1994

Connections, Symbols, and the Meaning of Intelligence

Peter M. Asaro '94
Illinois Wesleyan University

Honors Thesis:

Connections, Symbols, and the Meaning of Intelligence

Peter M. Asaro

Illinois Wesleyan University

Submitted: May 5, 1994

Running head: Intelligence

Peter M. Asaro
9-16-93

Honors Thesis: Statement of Intent

I intend to pursue an undergraduate thesis for honors through the Psychology department during the 1993-94 academic year. At this time, I wish to investigate Connectionist Theory, its assumptions, and its implications upon psychological methodology. More specifically, its theories regarding semantic processing. This will involve an investigation of neural networks and connectionist models of cognition and as both demonstrations of cognitive theory and as predictors of empirical findings.

I am currently a Philosophy major with an emphasis in philosophy of the mind I plan to pursue graduate work in this emerging inter-disciplinary field.

I have chosen Dr. Lionel Shapiro of the Computer Science and Psychology departments as my research advisor due to his experience in Artificial Intelligence. I feel that my honors review board should also reflect the inter-disciplinary nature of my research. In addition to Dr. Shapiro, I wish to have Dr. Laurence Colter and Dr. Ann Baker of the Philosophy Department and Dr. Susan Anderson-Freed of the Computer Science department.

Honors Thesis:

Connections, Symbols, and the Meaning of Intelligence

Peter M. Asaro

Illinois Wesleyan University

Submitted: May 5, 1994

Revised: May 18, 1994

Running head: Intelligence

With the dawning of the computer age in the early 1950's, researchers realized that computers were able to do more than mathematics. By 1955, Allen Newell and Herbert Simon were arguing that computers could instantiate the same functions as intelligent beings and Artificial Intelligence was born. AI embraces the Physical Symbol System hypothesis of intelligence, first formulated by Newell and Simon:

> *The Physical Symbol System Hypothesis.* A physical symbol system has the necessary and sufficient means for general intelligent action.
>
> By 'necessary' we mean that any system that exhibits general intelligence will prove upon analysis to be a physical symbol system. By 'sufficient' we mean that any physical symbol system of sufficient size can be organized further to exhibit general intelligence (Newell & Simon, 1981, p. 41).

There are two assumptions which are essential to this hypothesis: the first is that intelligence is a matter of performance, or functional intelligence, and the second is that mental representations can be encoded as the symbols of computational machines. This hypothesis represents the position of the Symbolic approach to AI.

More recently, debates in AI have focused on the implications of Connectionism. Connectionism is the hypothesis that distributed computations are capable of instantiating intelligent functions without relying on the representational character of symbols, but rather on the computational states themselves which are called *distributed* representations (Haugeland, 1991). This distinction puts connectionism at odds with symbolic theory. The current debates tend to be over which theory will yield intelligent systems--symbolic or connectionist? But as we will soon see, this really amounts to a debate over which representational scheme is required for general intelligence.

In this paper, I examine just what claims are required for the symbolic and connectionist hypotheses, and whether these claims are valid. I begin by examining the claim that general intelligence is essentially intelligent behavior--the claim of functionalism. I find that functional intelligence is an insufficient condition for general intelligence, which means being truly intelligent requires more than the ability to pass a Turing test. Both connectionist and symbolic theories argue that they can achieve genuine intelligence if they introduce the appropriate representations in their systems. I therefore examine the prospects of the two representational schemes--whether symbolic or distributed representations can provide sufficient conditions for general intelligence. I argue that neither representational scheme can be sufficient for intelligence. This implies that neither approach will yield genuinely intelligent systems. I conclude with a discussion of what symbolic and connectionist research is capable of achieving in light of my arguments. While AI, as traditionally conceived, cannot fulfill its goal of artificially creating intelligent systems, it can still do meaningful research. AI can very successfully model certain aspects of cognitive functions.

