AMERICAN PHILOSOPHICAL QUARTERLY
Volume 18, Number 2, April 1981

VIII. WHY COMPUTERS CAN'T ACT

LYNNE RUDDER BAKER

THERE are numerous claims against artificial intelligence: computers have no natural interests: they are not properly embodied; they cannot handle ambiguities of various kinds, and so on. Although such claims of inadequacy have been mounted and rebutted with fervor,[1] there is an equally profound deficiency that has not been noticed at all. Without denying that artificial models of intelligence may be useful for suggesting hypotheses to psychologists and neurophysiologists, I shall argue that there is a radical limitation to applying such models to human intelligence. And this limitation is exactly the reason why computers can't act.

My argument that machines cannot act is extremely simply. It goes like this:

P1: In order to be an agent, an entity must be able to formulate intentions.
P2: In order to formulate intentions, an entity must have an irreducible first-person perspective.
P3: Machines lack an irreducible first-person perspective.
C: Therefore, machines are not agents.

Since the argument is clearly valid, we need only to determine that the premises are true. The first premise is simply a matter of definition. All actions are performed by agents, and agents may be defined as beings capable of formulating intentions. Intentions often are formulated in language but they need not be. For example, I may intend to get to a lecture on time without ever putting into words, "I'm going to get there on time." But in order to be an agent, a being must be capable of formulating such thoughts, whether he expresses them in a language or not. H.-N. Castañeda provides a convenient model: Intending is a dispositional mental state of endorsingly thinking such thoughts as "I shall do A." Such thought-contents Castañeda calls "practitions" to distinguish them as the practical counterparts of propositions. In linking a subject and an action practically, practitions have a causal thrust which propositions—e.g., propositions expressing predictions about oneself—lack.[2] The second and third premises require elaboration and support. First, I shall explain the first-person perspective and then show why computers lack it; finally, and much more briefly, I shall argue that the first-person perspective is required in order to formulate intentions and hence in order to be an agent.

I. SELF-CONSCIOUSNESS AND THE FIRST-PERSON PERSPECTIVE

The first-person perspective necessary for agency is the one that enters into self-consciousness. The emphasis here on the first-person perspective is not a commitment to what might be called "Cartesian privacy." Rather, the view to be developed is compatible with theories that self-consciousness emerges from group activity and can not be understood apart from the social contexts in which it manifests itself. My aim is to analyze what is presupposed by views—propounded by philosophers as diverse as Dewey, Sartre, George Herbert Mead, among others—which need recourse to notions such as "seeing oneself in a certain light" or "imagining ourselves as others see us." By invoking the idea of the first-person perspective, I want to bring to light a presupposition common to a number of concepts associated with self-consciousness.

One manifestation of the first-person perspective is the ability to make first-person reference in a language. In English, the device for such self-reference is the set of pronouns 'I', 'me', 'my', 'mine'. These pronouns have the unique function of indicating the thinker or speaker without characterizing him in any way. First-person indicators are not simply substitutes for names or descriptions of ourselves: When a person thinks of herself in the first-person way, she is not thereby thinking of someone-who-fits-a-certain-description, such as the person who is talking, or the tallest woman in Vermont; nor is she thereby

[1] One of the most acute critics of artificial intelligence in its more extravagent guises is Hubert L. Dreyfus in *What Computers Can't Do*, 2nd ed. (New York, 1978). Many of Dreyfus's arguments have been rebutted by Margaret A. Boden in *Artificial Intelligence and Natural Man* (New York, 1977).

[2] Hector-Neri Castañeda, *Thinking and Doing* (Dordrecht, 1975), especially Chs. 6 and 10.

thinking of someone-who-is-named-such-and-such. One need not recognize oneself under any name or description in order to tender the indicator 'I' correctly. On the other hand, one may think of someone who fits a given third-person description (e.g., the killer of Laius) and that description may truly apply to the thinker (as it did, unfortunately, to Oedipus), without entering into the first-person perspective. Thus, when Oedipus said, "Find the killer of Laius," he was not making irreducible first-person reference even though the person who fit the description was himself. In short, thinking about oneself in the first-person way does not appear reducible to thinking about oneself in any other way.[3]

