This investigation is thus an effort to get behind questions such as Wittgenstein's: "What is left over if I subtract the fact that my arm goes up from the fact that I raise my arm?"[2] Wittgenstein's posing of this question takes for granted the issue dealt with here: What is it about the fact that we think of ourselves in the first-person that is relevant to our being agents?

An agent might be defined as a being which is capable of formulating intentions, based on certain kinds of beliefs and desires. Such a definition is almost empty, but not quite: it shows that agency is a matter of mental capacity, not a matter of sophistication of actual or potential behavior. Thus, an agent must have whatever conceptual apparatus necessary to formulate intentions. To speak of agency in terms of mental capacity is not to get the question against machines; if machines are to become agents, they must be "mental" in the sense developed here.

A conspicuous feature of intentions is that they are formulated in the first-person. If X intends to do A, then X endorses a thought of the form, "I am going to do A" or "I shall do A". Mental states of intending are rehearsed in conscious episodes, prominent among which are volitions. Volitions are rehearsals of intendings to do something here and now.[3] (E.g., "I shall now make reservations to go to Aruba.") First-person formulation is used not only for intentions but also for all kinds of practical reasoning: As an agent, I deliberate about my own future on the

basis of what I desire to do and what I believe, and what I take to be the circumstances that I am in. Therefore, to be an agent, an entity must be able to enjoy the first-person perspective.

The First-Person Perspective

The first-person perspective is intimately connected with the ability to make irreducible reference to oneself in the first person. Although language is not required for such self-reference, any language adequate for practical thinking must have a device for first-person reference. In English, we have the first-person pronouns, which serve the unique function of indicating the thinker or speaker without characterizing him in any way. That is, we use the first-person indicator to refer to ourselves qua ourselves, and not under any name or descriptions such as "the utterer of this sentence" or "the tallest woman in Nebraska" or "Edna Periwinkle". When a person thinks of herself in the first-person way, she is not thinking of someone-who-happens-to-be-herself (such as the tallest woman in Nebraska); nor is she thinking of herself as this-particular thing with a third-person demonstrative reference to herself. In short, thinking about oneself qua oneself cannot be reduced to thinking about oneself in any other way.[4]

With the ability to conceive of oneself qua oneself comes the ability to conceive of one's mental states as one's own. Such second-order consciousness is reminiscent of Kant's dictum, "The 'I think' must be capable of accompanying all my representations."[5] The ability to make irreducible first-person reference is clearly necessary for the ability to have second-order consciousness: if X lacks the first-person perspective, then X cannot conceive of his thoughts--or of anything else--as his own. That is, if X cannot make first-person reference, then X may be conscious of the contents of his own thoughts, but not conscious that they are his. In this case, X has no second-order consciousness. On the other hand, the ability to make first-person reference is sufficient for second-order consciousness in the case of beings which can conceive of thoughts at all: if X can conceive of himself from the first-person perspective, then if he can conceive of thoughts at all, he can be conscious that his thoughts are both thoughts and his own. Therefore, X is aware that his thoughts--regardless of the grammatical person of their expression--are his own if and only if X can make irreducible first-person reference.

The first-person perspective can come into play both in conceiving of properties of a thought qua mental episode ("It occurred to me today"; "I expressed it to myself in English," etc.) and in conceiving of properties of a thought qua representation of something else ("It was about myself"; "It was a wish that I liked Mint Juleps").[6] In deliberation, for example, self-reference is often manifested in both these ways. Say that Sal is deliberating about whether to go to the movies. Then she must conceive of her desire to go to the movies both as her own desire and as a desire about herself. To see that these two first-person elements are distinguishable, consider

(1) I want to go to the movies,

in which the first-person perspective has a double role: The desire is about the thinker, Sal, conceived in the first-person way, and the desire has the property of being a desire of hers. Now compare:

(2) I want you to go to the movies,

(3) He wants me to go to the movies.

No (1), (2), and (3) all require a first-person perspective on the part of the thinker. But in (2), the desire has the property of being the thinker's desire, although it is not a desire about the thinker; in (3), the desire is about the thinker, conceived in the first-person, but it does not have the property of being the thinker's desire. On the other hand, (1), (2), and (3) are all thoughts of Sal's. Since they all require for their expression the first-person pronoun to be used by Sal, they are all about Sal. We may generalize for any being X which can conceive of thoughts: Any thought about X, where X is necessarily conceived in the first-person, has the property of belonging to X, and X has the capacity to conceive of that thought as his own.