I. Functionalism

The most widely held paradigm in contemporary psychology is *functionalism*. Paul Churchland in his book *Matter and Consciousness* provides a clear definition of functionalism:

> According to *functionalism*, the essential or defining feature of any type of mental state is the set of causal relations it bears to (1) environmental effects on the body, (2) other types of mental states, and (3) bodily behavior. Pain, for example, characteristically results from some bodily damage or trauma; it causes distress, annoyance, and practical reasoning aimed at relief; and it causes wincing, blanching, and nursing of the traumatized area. Any state that plays exactly that functional role is a pain, according to

> functionalism. Similarly, other types of mental states (sensations, fears, beliefs, and so on) are also defined by their unique causal roles in a complex economy of internal states mediating sensory inputs and behavioral outputs.
>
> This view may remind the reader of behaviorism, and indeed it is the heir to behaviorism, but there is one fundamental difference between the two theories. Where the behaviorist hoped to define each type of mental state solely in terms of environmental input and behavioral output, the functionalist denies that this is possible. As he sees it, the adequate characterization of almost any mental state involves an ineliminable reference to a variety of other mental states with which it is causally connected, and so a reductive definition solely in terms of publicly observable inputs and outputs is quite impossible. Functionalism is therefore immune to one of the main objections against behaviorism (Churchland, 1984, p. 36)

So psychofunctionalism[1] is essentially an extended version of behaviorism that allows us to appeal to causal relations between different mental states in addition to causal relations between behavior and the environment. This would seem to avoid the problems of behaviorism's much narrower ontology; it would also seem to allow us to describe such mental processes which go into behavior formation as perceiving, interpreting, reasoning, deciding, and remembering. But functionalism will still be inadequate for describing at least one aspect of mental states necessary for intelligence--what they *mean.*

For the purposes of this paper I take psychofunctionalism to be the view that intelligence can be equated with or reduced to functions or functional states--that *being* intelligent is *behaving* intelligently. I argue that functionalism as a theory for intelligence is incomplete in that it fails to account for at least one important aspect of intelligence--the *meaning* of beliefs. This argument

[1] I shall use the term psychofunctionalism throughout this paper to refer to the theory of functionalism as a complete psychological theory as just defined by Churchland. This will be helpful in distinguishing it later from less ambitious doctrines which seek to provide a functional *analysis* or *description* of mental events without trying to provide a functional *reduction* of mental events.

rests on an assumption that I take as being generally uncontroversial. This assumption is that a necessary condition for intelligence is that behavior is essentially meaningful and that intelligent beings can access the meaning of at least some of their beliefs. Laurence BonJour (1991) defends this assumption in his "Is Thought a Symbolic Process" and I take his to be a convincing if unnecessary demonstration.

Given this assumption, I hope to show that psychofunctionalism is incapable of explaining the meaning inherent in our beliefs. The reason it cannot is that it cannot distinguish between those functions that are meaningful and those that are meaningless to a system. This argument rests on my showing that the only two options for a psychofunctional account of intentional semantics is to appeal to the causes of beliefs or the behavioral consequences of beliefs, and that both of these options leave intentional meaning indeterminant. The reason for this is that psychofunctionalism cannot determine which functions from causes to behaviors are intelligent and which are not, or which are realized and which are not. The result is that on this view all functions are equally intelligent. Calculators and thermostats, as functional systems, behave in the ways that they do based on *their* intentions just as much as humans act the way they do based on their intentions. This is just to say that thermostats can achieve intelligence and have beliefs if being intelligent just means functioning in the appropriate ways.[2] If my arguments here are sound, the result will be that psychofunctionalism cannot account for the fact that our thoughts

[2] The alternative conclusion might be eliminative functionalism, whereby humans don't have beliefs at all. But I find this to be unpalatable and uninteresting as well as counter to my assumption that we do have internally accessible beliefs.

are meaningful, and hence functional intelligence is an insufficient condition for general intelligence.