Closely related to the ability to conceive of oneself in the first-person is the ability to conceive of one's thoughts 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." 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 own. In this case, X has no second-order consciousness. On the other hand, if X can think of propositions at all, then the ability to make first-person reference is sufficient for the capacity for self-consciousness. That is, if X can conceive of himself from the first-person perspective, then he can be conscious that his thoughts are his own. Therefore, an entity which can think of propositions at all enjoys self-consciousness if and only if he can make irreducible first-person reference.[4]

It may be objected at the outset that machines cannot have thoughts at all. Since it is logically possible that an entity is conscious without being self-conscious,[5] I want to put aside the objection that machines cannot think of things at all. One reason to put it aside is that it is difficult to see how to adjudicate the point head-on; a second reason to put it aside is that I shall conclude that there is an important limitation to the analogy between minds and machines, and I want to concede to machines the benefit of the doubt here. Thus, for the purpose of this discussion, but not in general, when I say, "X can think of propositions," I mean it in the weakest possible sense—as weak as "X can have internal states that have propositional content"—so that we can say of computers that in this sense they can have thoughts.

Of course, thoughts may have propositional content without the first-person perspective. Examples of thoughts which have propositional content but which lack first person reference include: "The cow jumps over the moon," "2 + 2 = 4," "Smith believes that 2 + 2 = 4," and "It was obvious that anyone who knew the combination could have opened the lock."[6] When I say "A thing or person can conceive of its thoughts as its own," I mean it in a correspondingly weak sense—as "that thing or person can have internal states with irreducible first-person propositional content." Thus, an entity can conceive of thoughts as his own if he can have internal states expressible in the irreducible first-person as, for example, "I am wounded" or "I am now thinking that the sky is blue." Occurrences of such thoughts are episodes of self-consciousness. Self-consciousness as a dispositional state at least involves both the capacity to make irreducible first-person reference and the capacity to have thoughts which have propositional content

A caution is needed here. Although I have talked about the propositional content of thoughts involved in self-consciousness, what is important for intending, at least on Castañeda's view, is not the propositional content of thoughts, but rather their practitional content: Again, practitions have an operator on the copula so that subject and predicate are not joined as they are in simple predication, but rather in a practical way, suitable for action.[7] Since the first-person perspective which enters into the self-consciousness of propositional thinking is the same as that which enters into practical thinking, the above discussion of the first-person perspective in terms of propositions will suffice for purposes of showing how machines lack the first-person perspective in general.

[3] Castañeda, "Indicators and Quasi-Indicators," *American Philosophical Quarterly*, vol. 4 (1967); pp. 85–100; "On the Logic of Attributions of Self-Knowledge to Others," *The Journal of Philosophy*, vol. 65 (1968), pp. 439–456; "On the Phenomenologic of the I," *Proceedings of the XIVth International Congress of Philosophy*, vol. 3 (1968), pp. 260–266; "'He: A Study in the Logic of Self-Consciousness," *Ratio*, vol. 8 (1966), pp. 130–157.

[4] See my "First-Person Aspects of Agency," *SISTM Quarterly*, vol. 2 (1978), pp. 10–16.

[5] See Castañeda's discussion of Externus consciousness in "On Knowing (or Believing) that One Knows (or Believes)," *Synthese*, vol. 21 (1970), pp. 187–203. See also Castañeda, "Consciousness and Behavior: Their Basic Connections," *Intentionality, Minds and Perception*, ed. by H-N. Castañeda (Detroit, 1967), pp. 121–158, especially sections 9 and 10.

[6] For other examples, see Terry Winograd, *Understanding Natural Language* (New York, 1972), pp. 52–53.

[7] *Thinking and Doing*, p. 280.

II. Machines: No First-Person Perspective

Several writers[8] have seen an analogy between certain computers and self-conscious beings. I shall argue to the contrary that the analogy falters because machines lack the first-person perspective which is integral to self-consciousness. My evidence is largely linguistic: Computers cannot make the same kind of reference to themselves that self-conscious beings make, and this difference points to a fundamental difference between humans and computers—namely, that humans, but not computers, have an irreducible first-person perspective.