It should be clear that the first-person perspective, which is the capacity to make irreducible first-person reference, is the key to self-consciousness. The ability to make first-person reference is necessary for genuine self-belief as opposed to belief of a thing-which-is-in-fact-oneself. Likewise, any form of self-awareness requires that one be conscious of oneself qua oneself. Conversely, if an entity which can conceive of thoughts can make irreducible self-reference

and thus can conceive of his mental states as his own, he is thereby self-conscious. Thus, where X is an entity which can conceive of thoughts, we have the following principles connecting the ability to enjoy the first-person perspective to self-consciousness as a dispositional state:

(S1) X can conceive of a thought as his own if and only if X can make irreducible first-person reference.

(S2) X is self-conscious if and only if X can conceive of a thought as his own.

Therefore, it follows trivially that

(S3) X is self-conscious if and only if X can make irreducible first-person reference.

We need not be committed to the view that all thinking involves self-reference or that all consciousness is self-consciousness. Not only do we seem to have moments of being conscious without being self-conscious (e.g., in reverie), but it is possible that there be a conscious being which lacks the first-person perspective altogether. To such a being, there would occur thoughts, "There is a pain at place p and at time t;" "There are daisies at place p_1 and at time t_1".[7] Such thoughts essentially have no grammatical person: a being which could not refer to itself in the first-person could have no concept of "other-than-itself" either.

For example, an entity which would lack the first-person perspective--whether it is "conscious" in some sense or not--would be a computer, call it L. L may be able to scan states of itself and states of smaller computers, say, M and N. Although L cannot give a complete description of itself in the state of scanning M and N, a complete description of L would be available to some other scanner.[8] Now, L distinguishes itself, L, from M in exactly the same way that it distinguishes M from N. L's reports on itself, like its reports on M and N, are in the third person--e.g., "L is in state S"; or, if L is programmed to report "I am in state S," the first person report would be reducible to the corresponding third-person report: L cannot distinguish between "L is in state S" and "I am in state S" when the "I" refers to L. Lacking the conceptual apparatus to distinguish itself *qua* itself, L is unable to make irreducible first-person reference. In the absence of a capacity to conceive of itself indexically in the first-person, L cannot be said to have genuine self-belief. Moreover, vanity, pride, self-loathing, and all other attitudes which depend upon an understanding of oneself *qua* oneself are forever foreign to L--regardless of how "intelligent" L is.

To see how deeply the first-person perspective is embedded in intention, consider the ways an agent would formulate his intention to go home: He might formulate it as "I'm going to go home". He would not--and could not--formulate that very intention in the third person as, say, "Pete is going to go home". To see the nonequivalence of the first- and third-person formulations, we need only note that Pete's intention to go home in no way involves his even knowing any name or description of himself. His intention is about himself *qua* himself; he need not recognize himself under any name or description--e.g., he may have forgotten that his boyhood nickname was "Pete". Or again, knowing that the only person in Munroe Hall is going to go home is neither necessary nor sufficient for knowing that I am going to go home even though I am the only person in Munroe Hall. Therefore, the proposition expressed by "I intend to go home" is not equivalent to the proposition expressed by "The only person in Munroe Hall intends to go home". Thus, the first-person perspective is ineliminable in the formulation of intentions.

The first-person perspective is involved in yet another way in the attribution of intentions. If Pete attributes to himself an intention to go home, he might think, "I intend to go home." But this is elliptical for "I intend that I go home"--a sentence with two first-person references. Pete refers to himself in the first-person not only as the intender but also as the agent of the intended action; and the second first-person reference is no more eliminable than the first. Even a third-person attribution to Pete of an intention to go home implicitly attributes a first-person reference to Pete: "Pete intends to go home" is elliptical for "Pete intends that he (himself) go home,"[9] where the "he (himself)" attributes to Pete an irreducible first-person reference. That is, Pete would express his intention to go home in the first-person. Thus, the first-person perspective is fundamental to both the formulation and attribution of intentions, and is a central feature of agency.

A Minimal Agent

Although all intending is unavoidably tied to the first-person perspective, there still may be different levels of agency. Consideration of possible beings with limited kinds of agency will illuminate our concept of full agency. For example, it appears to be logically possible that there be an agent which can refer to itself in the irreducible first-person way, but which lacks the concepts of agency and intention. Such a being would in fact be able to endorse intentions such as "I am going to A," but would not be able to see itself as an agent. Or again, such a being could think endorsingly without being able to conceive of this thinking as an endorsement. In short, such a Minimal Agent would be able to intend but would lack the concept of intention.