II. Functionalism and Intentional Indeterminacy

In trying to express his concern that philosophy of the mind is not addressing the issues revolving around consciousness, John Searle presents the problem of meaning in psychofunctionalism (Searle, 1990). Basically, psychofunctionalism cannot precisely fix the meaning of mental states--more specifically the subjective meaning that a belief has for the believer. Searle's despair is that this has led to the general avoidance of the consciousness "problem" and thus his argument is presented in terms of consciousness. Searle seems to think that by turning certain brain states into conscious states, we will easily and decisively explain mental phenomena. I think his conclusions are mistaken but find interesting implications to be drawn from his arguments. I have thus adapted his lucid discussion for my own purposes. However much of the argument remains his, I doubt he would endorse my ultimate conclusions. My concerns in this portion of paper are to illuminate the failed attempts of functionalism at explaining thought contents or meanings, while Searle is concerned with explaining subjective conscious experience.

Searle reminds us that in explaining any phenomenon, we need to address three aspects of it: its ontology, causality, and epistemology. The fundamental problem that strict behaviorism ran into was a confusion between ontology and epistemology (Searle, 1990). Noting that the only objective, empirical method for investigating human thought is to study the overt behaviors of people, the behaviorists quickly turned to explaining thought in strictly behavioral terms. They took what was an epistemological constraint and turned it into a metaphysical conclusion about the mind. According to behaviorism, mentality is *nothing more than* behavior and dispositions to behave in certain ways.

Functionalism is guilty of making the other possible mistake: confusing causation and ontology. By observing that the universe-according-to-physics is nothing but a complex interaction of causal sequences, the functionalists concluded that so too must mental events be actors in the causal play. So minds have inputs, such as sense experience, and outputs, such as behavior, and mental states that can influence other mental states. Including mental states allows psychology to include beliefs and desires in the story of what goes on between the inputs and outputs. On this construal, mentality is *nothing more than* having the proper causal relations between inputs, outputs and mental states. This is to say that there is nothing more to a thought or idea than its causal significance or functional role.

I think both of these confusions result from scientific sympathies to the assumptions of physicalism. Physicalism would have us believe that our ontology contains only one category (the physical) and we can thus ignore or assume this phenomenal aspect for all phenomena.

There is thus a faith in the sciences that we can understand all phenomena in physicalist terms.[3] In trying to be accepted as a genuine science, psychology has largely adopted this physicalism. Public opinion and ever advancing research in psychology have consumed many scientists and philosophers in the chore of trying to jury-rig physicalism's mental ontology to fit with psychological theory and hence physicalism.

Psychofunctionalism is appealing to the physicalist because it is a way of describing psychological phenomena or descriptions as supervening[4] over physical descriptions. A functional description avoids the problems of physical descriptions in identifying meaning with specific physical states by identifying meanings with functional states. This amounts to saying that certain physical states alone are not sufficiently meaningful, but a physical state properly "hooked-up" in a larger system can be meaningful. This is achieved by the grace of functional equivalence. Functional equivalence is the idea that it does not matter how we physically instantiate a function; as long as the function is the same the meaning is the same.[5]

[3] Interestingly (and perhaps fittingly) enough, it is probably the theoretical mathematicians and physicists who are the most likely scientific group to disbelieve logic and physicalism respectively. Fuzzy logic is trying to solve the paradoxes created by formal logic and set theory while particle physics has demonstrated that electrons and photons have indefinite ontological status. Electrons are subject to the Uncertainty Principle while photons can be waves or particles but not both at the same time. Physicalism may be in vogue, but it doesn't necessarily match all of the empirical evidence available and is not embraced by all physical scientists.

[4] Supervenience is a notion by which we can give generalized functional descriptions independent of precise physical descriptions without having to appeal to ontologically distinct entities such as minds.