Earlier I claimed that thinking about oneself in the first-person way is not reducible to thinking about oneself in any third-person way. As further support for that claim, consider an attribution of self-belief: "J. Johnson believes that he (himself) is wealthy." The "he (himself)" or Castañeda's "he*" in indirect discourse is called by Castañeda a *quasi-indicator*: it attributes irreducible first-person reference to the person referred to by the antecedent of "he*," where the antecedent lies outside the scope of the cognitive or linguistic verb. Thus, "J. Johnson believes that he* is wealthy" attributes to J. Johnson the first-person belief which he would express as "I am wealthy." Now contrast "J. Johnson believes that J. Johnson is wealthy." J. Johnson would express the latter belief in the third-person as, "J. Johnson is wealthy."

Perhaps surprisingly, "J. Johnson believes that he* is wealthy" is not equivalent to "J. Johnson believes that J. Johnson is wealthy." To see this, consider the following little fantasy: J. Johnson, a New York multi-millionaire, is abducted, bopped on the head, and left on the side of the road in Vermont. When he recovers, he cannot remember his immediately prior life. Eeking out a living on a sheep farm in Vermont, he regularly reads of J. Johnson, the missing millionaire, in the newspaper. J. Johnson thus comes to believe that J. Johnson is wealthy, but, not knowing that he* is J. Johnson, he does not believe that he* is wealthy. Then, our sheep-farming J. Johnson wins the Vermont lottery; at about the same time, he reads in the newspaper that, due to mismanagement, the financial empire of J. Johnson has crumbled and that J. Johnson is now a pauper. Thus, still not believing that he* is J. Johnson, J. Johnson believes that he* is wealthy (since he won the lottery); but he does not believe that J. Johnson, whose empire has been lost, is wealthy. Therefore, "J. Johnson believes that J. Johnson is wealthy" is not equivalent to "J. Johnson believes that he* is wealthy": either can be true while the other is false. And J. Johnson's genuine first-person belief, "I am wealthy," is thus not equivalent to J. Johnson's third-person belief, "J. Johnson is wealthy." Analogous examples, showing the non-equivalence of a first-person formulation and *any* third-person formulation, can be construed for any self-conscious state. Since there is an ineliminable difference between attitudes about oneself from the first-person perspective and attitudes about someone-who-is-in-fact-oneself, there is an irreducible first-person perspective that cannot be analyzed in terms of the third-person. This irreducible first-person perspective is enjoyed by self-conscious beings. As a manifestation of this irreducible first-person perspective, the indicator 'I' is not simply a replacement for a third-person name.

Computers do not share this irreducible first-person perspective. Of course, computers may be programmed to use 'I' in grammatical sentences; in that case, 'I' is self-referential in the sense that 'I' always refers to its apparent user. (Compare the poison labeled "Don't drink me.)" "I am in state S" printed out by computer C refers to C—and hence to itself—just as "C is in state S" printed out by computer C refers to C—and hence to itself. But, and I will argue for this, C's use of 'I' is not the use of 'I' characteristic of self-conscious people. A machine's production of 'I' no more indicates the first-person perspective associated with self-conscious beings than its production of the word "pain" is evidence that it has feelings.

What, then, *would* induce us to say that a machine has a first-person perspective? A critic might request a specification of empirically ascertainable conditions (perhaps in the form of a so-called Turing Test) under which we would be required to attribute the first-person perspective, followed by an argument to show that the conditions cannot be fulfilled by machines. But such a request would be unreasonable: the problem of the first-person perspective, like the traditional problem of other minds, precludes enumeration of such sufficient conditions. (At best, a Turing Test only reveals that we can make mistakes—an unstartling result.)

