DANIEL DENNETT

# COMMENT ON WILFRID SELLARS

I might say, by way of introduction, that unlike the earlier commentators, both Putnam and I did have a chance to see a version of this paper before the session. In fact about two weeks ago I got the wind up about the conference and called Wilfrid to see if a paper was forthcoming. He immediately, very generously, sent me not one but six papers. The paper that he read today was not among them. But he did tell me that the paper he would read would consist of various parts of the six that he did send me and would come mainly from one paper central to which was an analysis of transparent belief in terms of opaque belief. I thought this analysis was entirely wrong, so I worked very hard to produce an elegant and concise refutation of it. When the real paper arrived a week later, I found that wasn't the part I was supposed to take. At that point I was so mentally exhausted that I put away the paper until last night about midnight, when I sat down to write these comments. I think I may have been in a hallucinatory frame of mind by the time I got through. So if a look of horror and disbelief comes over my face as I read these remarks, you should not be surprised. Of course, maybe I'm still in a hallucinatory frame of mind. I rather hope I am.

Yesterday we saw David Lewis' ingenious suggestions about how to put Gil Harman's three levels of meaning together into a single theory. Sellars explicitly is setting out to do part of that job; namely to provide a level-one theory of meaning. Which is roughly, I take it, in terms of Lewis' diagram, a matter of filling in the relations between **Ao**, agent's beliefs and desires as expressed in our language, and **P**, agent's physical system. But Sellars is going to do this via a functional analysis of the agent's thoughts, for these are viewed as being in language, in some sense. In setting out in this way Sellars is, I think, explicitly confirming an interpretation of his program offered by Harman in 'Three Levels of Meaning'. There Harman said, speaking of Sellars' early paper 'Some Reflections on Language Games': "Sellars is simply offering a functional account of psychological states in the guise of a theory of meaning." The program, I

*Synthese* **27** (1974) 439–444. 

take it, is to isolate a certain variety of agent's speakings and dispositions to speak, namely his candid or spontaneous thinkings out loud and proximate propensities to do so, and to identify these *as* the agent's thoughts, even though this identification is acknowledged to be an oversimplification and distortion in some regards. I want to say why this seems reasonable to Sellars and, to some extent, to myself. We want, it seems, a psychological theory of an agent's thoughts, and thoughts are individuated by their contents or meanings. Since Sellars' proposed theory of meaning is that meaning is functional classification, psychological states and events are to be individuated by their function. But how are we to examine psychological states and events to determine their functions? Lacking, at least temporarily, a detailed neurophysiological identity theory, we cannot very well examine psychological states and events the way we can examine hormonal or metabolic states and events, i.e. by seeing what happens mechanically under various conditions and then doping out a functional account of internal design. That, in this instance, would be putting the cart before the horse. Rather we must start with the *meanings* of thoughts as psychological states and events as the given, and work out functional accounts of these (perhaps later to find realizations of these functions, of whose existence we are assured by our theory of meaning, in the hardware of the brain). Thus, the program is roughly analogous to that of pre-Watson-Crick genetics. So how do we get at the meanings of the thoughts in order to do our functional analysis? We cannot look at the thoughts, and our access to our own thoughts, let alone the thoughts of others, is complicated by their fleeting and protean passage through our mind. If one is a Cartesian of the sort Sellars alludes to (I'm not quite sure what this sort is), or if at any rate one maintains the view that thoughts, at least some of them, are imageless, wordless, speechless meaning-bearers, then function would appear to be utterly inscrutable. For such thoughts would appear to be both atomic to analysis and indefinitely bountiful and various; lacking parts, their functions cannot be determined to be a function of the function of their parts. And there are too many of them to catalog their functions individually. If, on the other hand, it is true that language is the medium in which we think, the psychological theorist will be most fortunate. For language does have parts, and these seem just the right sorts of parts to be subject to analysis in terms of functional individuation. But the claim that language *is* the medium of our thought is

multiply ambiguous and, I think, vulnerable to doubts and misgivings on all its readings. Sellars acknowledges this to some extent, and has some suggestions for qualifying the thesis to make it more attractive; but not yet attractive enough for me. However I am not at all sure that he needs the thesis that language is the medium of our thoughts in any form, either as an implication of some larger thesis in his proposed theory of meaning or as a permitting condition for his program.

I may have badly misunderstood Sellars here, but in any case I think he commits himself to the view that thoughts that are somehow verbal are, as psychological events and states, actual mediators in human behavioral reaction to experience. That is, I take him to hold that the actual processes that take a person from experience to action are perspicuously divisible into parts analogous to the functional parts of the verbal chain of reasoning we can or do use to predict and explain that passage from experience to action. Perhaps, though, I put it too strongly. Perhaps Sellars is only claiming that the thoughts he is talking about, the thinkings out loud and so forth, are in fact actual episodes and are properly verbal in an extended sense, and that while they are not themselves always mediators in our action, they provide a model for the other real episodes that are. The latter claim in any case permits a certain division of the issue which allows me to say what I'm inclined to agree with in Sellars' view. The candid thinkings out loud, or inhibited would-be thinkings out loud, are, I think, a real and important category of episode. And Sellars' description of them is acute and accurate. Such thoughts thought out loud are the protocol statements used as data in many psychological investigations of problem-solving and the like. Such thoughts are not actions, but there are reasons, as Sellars points out, for calling them acts. They are not, or need not in volve, the occurrence of a sequence in time, a phonological analog; nor are they necessarily bound to a particular natural language, even if the agent is not multi-lingual. I would say that these thoughts realize our semantic intentions of the moment, in this sense: they are that against which we test our utterances' execution. Slips of the tongue, misuse of words, even distorted expressions of what we intended to say at the moment, are judged by us against the standard set by these episodes. That is, these episodes are, you might say, feedback standards.