[5] If the function is, for example, the addition function of arithmetic, it doesn't matter if we carry it out on our fingers, an abacus, a calculator, a supercomputer or the human brain if we get the same outputs for the same inputs, then the functions instantiated are the same. Artificial

Some functionalists contend that some properties of mental states such as meaning simply supervene upon functional states. There are many versions of supervenience, but I think it is safe to say that all of them ultimately say that meaning is something that is causally divorced from the functional process. Meaning has no causal efficacy and hence is epiphenomenal. Many functionalist theories, however, wish to maintain that beliefs and intentions affect our behavior in the ways that they do as a consequence of their meaning, and it is this sort of functionalism that is important to the Physical System Hypothesis and which I wish to address.

Searle draws a very useful distinction between *intrinsic* intentionality and *as-if* intentionality.[6] As-if intentionality is just the sort of intentionality that you can ascribe to anything whatsoever. Anything's behavior can be explained as if it were a conscious thinking thing. As Searle demonstrates:

> Water flowing downhill behaves *as if* it had intentionality. It *tries* to get to the bottom of the hill by ingeniously *seeking* the line of least resistance, it does *information processing* in order to *calculate* the size of the rocks, the angle of the slope, the pull of gravity, etc. (Searle, 1990, p. 274).

If you are at all uncomfortable with saying that everything in the world is thinking about what it is doing, and you should be, then there needs to be a kind of intentionality that only truly

Intelligence is founded on the notion that we can design computers to carry out those functions which make humans intelligent and strong AI theorists argue that such a computer can then be said to have intelligence in virtue of its capability to perform those functions.

[6] Philosophers like to talk about minds in terms of intentionality. An intention is just the believing in or desiring for a particular state of affairs in the world. This generally takes the form of asserting or believing some mental proposition. Having intentionality just means having genuine beliefs or desires. See Searle (1983) for a thorough study of intentionality.

mental things can have, namely intrinsic intentionality. Tied up in this notion of intrinsic intentionality are our deepest notions of what mentality is. Thought is essentially meaningful and this meaning potentially, plays a role in conscious experience. We call something intrinsically intentional only if it has the capacity to entertain and understand genuine ideas which is to say it has some sort of colorful mental life.

The inability of functionalism to capture the subjective aspects of thought has been discussed at length in the literature regarding *qualia*. The idea here is that there are qualitative mental experiences such as color sensations that are essentially subjective, and private. Since we can never experience someone else's sensory *qualia*, we can never know what someone else's *qualia* are like nor can we be justified in saying what has them and what does not. The functionalist's short answer to this problem is to regard such subjective experiences as non-essential, epiphenomenal, or altogether non-existent. If there were properties inherent to a thought that cannot be realized in the functional roles of that state, then those properties aren't a part of the explanation of why that thought has the consequences it does. Since there are no objective phenomena, and the subjective phenomena do not seem to be required to explain anything objectively relevant, *qualia* just do not matter to science (perhaps it is an aesthetic issue).

I believe that this problem extends itself beyond the *qualia* issue to the content issue. While it may not be necessary to an account of the mind to specify what our thoughts "feel" like, it will certainly be necessary to specify what those thoughts *mean*. The fact that we can understand anything, that we can formulate a philosophical theory and even functionalism, implies that our thoughts have meaning such that we can access and understand them. It turns out that meaning is actually paramount to a pretheoretical notion of psychology. Almost all psychological theory, and especially functional psychology, is based on the notion that our behaviors are based upon our beliefs and more specifically that we frequently behave the way we do based upon what our thoughts mean. So, unlike *qualia* which may not need to be part of the psychological story, content and meaning are absolutely essential for an understanding of the mind-brain and human history in general. But as Searle points out, meaning is just as subjective and externally inaccessible as *qualia*.