Without attempting to give a full account of the criteria for saying "X has the concept of C," let me give what seem to me to be sufficient conditions for saying, "X has the concept of intention". X has the concept of intention if X can conceive of his own thinking endorsements of thoughts of the form "I shall do A" as causally efficacious. This is sufficient because causal efficacy is the heart of intending. What distinguishes intending from the fundamental contemplative state of believing is that intending is at least "the re-arrangement of the causal powers within oneself in the direction of the action one intends to do."[10] Thus, the ability to conceive of one's own endorsements of thoughts such as "I shall do A" as having causal powers guarantees that one has the concept of intention.

Notice that this criterion is not simply that X conceive of his mental states generally as having causal efficacy: for some mental states may have causal efficacy without being intentions. For example, thinking of the foul-tasting soup may cause Smith to grimace, but Smith's thought of the foul-tasting soup is certainly not an intention; and thus Smith's conceiving of the causal power of his thought of the foul-tasting soup would not show that Smith had a concept of intention. The criterion also is noncircular. Ex hypothesi, the Minimal Agent has thinking endorsements of the form, "I shall do A," where there is a causal thrust not only to the endorsement but also to the thought of the agent's linking himself practically to the action. Intentions have a causal dimension internal to them which thoughts of the foul-tasting soup lack. The claim here is that X has the concept of intention if he can conceive of those states which are in fact his intendings as causally efficacious.

The outlook ascribed to the Minimal Agent is one common enough on various occasions for normal agents. It is not unusual for us to act on intentions which we have endorsed without being conscious of ourselves as agents: Say that Max is scanning the desserts in the cafeteria, trying to resist the temptation to take one; he may not realize that he has endorsed any intention at all until he reaches out to pick up the largest piece of chocolate pie. But the question is not whether we can endorse thoughts of "I shall do A" without being conscious of ourselves as agents or intenders; it is rather stronger: the Minimal Agent endorses thoughts of "I shall do A" without ever being conscious of himself as an intender, without even having the concept of intention.

A Minimal Agent would be more like a very young child than like the weak-willed Max. Just as a child can have beliefs without knowing the concepts of belief or truth, so too a child might be able to formulate intentions and think endorsingly, "I shall do A" without having a concept of the causal power of his endorsement. And it is logically possible that the child's development be arrested and that he never come to have the concept of intention. The case of the child is problematic; it seems to me that we do not actually attribute agency in the full sense of the word to young children. Nevertheless, let us see what kind of entity a Minimal Agent would be.

For a Minimal Agent, call it MA, it must be possible that the following all be true: (a) MA thinks endorsingly a thought expressible as "I shall do A". (b) MA conceives of his mental state of endorsing the thought, which is in fact one of intending, as his own. (c) MA does not have the concept of intending, and hence cannot conceive of the thought as an intention.

Now let (a) and (b) be true. If MA's conceiving of his mental state of endorsing the thought includes a conception of his endorsement as causally efficacious, then (c) is false. This follows from the sufficient condition given above for X's having the concept of intention. So, say that MA's conceiving of his mental state can not include conceiving of it as causally efficacious. Now, if MA cannot conceive of his endorsings as causally efficacious, there is no difference from MA's point of view between endorsing and merely entertaining the thought, "I shall do A"; this is so because the content of the thought, "I shall do A," is the same in both cases.

Agents must be able to consider alternative courses of action without endorsing them, i.e., without initiating a causal sequence leading to their performance. But, as we have seen, MA cannot conceive of the difference between his mental states of intending to A (with their causal thrust) and his mental states of practically considering doing A. He lacks the conceptual machinery to formulate the thought, "Now I shall consider whether or not to do A."

An MA would differ from the normal agent in other far-reaching ways as well. Most important is this: One can intend to do only what one can conceive of; MA cannot conceive of intending. Therefore, MA cannot intend to come to intend to A. To see the depth of this deficiency, we need only note that MA could not be a Reformed Uncle Scrooge: Uncle Scrooge, known mostly for his wealth and meanness, decides to mend his ways in his old age. So he sets for himself a new policy: henceforth, he will do at least one charitable deed, unbregudgingly, a week. Uncle Scrooge's plan of action is schematic; at the time he endorses the plan as a whole, he does not foresee (much less endorse) the individual good deeds which he will do.

