This work originally appeared in:
Dennett, Daniel. 1986. Commentary on Newell: Is There an Autonomous "Knowledge Level"? In *Meaning and Cognitive Structure: Issues in the Computational Theory of Mind*, ed. Z. Pylyshyn and W. Demopoulos, 51-54. Norwood, N.J.: Elsevier.

Commentary on Newell's Paper

Is There an Autonomous "Knowledge Level"?

by Daniel Dennett

I'll try to be brief since lunch awaits us. I've rewritten my comments about five times this morning. I agree in very great measure with what Professor Newell has to say in his paper on the knowledge level. Indeed my idea of the intentional stance as a very close cousin of his idea, as he has said. Our agreement is very widespread on that point. I'm afraid our agreement even consists in our both being fuzzy in the same places about some of the really important issues. Perhaps the most useful thing I can do here is to point out some of those places and try to shore up Professor Newell's fuzziness in a few places with a little bit of slowly emerging clarity of my own. He may disagree with my attempts to shore him up, of course.

Is there a knowledge level? Do we really need to talk this way? I think the answer is "yes," and I'll try to say why; I'll also try to underline and support a few of the things Newell said about that. Here's one of the questions he's a bit fuzzy on. Is the knowledge level in fact a level of system description? On one of the slides he doesn't show it, on the next slide he does show it, as a level of systems description. But then when he says it's a level of system description, he says it's a competence model; I think that's just right. It describes the system only by saying what the system ought to be without saying how you're going to achieve that. It is a task-setting description, as it were, setting a task for the designer rather than telling the designer any feature that the design is supposed to have other than this particular competence, this particular bit of knowledge. Knowledge is an extremely abstract commodity, but we do have it. Each of us has different amounts of it on different topics. But in virtue of what do we have it? Many people when faced with a question of this sort get scared. They are afraid of becoming philosophers or they are afraid of dualism. They're afraid of abstraction. So they try to answer the question of "In virtue of what do we have knowledge?" by reducing knowledge to something a little more concrete. I won't review the history of various forms of crude materialism that were replaced with forms of less crude materialism which were replaced with forms of still less crude materialism. As we march up the ladder through Turing machine functionalism we finally get to a view such as the view in Fodor's Language of Thought (1975) which I think in this regard is very similar to Newell's view that what we're going to reduce knowledge to is the physical symbol system level. This question of

what's the relationship between the knowledge level and the physical system level is really another way of asking of Fodor whether he really means to reduce psychological features such as believing that p, for some proposition p, to being in a computational relation to a particular formula, which then, via the semantics of the symbol in the formula, can be seen to mean that p.

My view is that the attempt to reduce the knowledge level to the physical symbol system, or Fodor's attempt to reduce it to a relation between a system and a syntactic expression, is a mistake; it is the last vestige of the reductionism that we want to get rid of. We want to agree that the knowledge level is an irreducible level of characterization, and we should learn to get over our fear of that. We should be willing to accept that we can have an abstract characterization of a physical system in terms of the knowledge that it has in it without then reducing that, via the cascade of system levels, down to the hardware. What we're left with is the idea that every piece of particular knowledge, every particular intentional system or physical symbol system with knowledge, is after all just some physically realized system, and it will not be any accident that it has the knowledge it has. But we won't suppose that we can achieve anything like a type-type reduction. Newell spoke very much in passing about one of the important reasons for this. You need the knowledge level and you need it unreduced because I know things that you know. That's something we can have in common. Now at the physical symbol level we may be very different. If that's not obvious in the case of two people just consider two computer systems, one of which uses production systems and one of which uses some other LISP-based virtual architecture. At the knowledge level they may have something in common. Maybe they both "know" that higher is transitive, or maybe they both "know" that you shouldn't sacrifice your queen for a couple of pawns under most circumstances. These are features which they have in common and which are only describable at the knowledge level because their architecture, and their symbol systems, are different. They have different languages of thought. This, I think, has always created a problem for somebody like Jerry Fodor who has the problem of how to describe the transaction that occurs when one person speaks a sentence of natural language to another person and thereby manages to bring about a sharing of a belief. What if they don't have the same language of thought? How is he going to characterize that?

Another reason I think that the knowledge level is autonomous and irreducible (and important) has to do with the point that Brian Smith was getting at. I really can't take the line on designation and on reference that comes out of Professor Newell's account of symbols. We get the idea that a symbol designates if it gives access to a certain object or if it can affect a certain object. And this almost looks all right as long as what we're talking

about is internal states. If you have a symbol that you want to say designates some subroutine and you say that the proof of the pudding that it designates that subroutine is that when it is tokened in the system it calls that subroutine, it actually brings it on the scene and gives you access to it or it changes something in it, then it looks all right. But of course the real problem is that that isn't what reference is all about. If that were what reference were all about, then what would we say about what you might call my Julie Christie symbol problem. I have a very good physically instantiated symbol for Julie Christie. I know it refers to her, I know it really designates her, but it doesn't seem to have either of the conditions that Professor Newell describes, alas. How do you solve the Julie Christie problem, the problem of intentionality, how do you get the aboutness into your physical symbol when the physical symbol is to refer to something outside the physical symbol system in which it resides?

Here I think there are two fundamentally different ways that are being considered by various people. Intentionality is an ancient problem, and it is very murky if you try to understand it in the old-fashioned Brentano sense. One response to the murkiness is to declare that intentionality is just aboutness, and we have one clean model of aboutness: the reference of terms in a language. We can take the mysterious notion of the aboutness of knowledge, or belief, and try, via something like the language of thought hypothesis, to reduce the having of knowledge to the having, quite literally, of formuli, mentalese sentences, in the head. And then the problem of the aboutness of knowledge reduces to the problem of the semantics of the language of thought. Then we call on linguistics to solve that problem for us, so if we have the semantics of the language of thought, we can do the semantics for the symbols at the symbol level.

I suggest instead we should try to solve the problem the other way around. You can't really make sense of something as a symbol in a system unless you've got lots of other symbols in the system and they're interacting in all of their various ways. The question of whether they are symbols and if so what their reference is can only be answered by going back up to the knowledge level, and seeing whether or not the activities of these symbols subserve the control activities of a system which is understandable as a system that has knowledge. Then you see what the knowledge is about by looking out in the world, to see what things in the world are dealt with knowledgeably; by having understood the intentionality of the system at the knowledge level, you can then go back down to your symbol level and say, "Well" (and here is where approximation comes in) "it's because these actual physical patterns play the roles they play in subserving the control of this system which has this knowledge that we claim that this little bit in this symbol really is about, as it might be, Julie Christie." What I'm suggesting is that in order to do anything remotely like procedural semantics

you can't do it directly at the symbol level at all; you have to go back up and talk about the global-knowledge level characterization of the system. Once you see what it has beliefs and desires about at that global level, you can go back down and talk about the interpretation of the physical symbols.