Why are we unable to specify empirical criteria sufficient for the first-person perspective? It is assuredly *not* the case that the only way to ascertain

[8] See, for example, D. M. Armstrong, *A Materialist Theory of the Mind* (New York, 1968) and Keith Gunderson, "Asymmetries and Mind-Body Perplexities," in *Materialism and the Mind-Body Problem*, ed. D. Rosenthal (Englewood Cliffs, N. J., 1971), pp. 112–127, especially pp. 121–123. Some of the most sensitive papers along these lines are in Daniel Dennett's *Brainstorms* (Montgomery, Vermont, 1978).

whether something is self-conscious is, *per impossibile*, to slip into its "mind" to check that it has the proper kinds of experiences (or even that it has experiences at all). Rather, the first-person perspective is displayed in our patterns of action, language and thought, and in the myriad conventions that regulate our common life. Suppose that it were possible to enumerate conventions (and what is to count as conforming to each one) such that conformity to some subset of them would suffice for ascribing the first-person perspective. Such a list would still be useless to determine whether computers have a first-person perspective, because the very language of convention is already laced with the idioms of self-consciousness.[9]

There may appear to be a circularity here: our attributions of self-consciousness to others are rooted in our common participation in the conventions that define our life: but the very language used to describe and conform to those conventions already presupposes that the participants are self-conscious. This apparent circularity is not a fault of my argument; it is only the commonplace that intentional terms can not be defined without using the language of intentionality. This is not an argument that we could never conceivably be justified in attributing to machines the first-person perspective; the point here is only that we cannot specify empirically ascertainable sufficient conditions for the first-person perspective any more than we can for other minds generally.

To put the point another way: suitable empirical conditions for the first-person perspective would require an entity to be in some observable state to justify attribution of the first-person perspective; to be noncircular, such conditions must be statable in the third-person without invoking the first-person via quasi-indicators. But since the case of J. Johnson shows that we cannot specify criteria for the first-person perspective in the third-person (without quasi-indicators), we can never know whether any proposed set of empirical conditions is sufficient for the first-person perspective. It may well be possible to state in the third-person (without quasi-indicators) some empirically necessary conditions for the first-person perspective—perhaps in terms of complexity of physical structure or of complexity of behavior—but these would be no help here, where we want to know under what conditions we ought to attribute a first-person perspective. For these reasons, it appears illegitimate to request emuneration of suitable conditions of the first-person perspective coupled with an argument to show that machines can not satisfy them. So my argument must take another tack.

From what has already been said, it is apparent that a computer can not simply be *programmed* to have a first-person perspective. Since a program is a set of instructions to be carried out sequentially, computers can be programmed to perform tasks governed by formalizable rules, whether such rules are algorithms or heuristic programs. Now consider a dual presupposition of the claim that the first-person perspective is programmable: first, the ability to have first-person episodes would have to be a result of information-processing; and second, input, in the form of discrete items, would have to be transformable by means of specifiable rules into first-person episodes. This double presupposition can be seen to be unwarranted on several grounds.

First, consider the difficulty of finding the appropriate data for input, on which the rules would have to operate. Dreyfus has argued convincingly that the world we dwell in can not be represented as some number of independent facts ordered by formalizable rules.[10] On the one hand, facts are not detachable from the situations that give them their significance and relevance; on the other hand, the situations cannot be made wholly explicit in terms of rules. Thus, the role of context in knowledge and perception inhibits the isolation of relevant data to be used as input for the first-person perspective.

Second, and perhaps more important, is the fact that the first-person perspective is not the result of any rule-governed process. A necessary (but not sufficient) condition for programming a first-person perspective would be the discovery of heuristic rules according to which the first-person perspective is achieved. But there are no heuristics for attainment of the first-person perspective. Even for admittedly self-conscious beings, no amount of third-person information about oneself ever compels a shift to first-person knowledge: Oedipus's being aware that Laius was killed at a crossroads, even coupled with his first-person knowledge that he* had killed a man and his party at a crossroads, did not lead Oedipus to the conclusion that he* was the killer of Laius. Because there is always a gap between third-person knowledge about oneself and the corresponding first-person

[9] For a provocative analysis, see David Lewis's *Convention* (Cambridge, Mass., 1969). Even if Tyler Burge ("On Knowledge and Convention," *Philosophical Review*, vol. 84 (1975), pp. 249 255) is correct in his criticisms of Lewis for overstating the role of self-consciousness, it may still reasonably be claimed that self-consciousness is required to be the sort of being that can have its life governed by conventions.