Searle describes thoughts as having a particular aspectual shape. The aspectual shape is the precise meaning one has in mind when one entertains a thought. For example, Searle entertains a desire for a glass of water but does not entertain the desire for a glass of H_2O. These intentions are coreferential yet differ in their mode of reference or aspect; we can imagine someone who does not know that water is H_2O and would not agree in the equivalence of these beliefs. Searle is clear that "This aspectual feature must matter to the agent. It must exist from his/her point of view" (Searle, 1990, p. 275). This is really just another way of talking about the referential opacity of intentions and narrow content.[7] This is to say that there is meaning to

[7] The semantic opacity of thoughts is simply the idea that we can understand the meaning of the thoughts in our head in a specific way even though there are aspects of their meaning which we do not necessarily understand. The fact that we can believe something about the Evening Star but not believe the same thing about the Morning Star is an example of this opacity. This is because we might not be aware of the fact that the Morning Star is really the same thing as the Evening Star. Narrow content is the specific content of a given belief. By specific, I mean the content of a belief that can be about the Evening Star and not the Morning Star.

a belief which has efficacy in the reasoning processes of the subject. There may be many other meanings that a belief could have, but there is at least one which the subject actually does have, and this is the one that determines behavior.

The consequence psychofunctionalism's inability to describe subjective intentional ascriptions is that it results in intentional indeterminacy. The fact that psychofunctionalism results in intentional indeterminacy is due to the methods of functional analysis. There are many equally accurate levels of objective functional description for any complex functional system. A functional description of the kidneys could describe their role in any of a number of different bodily systems (the circulatory, excretory, or endocrine). In describing the functional role of a process or mental state, it becomes necessary to define the scope of the system for which the role of the function will be determined. A linear transformation algorithm in computer science is a well defined function, but this same algorithm will have a different functional role when it is considered as part of a ray-tracing program then it will in a multi-variable statistics program. How one defines the system determines in part what the functional role will be; part of defining the system includes division into subsystems, the situational context which the system or subsystem is in, and the purposes we have in seeking the function. Because we can analyze a system for different purposes, we can generate equally valid functional roles for the same process or state. The indeterminacy that results from functional role analyses will give us a great many, if not potentially infinite, number of possible 'meanings' for a given state.

The basic tenet of psychofunctionalism being that everything is part of a causal chain with beliefs being steps in some causal chains linking stimuli to behaviors, it becomes very important how we "cut-up" the causal chains for our descriptions of beliefs. Fred Dretske, an advocate of the functionalist cause, describes the difficulty in this process:

> One can be easily misled into thinking that the cause of a behavior is necessarily the cause of output. And once this confusion is in place, one will have no option but to identify causal explanations of why we *do* the things we do with causal explanations of why our body *moves* the way it does. One will, in other words, have succeeded in confusing psychological explanations of behavior with neurobiological explanations of motor activity. Reasons--our *thinking* this and *wanting* that--will have been robbed of an explanatory job to do, reasons--and by this I mean the beliefs, desires, intentions, and purposes that common sense recognizes as reasons--will have been robbed of any scientifically reputable basis for existing...Thinking about behavior as a process having output as its product, is, if nothing else, a useful way of avoiding this mistake (Dretske, 1988, p. 36).

Dretske is really doing two things here. The first is setting up a distinction between intended behavior and actual response output which he uses as a starting point for his covariance theory of meaning. The second is to set up a supervenience of functional belief descriptions over causal sequences. I will address covariance theories of meaning in the next section, so I am only concerned here with his second objective which is the embedded nature of intentions and causes.

The supervenience of beliefs as functions involved in causal chains is quite important to us because it matters a great deal what we pick out as the intention in a causal chain of events. This is due to the fact that the causal chains of the world stretch all the way back to the Big Bang and forward to seeming eternity, as well as intertwine and overlap one another, while our intentions are temporally limited and singular. Humans are limited beings and a human intention is generally aimed at bringing about a *particular* state of affairs or having a *particular* consequence. The indeterminacy of psychofunctionalism stems from our inability to determine precisely which functional analysis will give us the particular intention that an individual has.