It is impossible for MA to be a Reformed Uncle Scrooge on two separate grounds: (i) Lacking the concept of intention, MA cannot try to change his intentions; for one cannot try to do what one cannot conceive of. Like Frankfurt's "wanton" who has desires but does not care which of his competing first-order desires dominates,[11] an MA would be a wanton agent. Unable to conceive of himself as an agent, MA could not care what intentions he has. (ii) More fundamentally, lacking the ability to have second-order intentions, MA could not endorse <u>any</u> action plan which called for subsequent decisions. The endorsement of such an action plan is in part the intention to come to have appropriate intentions later. But, of course, MA cannot intend to come to intend later; for one cannot intend to do what one cannot conceive of.[12]

Whether or not we say that a being with the abilities and characteristics of an MA is really an agent is mostly a verbal matter at this point. The important thing is to see what such a being would and would not be able to do. In summary, there seem to be two large areas in which the MA is deficient: first, he cannot conceive of the difference between intending to do A and practically considering doing A; and, second, he cannot have second-order intentions, and hence he can neither try to change his character nor endorse action plans which require later intentions.

A Purely Contemplative Being

The practical thinking enjoyed by agents is often contrasted with theoretical or purely contemplative thinking. It seems possible that there be a self-conscious being who could have beliefs, and generally engage in the activities of contemplative thinking without having the ability to entertain or endorse any thoughts of the form "I shall do A". If such a being--call it a Contemplative Nonagent, or CN--is possible, then self-consciousness does not suffice for agency. But the kind of being that a Contemplative Nonagent would be shows that self-consciousness does not suffice for contemplative thinking either.

Although intending is the mechanism by which we initiate changes in the physical world, it plays a significant role in our purely mental lives as well. A CN, without the ability to formulate intentions, would not resemble a person whose life was confined to a sensory deprivation chamber so much as a person whose thoughts passed before him unbidden, as if on a screen. Consider the related ways in which the mental life of pure reasoning of a CN would be attenuated:

1. A CN could not direct his own thoughts. Although he could recognize his thoughts as his own, they would simply occur to him. He could not think, "I intend to consider that problem tomorrow." He could only predict that he will think about the problem tomorrow; tomorrow he might remember the prediction, but he could not set about to bring it to pass. Any memory he has would simply occur to him as something he has thought about in the past; he could not summon up a memory. He could not, for example, think, "What did I predict yesterday? Let me see."

2. A CN could not attempt, try to seek to do anything. The idea of trying is closely tied to the idea of intending (as we saw in the case of the Minimal Agent): it is a necessary condition of trying to do A that one intend to do A. Not only could the CN not set about, e.g., to prove Goldbach's conjecture, but if he found himself doing what we would call "trying to prove Goldbach's conjecture," he could not conceive of what he was doing as trying.

3. A CN could not reason, or engage in any normative activity, except by happenstance. The idea of reasoning involves the idea of conforming to certain standards. If the CN happened to reason correctly, then so much the better. But he could not intend to conform to the canons of correct reasoning any more than he could intend to avoid mistakes.

Lacking the ability to have intentions, but not to think, the CN would not enjoy a life of the mind in anything like its ordinary sense. Thus, it seems that agency in some sense is necessary for much contemplative thinking as well as for practical thinking. A fully contemplative being would at least be able to endorse thoughts such as, "I shall try to prove or disprove Goldbach's conjecture tomorrow, and I shall try not to make any errors in the proof." Let us call such a fully contemplative being a Contemplative Agent, or CA.

Now we have a rich sense of contemplation, but a truncated sense of agency. Although the Contemplative Agent could endorse thoughts such as, "I shall do A," the expressions which can replace "A" in the schema are severely limited. They are limited, in fact, to thoughts about the Contemplative Agent's own mental life. This is not a structural deficiency, however: "I shall now prove the theorem" is no less an intention than "I shall now raise my arm." But the Contemplative Agent would still not be able to interact with the world (except insofar as events in the world caused his various mental states), and thus would be a somewhat anemic agent.