[10] Dreyfus, *What Computers Can't Do*. Ch. 6.

knowledge, there is no way to specify how much information about oneself in the third-person is sufficient to lead one to first-person belief; the variety of conditions under which we make the leap to first-person belief is so extraordinary that it is futile to look even for rules to govern all the appropriate transformations of third-person sentences to first-person sentences. Thus, assuming that one already has some first-person beliefs, there are no rules for increasing one's store of first-person beliefs on the basis of knowledge about oneself in the third-person: J. Johnson's realization that he* is a millionaire—as opposed to his discovery that J. Johnson is a millionaire—is not the outcome of a heuristic process.

But if there are no heuristics for the having of first-person episodes by unquestionalby self-conscious beings, much less can there be heuristics for initially reaching the capacity to have first-person episodes. *Any* inference to a first-person belief seems to require prior first-person beliefs as premises. Oedipus could never have discovered that he* was the killer of his* father and the husband of his* mother if he had had no genuine first-person beliefs at all.[11] Whatever the genesis of the first-person perspective, it is not a rule-governed inference from non-first-person premises to a first-person conclusion. The ability to see oneself from the first-person perspective is not the sort of thing that can be arrived at by following instructions; for a first-person stance is no more the outcome of any procedure than is the ability to feel anxiety. In a word, the ability to have first-person episodes turns out to be what Gunderson would call a program-resistant feature of mentality.[12]

Granting that the first-person perspective cannot arise from programming, some (including Gunderson) would still counsel agnosticism: to say that a machine can not be programmed to have the first-person perspective is only to say that computers at this time lack a certain capacity; who is to say what future developments of *hardware* may bring? But in order for improvements in hardware to warrant attribution of the first-person perspective to computers, computers would have to be capable of being in situations, like that of J. Johnson, whose descriptions require first-person language. Beings whose states can be completely described without recourse to first-person language (via quasi-indexical reference) have no first-person perspective.[13] I believe that third-person language will always suffice to describe a computer's states because of the following difference between persons and computers:

There is a variety of referential error to which beings acting self-consciously are logically immune. This immunity—which has nothing to do with traditional arguments about the alleged infallibility of reports of one's own mental states—is the characteristic feature of the first-person perspective. By contrast, it would seem that machines are always both logically and physically liable to error. If so, machines must lack the first-person perspective.

The kind of referential error at stake can be explained as follows. Every time anything—a person or a computer—uses 'I' in a grammatical English sentence, there is something to which 'I' refers. But this is not all. When 'I' is used in its usual way by self-conscious beings, it refers to the thing to which the user takes it to refer. From the first-person perspective, Smith could never use 'I' and take herself to be referring to someone other than herself. What Smith takes to be herself is herself. Say that Smith complains, "I have heartburn." Now she may mistake her internal state as heartburn when it is really a mild heart attack, but when she refers to herself in the irreducible first-person way, she cannot misidentify *whose* internal state it is. It is indisputably her own. There could never arise an occasion for someone to say, "She is mistaken; that isn't *her* heartburn (or heart attack, or whatever); it is Ralph's." Again: if I say "I am six feet tall," I may be mistaken in attributing that height to myself, but when speaking from the first-person perspective, I cannot take myself to be referring to someone different from myself. So first-person pronouns, in their typical use by self-conscious beings, are immune to the kind of referential error to which names are susceptible. This explains why the device for irreducible first-person reference does not function simply as a name for the user. By contrast, for computers which issue first-person sentences, 'I' does function as a name. Writers such as Winograd say as much.[14] Moreover, we can construct

[11] For other episodes whose descriptions require indexicals (first-person and otherwise), see John Perry's "The Problem of the Essential Indexical" in *Nous*, vol. 13 (1979), pp. 3–22. I suspect that problems involving such episodes are generally unsolvable by computers.

[12] See Keigh Gunderson's *Mentality and Machines* (New York, 1971), Chs. 3 and 5. If my argument is correct, most of the above paragraphs would have to be taken metaphorically, since none of the language of action (e.g., "solving a problem", "following heuristics") would apply literally to machines.