Let's take Dretske's own example to clarify what I mean:

> Suppose Clyde accidentally knocks his wine glass over in reaching for the salt. The glass falls to the carpet, breaks, and leaves an ugly red stain. Clyde has done a number of things: moved his arm, knocked over the wine glass, broken it, spilled the wine, and ruined the carpet. He did all these things--the first intentionally, the others inadvertently. In speaking of these as things Clyde did, we locate the cause of these various events and conditions in Clyde. In each case the effect is different--arm movement, the glass toppling over, its breaking, the wine's spilling, and the carpet's being stained--and hence the behavior, the process having these different events or conditions as its product, is different. But the causal origin, some event or condition in Clyde, is the same (Dretske, 1988, p. 37).

Dretske tells us that Clyde intended only to move his arm while the other events were unintended. This clearly seems to be the case, but how is the psychofunctionalist to determine precisely what Clyde meant to do when he moved his arm, what he actually intended? Perhaps an appeal to previous cases in Clyde's causal history such as Clyde's desire for salt or the fact that the salt was behind the wine glass could more strictly determine his beliefs in this case, but the next case will show that it will often be impossible to determine one's intentions based on causal features.

Suppose you are playing a game of chess with me. We have established a small wager and you have no idea how good I am at chess. The game is going quite well when I make what you find to be an absurd move with my King's Knight: I expose myself to a checkmate in two moves. Now you get wary and try to figure out my intentions based upon the function of this move. First you may think I made a mistake: I intended to threaten your Queen and failed to see that I had opened myself up to a checkmate. Then you may come to believe that such an obvious flaw must be a trick, and you begin looking for the ways in which I might be trying to get the better of you. A narrow view of my intentions will yield my intention to make a particular move. But chess is a game of strategy, and a particular move is generally part of a broader strategy which I have in mind, the broadest being that I intend to win the game. But perhaps I am a sort of chess hustler, and I am intending to throw this game in the hope of getting you to wager double on the next game; so there is a chance that I do not have even the broad intention to win. A mid-level analysis might be that I am simply trying to execute some gambit that you have never seen. Which analysis of my move is the correct one? I argue that there is no way for a functional account of the chess game to determine what my intention is because each level of analysis is equally plausible in describing the causal events, but I only have one set of intentions.

The psychofunctionalist's first response is probably going to be that I have multiple, overlapping, or embedded intentions when I move my piece. It will probably go something like: I intend to win the game, I believe that my avant-garde gambit will win me the game, I believe that moving my King's Knight will result in an execution of my gambit, and so I intend to move my Knight in order to satisfy all of these intentions. My intentions are many, but my behavior is singular. The issue at hand is how to precisely determine all and only those intentions I have in moving my Knight by means of a purely causal account of the situation. It may be the case that I am a novice, my earlier moves were beginner's luck, and this move was capricious; I only intended to move my Knight and hoped it would help me to win. The psychofunctionalist will have to appeal to my previous chess experience to judge how much I might know about the game, but even this cannot conceivably be sufficient to determine my intentions, so long as the game situation is sufficiently novel to me. Even if I am a grandmaster, I could be trying a newly devised gambit or have overlooked a piece.

There is, in fact, a much deeper issue lurking here. Even if my intention were merely to move my Knight, what does this intention *mean*? Without the situational rules surrounding the game of chess, the movement of chess pieces is meaningless.[8] Given the goals and rules of the game, a particular move has significance, or a role in the overall game. Given that this game is part of a social interchange between you and me, it is also subject to the rules of personal protocol, and the laws of the city, state, and country we are playing in.[9] A large part of determining my intentions is going to involve determining my understanding and awareness of the rules of the game. It seems that I might intend many different things by my move, but I surely do not intend every thing it signifies. Dretske mentions this in his book *Knowledge and the Flow of Information*:

> It makes little sense...to speak of *the* informational content of a signal as though this was unique. Generally speaking, a signal carries a great variety of different informational contents, a great variety of different pieces of information, and although these pieces of information may be related to each other (e.g., logically), they are nonetheless *different*

[8] Physical acts which are given meaning by their role in defined systems such as games or laws, ruled-governed acts, are called Conventional Acts. There is a sizable literature on Conventional Acts and Speech Acts which has implications for belief and linguistic content which I do not have the space to address in this paper. See Searle (1969) for more on these implications.