Now perhaps Sellars would not agree with all of this characterization I've given so far to his episodes. But I suspect our disagreement would be

minor and readily resolvable. Now to the disagreement. I think that these episodes, while actual, and while they play an essential role in the mediation of our actual linguistic behavior, do not mediate behavior generally. They are relatively peripheral, and their absence or presence on a particular occasion is seldom critical to the accuracy of intentional explanation of the non-linguistic behavior of the agent at the time. Some of Sellars' remarks tend to suggest that he disagrees. In discussing language departure transitions, he says: "The speaker responds to such linguistic conceptual episodes as 'I will now raise my hand' with an upward motion of the hand, etc.," and later he says: "according to the verbal behavior it supports, a volition is a thinking out loud or a proximate propensity to think out loud, 'I shall' ...." But perhaps, as I suggested before, Sellars holds a more circumspect view. Thinkings out loud or proximate propensities to think out loud are only a model for the deeper, less scrutable and accessible episodes, and it is just here that verbal behaviorism is to be viewed as a permissible oversimplification. But then why should we suppose that the rules that govern functioning that form the substance of the proposed theory of meaning govern our model episodes at all? That is, the theory of meaning has rules in it which govern function and thereby determine meaning, but why should we suppose that these rules govern the thinkings out loud that actually occur? We think enthymematically, in the sense of thinking that we're talking about. The transitions exhibited in our actual thinking are almost never warranted directly by any plausible rules. These loose-jointed transitions of our protocols are hints, but as hints of deeper episodes they are unreliable. Consider an example. Mr. Tortoise, asked how much $12 \times 37$ is, thinks out loud as follows: "$2 \times 7$ is 14, carry the 1; $2 \times 6$ is 3 and 1 is 7; $1 \times 37$ is 37; add them up, thats 444." There's a protocol par excellence. That's Mr. Tortoise. Now Mr. Hare, not to be confused with Professor Hare, is asked the same question and says straight out, "$12 \times 37$ is 444". Now this act of his may puzzle us. But certainly it can happen. He may have an explanation. He may say, "Well I once worked at S. S. Pierce where they sold Mouton Rothschild at $ 37.00 a bottle and most people bought it by the case." (Devaluation has forced S. S. Pierce to discontinue case discounts.) Or he may have no explanation. He may simply say, "I just can do that." "You asked me and I told you; $12 \times 37$ is 444". Perhaps he's a mathematical prodigy. Now the fact that he says "I just can do it" and can give no

other explanation shouldn't puzzle us, because after all if we ask Mr. Tortoise how he knew the first *step* of his protocol, "2 × 7 is 14", he has no answer to that. "I just can do that." Or if he does offer an answer, he's just guessing or providing some sort of a theory that he has no special access to himself.

Now it seems more probable, I suppose, that in Mr. Tortoise "2 × 7 is 14" is somehow a fact stored raw. This is more probable than that Mr. Tortoise unconsciously added 7 pairs and then gave the answer, but the latter is presumably possible. Now when it comes to hypotheses about how the prodigy, Mr. Hare, answered the question, we have initially no clues at all. That is, neither we nor he has any clues about how he does it. It is conceivable, for all we know, that he unconsciously converted the two numbers into binary notation and did a series of additions computer-style. Very possibly, however, he solved our mathematical problem in just the way Mr. Tortoise did, only unconsciously. Now, if so, then there is reason to suppose that some things function in Mr. Hare in such a way that Mr. Tortoise's protocol could tell the story of their operation as well as the story of operations in Mr. Tortoise. They could tell the story of the covert operations in Mr. Hare as well as they could tell the story of the more accessible operations of Mr. Tortoise. But from the fact that Tortoise's protocol could be used to make a prediction of the answer Hare gave (and it will do this however Hare has arrived at his answer), we cannot assume it is a model of any function in him. If there are functional parts Tortoise-style in Hare, then they can be assigned the meanings of the words in Tortoise's protocol. If there are different functional parts which mesh with a different rational reconstruction taking us from 12 × 37 to 444 then these parts can be endowed with those meanings. And if we can't find a reconstruction that matches well with whatever engineering we might find in Hare's brain, if the sort of functioning that does go on in his brain is not amenable to this sort of analysis, what should we conclude? That language does not express our thought? Hardly. That since no actual physical things in Hare have the function of the words in our rational reconstruction, there must be some other actual non-physical things that do? No. A failure to find the presumed functional isomorphism would not diminish the utility of intentional explanation of behavior via a Sellarsian, or other, theory of meaning. Nor would it impugn the correctness of such an account in particular cases. That is, it seems to me the best way to put

this is that Sellars distinguishes between the coarse-grained theory of verbal behaviorism, and the fine-grained theory, but there is no reason to view Sellars' coarse-grained theory as an empirical theory, in any sense, of any actual episodes. Now in talking with Hilary about this, I said that it seemed to me that one probably ought to be able to construct a compelling argument to the effect that the brain couldn't function in such a way that intentional explanations work to the extent that they do unless there were in fact perspicuous functional divisions, with the functions of larger divisions being functions of functions of smaller or subsidiary brain parts, such that we could consider these functional parts as representations in some sense. The argument that Hilary suggests is that it's just clear that the brain has to have some sort of what he calls 'generaticity' which can only be gotten in this way. I'm not quite sure how this argument would run, but in any case, if there is such an argument with such a conclusion, and it's a sound one, it still would not follow (and this is I guess my main disagreement with Sellars), from the adequacy of a theory of meaning as functional classification, that *that* theory was also a psychological theory of the entities that used the language for which Sellars had the theory.