The possibility of the Contemplative Nonagent shows that self-consciousness is not sufficient for agency. What more is required? The mere ability to formulate intentions about one's own mental states, a la the Contemplative Agent, seems not enough for full agency. What is needed is the ability to endorse thoughts leading to changes in the world, changes perceptible to other self-conscious beings. Moreover, a complete account of agency would have to include an analysis of the normative thinking that we engage in when we deliberate. Thus, first-person considerations alone do not exhaust the concept of agency. Nevertheless, the first-person aspects of agency are so fundamental that to overlook them, as many philosophers have done, is to neglect the cornerstone of agency.[13]

NOTES

[1]For an account of agency in terms of causation by wants and beliefs, see Alvin Goldman, A Theory of Human Action (Princeton: Princeton University Press, 1970), pp. 80-85, especially p. 83; and Donald Davidson, "Actions, Reasons and Causes," Journal of Philosophy 60, 1963: pp. 685-700. For an account of special agency causation, see Roderick Chisholm, Person and Object (LaSalle, Ill.: Open Court Publishing Co., 1976), Chapter 2.

[2]Ludwig Wittgenstein, Philosophical Investigations (New York: 1953), section 621.

[3]Hector-Neri Castaneda, Thinking and Doing (Dordrecht, Holland: D. Reidel Publishing Co., 1975), pp. 277ff.

[4]Hector-Neri Castañeda, "Indicators and Quasi-Indicators," American Philosophical Quarterly 4 (1967): pp. 85-100; "On the Logic of Attributions of Self-knowledge to Others," Journal of Philosophy 65 (1968): pp. 439-456; "On the Phenomenologic of the I," Proceedings of the XIVth International Congress of Philosophy 3 (1968): pp. 260-266; "'He': A Study in the Logic of Self-Consciousness," Ratio 8, pp. 130-157.

[5]Immanuel Kant, The Critique of Pure Reason (B, 131).

[6]Compare Descartes' distinction in the Third Meditation between the inherent reality and representative reality of our ideas.

[7]Castañeda has described such an "Externus consciousness" in "On Knowing (or Believing) that One Knows (or Believes)," Synthese 21 (1970): p. 192. See also Castañeda's "Consciousness and Behavior: Their Basic Connections," in Intentionality, Minds and Perception, ed. H-N. Castañeda (Detroit: Wayne State University Press, 1967), sections 9 and 10. Castañeda attributes to Externus the ability to make indexical references to places, e.g., "There is a pain here". I am being more conservative on the chance that all indexical reference may ultimately presuppose the ability to be self-conscious.

[8]See Keith Gunderson, "Asymmetries and Mind-Body Perplexities," in Materialism and the Mind-Body Problem, ed. David M. Rosenthal (Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1971): pp. 112-127. Also, G.E.M. Anscombe uses an example which makes a point similar to the one I am making with computer L in "The First Person," in Mind and Language, ed. Samuel Guttenplan (Oxford: Clarendon Press, 1975): pp. 49-50.

[9]Castañeda has called this use of "he (himself)" in indirect discourse a quasi-indicator. Likewise, the second occurrence of "I" in "I intend that I go home" is a quasi-indicator. In a subsequent paper, I plan to show that all attributions of intentions have this quasi-indexical element.

[10]Castañeda, Thinking and Doing, p. 41. Castañeda in discussion has described a less-than-full agent, after which the Minimal Agent here is fashioned. Als, the forthcoming example of Smith's thought of the foul-tasting soup is taken from remarks of Castañeda's.

[11]Harry G. Frankfurt, "Freedom of the Will and the Concept of a Person," Journal of Philosophy 68 (1971): pp. 5-20.

[12]Normal agents intend to come to intend to A when, e.g., A is something that an agent intends to do next week, and the agent now believes that a necessary condition for his doing A is that he come to have an intention next week, which he would formulate then as "Now I shall do A".

[13]Much of this essay was written during Castañeda's NEH Summer Seminar, 1978. I want to thank H.-N. Castañeda, James Anderson, Charles Crittenden, Baylor Johnson, Kenneth Lucey, Jane L. McIntyre, and Roger Rigterink for their helpful comments.

DANGEROUS BEHAVIOUR

J. M. Brady
Department of Computer Science
University of Essex
Colchester CO4 3SQ, England

Introduction

The ideas presented in this paper derive from the author's experience on a project, extensively reported at the last AISB Conference (Bornat and Brady, 1976 a, b, Bornat, 1976, Brady and Wielinga, 1976) to develop a program capable of reading handprinted Fortran coding sheets (see Brady and Wielinga 1976, 1979 for fuller accounts). The Fortran coding sheets project (FCSP) was originally based on three ideas: the active deployment of knowledge to guide the perception process, the exploitation of redundancy which is seemingly ever-present in real perception, and the "heterarchy conjecture".

This paper was presented at the Third Conference of the AISB held in Hamburg, July of 1978. It has previously appeared in the Summer 1978 edition of the AISB Quarterly.