[9] If you happen to be my boss or superior, or have a bad temper, it may be part of the rules of social protocol for me to throw the game. It may also be a subtle slight against you to throw the game in an obvious manner in order to indicate that I am deliberately losing. Because we have placed a wager on the game, all of my intentions regarding the game may rightly qualify as "illegal" according to the gambling laws of the state we are playing in. Do I thereby *intend* to commit a misdemeanor by moving my Knight? We could probably go on and analyze this game's significance in politics, economics, or world history, but the point is that it matters a great deal where we draw the line as to what my move signifies or represents.

> pieces of information... This feature of information serves to distinguish it sharply from the concept of *meaning*--at least the concept of meaning relevant to semantic studies of language and belief. The statement "Joe is at home" may be said to mean that Joe is at home (whatever the person who made it happened to mean or intend by making it). It certainly does not mean that Joe is either at home or at the office. The statement *implies* that Joe is either at home or at the office, but this is not what it means. On the other hand, if the statement carries the information that Joe is at home, it carries the information that Joe is either at home or at the office. It cannot communicate the one piece of information without communicating the other. One piece of information is analytically nested in the other (Dretske, 1981, p. 72).

If we look closely at what Dretske is telling us, it is that there are many implications or significances to any behavior. What a behavior, especially a linguistic behavior, signifies is not what it *means*. What it means depends on what it was intended to mean. What we can figure out empirically (maybe) is what a behavior signifies. What the functionalist is claiming is that these empirical methods can give us the intention behind the behavior. But a determination of what my intention is requires an understanding of what my move means and, ala Quine, this cannot be had without first knowing what my intention was--a clearly circular conception of intentional content.

The indeterminacy of intentions rests on the inability to determine which level of analysis will give us the intentions that an individual is using as a basis for his/her actions. This has led to a philosophical notion termed narrow content. Narrow content is opposed to wide content.[10] Wide content is a notion put forth by Hilary Putnam (1975) according to which there is an aspect of meaning to our thoughts to which we do not have access. Roughly, intending to buy a lottery

[10] These terms are roughly equivalent to the opaque and transparent construals of meaning discussed earlier. Wide content is content transparently construed while narrow content is content opaquely or nontransparently construed.

ticket is usually the intention to buy a winning lottery ticket, but there is no way of knowing which you are actually buying so you can't be sure which kind you intended to buy in the wide sense--it depends on actual states of the world which you don't have access to (at least until after you've bought the ticket). Narrow content is the aspect of intentional meaning to which we necessarily have access; indeed it is the aspect which is argued to determine behavior (Fodor, 1981). According to this construal (or scope) of content, it can be your intention to shoot "the deer" even though "the deer" turns out to be your hunting companion, Bob. The fact that "the deer" is your friend doesn't affect the fact that you thought it was venison.

This amounts to a demonstration that there are different ways to ascribe meaning to mental states functionally. This motivates Fodor to argue that it is only the narrow construal that should be utilized in psychology. But there are also different ways to identify narrow content, and it turns out that we cannot strictly determine even the narrow contents, much less intentionality. As Robert Stalnaker (1990) puts it:

> Even if we could individuate thought or beliefs independently of their contents, this would not necessarily suffice to yield a determinate narrow content for them by the procedure I have suggested [which is considering contents for other possible worlds]. Suppose we identify mental thought tokens by their physical or syntactic properties. These properties surely will not be sufficient to determine even the narrow content of the thought token (Stalnaker, 1990, p. 135).