[13] My point here is a kind of converse of Dennett's in "Intentional Systems" (reprinted in *Brainstorms*, pp. 3–22). Dennett holds that any system *may* be described as having beliefs and desires if it is convenient to do so. I hold that some things (such as persons) *must* be described not only as having beliefs and desires but also as having irreducible first-person beliefs and desires if they are to be understood.

[14] For example, Winograd's robot SHRDLU uses 'I' as simply another name referring to SHRDLU. See *Understanding Natural Language*, p. 143 and p. 158.

cases in which machines issuing first-person English sentences can systematically misapply 'I' in the way that our self-conscious Smith above could not. E.g., say that a computer Q answers questions about itself, Q, and another computer R. When these questions require a third-person answer, Q correctly distinguishes between Q and R. But when the questions require a first-person answer, Q systematically answers as if it were R. Thus, in issuing first-person English sentences, computers are liable to a kind of error in reference that would be impossible from the first-person perspective.

In the absence of a capacity to conceive of itself from the irreducible first-person perspective, a computer cannot be said to have genuine self-belief. Moreover, vanity, self-deception, self-esteem, self-loathing, and all other attitudes which depend upon a regard of oneself in the irreducible first-person are forever foreign to the computer—no matter how "intelligent" it is.

Thus, a crucial difference between machines and self-conscious beings is this: for self-conscious beings, there is an irreducible distinction between genuine self-consciousness and consciousness of someone-who-is-in-fact-oneself; for machines, on the other hand, there is no corresponding distinction between say, genuine self-scanning and scanning a unit-which-is-in-fact-itself—just as in the case of self-defrosting refrigerators, there is no distinction between genuine self-defrosting and defrosting a refrigerator-which-is-in-fact-itself.

We can summarize this difference between computers and self-conscious beings *vis-à-vis* the irreducible first-person perspective in terms suggested earlier:

> Attributions of self-conscious states require the irreducible quasi-indicator, e.g., 'X is conscious that he* is F', where 'X is conscious that he* is F' is not equivalent to any proposition—e.g., 'X is conscious that X is F'—which lacks a quasi-indicator.

No attribution of any state to a computer requires a quasi-indicator. For Q, "Q believes that it is F" is equivalent to "Q believes that Q is F", or to some other proposition of the form 'Q believes that a is F,' where 'a' is a name, description or indicator with no occurrence of 'he*'. So computers lack the irreducible first-person perspective, and the analogy between minds and machines founders on the facts of the first-person perspective.

This conclusion fits comfortably with other intuitively plausible positions. First, it is not claimed that the human species is unique in enjoying a first-person perspective. Indeed, certain experiments on chimpanzees suggest that they may be trained to recognize themselves in the first-person way;[15] it is plausible to hold that such chimpanzees have a sort of rudimentary self-consciousness. On the other hand, nonhuman higher animals are not agents in anything like the way that we are. In general, to the higher animals, I would apply Malcolm's distinction between thinking and having thoughts in the sense of entertaining propositions.[16] Dogs, as well as chimpanzees, do things intentionally in the sense that, according to Malcolm, dogs can think. But neither dogs nor the trained chimpanzees can entertain propositions at all and hence can not formulate the thoughts required for full-fledged agency.

Second, the first-person perspective is not claimed to be either logically or temporally prior to the third-person perspective.[17] One can not have a first-person point of view without a concept of otherness by means of which to distinguish things as different from oneself; conversely, one cannot have a concept of things as different from oneself without the ability to think of oneself from the first-person point of view. Thus, lacking a first-person perspective, a computer has *no* genuine perspective.

III. The First-Person Perspective as Necessary for Intentions

The language of action is filled with presuppositions about the first-person perspective: part of what makes something the kind of action it is (weeding a garden, playing a prank, apologizing and so on) is what the agent believes that he* is doing. All that remains to be shown is that the ability to formulate intentions and hence to be an agent requires the first-person perspective.

To be capable of formulating intentions is to be capable of endorsing genuine first-person practitions—thoughts of the form 'I shall do A' or 'I'm going to do A'. Since the genuine first-person point of view is irreducible, it follows that beings which lack

[15] Gordon Gallup, Jr., "Self-Recognition in Primates: A Comparative Approach to the Bidirectional Properties of Consciousness," *American Psychologist*, vol. 32 (1977), pp. 329–338.

[16] "Thoughtless Brutes," *Proceedings and Addresses of the American Philosophical Association*, vol. 46, 1973, pp. 5–20.

[17] In her excellent "Second Person, Past" (forthcoming in *Philosophia*), Annette Baier rightly points out that the personal pronouns have sense only in relation to each other. She puts somewhat heavier emphasis on the role of the second-person than I do.

the first-person perspective are not capable of intending and hence are not agents. Although correct, this conclusion is a little hasty. Let us consider briefly the elements of intending.

If Jones rehearses an intention to go home, his thought must include an idea of himself connected to his idea of the action by means of the practical operator on the copula; for the causality of intending resides in the practical way in which the agent links his idea of himself to his idea of the action. But notice: Jones must conceive of himself in the first-person way. His intention to go home does not simply link practically someone-who-is-in-fact-himself with the property of going home; rather the connection between the agent and the action involves Jones conceived in the first-person way.

But, it might be objected, some attributions of intention do no seem to ascribe a first-person perspective. Consider a case such as "Jake intends for Dan to go home," which does not appear to attribute to Jake a first-person reference. Without stopping to analyze this, let me suggest that it is partly a prescription and partly an intention. To the extent that it is an intention, Jake intends that he* do something which will lead to Dan's going home. Perhaps one of the things which Jake intends to do is to utter the command, "Dan, go home." Of course, to say that Jake intends to do something which will lead to Dan's going home is to attribute to Jake the first-person perspective which appeared to be missing from "Jake intends for Dan to go home"; for Jake would formulate his intention to urge Dan to go home in the irreducible first-person.

Not only is the first-person perspective required for the formulation of intentions, but also each instance of intending requires that an agent have first-person beliefs. Consider, e.g., "Woodrow Wilson intended to make the world safe for democracy." Among other things, Wilson must have had beliefs about the circumstances that he* was in e.g., he must have believed that he* had relevant abilities, that he*was in a position to influence other nations, etc. Moreover, Wilson could not have believed that his future was closed. That is, he could not have believed that the world's being safe for democracy (or not being safe for it) was a foregone conclusion independent of his intention; nor could he have believed that his future precluded alternative courses of action. This is not to say that determinism is false, but rather that an agent, from his own first-person point of view, must not conceive of that part of his future about which he has intentions as already fixed regardless of what he intends.[18] The relevant beliefs about the future must be from the first-person perspective: if Wilson has an intention, he cannot believe that *his own* future is closed, regardless of any beliefs he has about himself conceived in the third-person.

Since the first-person perspective enters not only in the formulation and attribution of intention, but also in the beliefs presupposed by any given intention, no entity which lacks the first-person perspective can be an agent.

IV. Conclusion

So machines cannot engage in intentional behavior of any kind. For example, they cannot tell lies, since lying involves the intent to deceive; they cannot try to avoid mistakes, since trying to avoid mistakes entails intending to conform to some normative rule. They cannot be malevolent, since having no intentions at all, they can hardly have wicked intentions. And, most significantly, computers cannot use language to make assertions, ask questions, or make promises, etc., since speech acts are but a species of intentional action. Thus, we may conclude that a computer can never have a will of its own.[19]

Middlebury College, Vermont

Received July 31, 1979

[18] Some would state this condition more strongly and require Wilson to believe that it is within his* power to make the world safe for democracy. I prefer the negative formulation for this reason: the stronger formulation implies that Wilson has the concept of causal efficacy; but I think that we want to attribute intentions to some (e.g., young children) who may still lack the concept of causality.

[19] Versions of this paper were read at the Western Division of the American Philosophical Association (1979), the University of Rochester, Union College and the Creighton Club. I wish to thank my commentators, Martin Ringle and Rew Godow. Also, I am indebted to Hector-Neri Castañeda, Richard Taylor, Annette Baier, Philip Kitcher, Victor Nuovo and Stanley Bates for helpful comments along the way.