Stalnaker goes on to argue that to limit the causal relations of thoughts to "in-the-head" functional relations, as narrow content seems to require, will result not in a narrow content but in no content whatsoever. Stalnaker is an advocate of externalism and his argument claims that the thoughts in our head ultimately depend on the outside world for their meaning. I will discuss the different approaches of internalism and externalism in the section on representation.

The problem for narrow content is ultimately one of perspective; when trying to determine someone's narrow intentional state, different perspectives will yield different intentions. Ned Block (1991) points in the direction of this problem with the following example:

> If you and I both say "My pants are now on fire," we express contents (truth conditions) that are importantly different. What you say is true just in case *your* pants are on fire; what I say is true just in case *my* pants are on fire. Nonetheless, there is also an important semantic commonality (Block, 1991, p. 34).

Block goes on to distinguish the *character* of a thought from its *content*, where the character is the functional role of the thought and the content is the actual subjective meaning of the thought. Thus, it is really the narrow character which gets involved in behavior:

> Character is relevant to psychological explanation in a way that content is not. Suppose you and I think thoughts with the same character, thoughts that we both would express with "My pants are on fire," and as a result, we jump into a nearby pool. The common character of our thoughts would seem to be part of the explanation of the commonality of our behaviors. By contrast, if we both had thoughts with the same content, the content that I express with "My pants are on fire," we would have done quite different things: I jump in the pool, but you don't jump--you push me in (Block, 1991, p. 34).

Block is arguing that it is the character that can be determined by psychologically objective observations, not the content that the subject has.

This goes back to Searle's arguments about an objective science of subjective experience. Quite simply, we are only inclined to call these thoughts similar from an external, objective perspective--it is the thought's character which is easily generalizable. The functionalist would say that the thoughts are similar because they result in similar behaviors. But subjectively, I think about myself in very different ways than you think about yourself. Thus, our thoughts are importantly different in content. The problem of perspective with regards to narrow content is that a narrow determination of content will have to be in terms which are defined only for the system under scrutiny. While it is conceptually possible to derive the necessary terms in a holistic fashion, these terms will be meaningless when we consider the system objectively--we cannot use these terms to describe any other system. Thus, even narrow content, which is supposed to be the subject's content, is underdetermined by functional analysis.

Block examines the consequences of choosing either perspective for psychology. He argues that there are only two possible routes for narrow content (he doesn't question his functionalist presuppositions). On the first account, narrow content is a matter of the thought's function in a given situation and content collapses into the syntax for that situation. On the other hand, narrow content can include the content of all the employed terms as they are used in the system in general. On this account, we end up with holistic semantics. But according to holistic semantics, every individual will have slightly different understandings of their proper name terms giving different subjective meanings to all thoughts to such a degree that no one is ever thinking the same thing as someone else or even at two different times. Block argues that this is an unfortunate result, but necessary if we want to get at true subjective content.

Besides the infelicitous result that nobody ever has the same belief as anyone else (nor can we account for belief *similarities* between individuals), there are deep problems for a holistic semantics. First, there is no reason to think that anyone's holistic semantic "map" will be complete or self-consistent. Indeed, it seems that we very often have inconsistent beliefs about something, yet these beliefs are still contentful, and holism cannot account for this. Consider X's system of beliefs which does not distinguish real numbers and integers, and which contains the following beliefs:

1) There are infinitely many numbers between two sequential whole numbers.
2) The numbers between 1 and 5 are 2, 3 and 4.
3) 1 and 5 are whole numbers.

At a given time, X might assert any either one of these beliefs, but taken together they are formally inconsistent. If we appeal to the entire system to give belief 2) meaning, it will have an ambiguous meaning. But it seems that a system can hold belief 2) unambiguously. To avoid this paradox, we must say that we should only use the relevant conceptual roles for meaning determination. But choosing the relevant roles requires that we already know what the belief means. Furthermore, holism cannot account for how any of the terms have in-the-world or objective meaning without switching perspectives, which is highly problematic.

I believe I have made it quite clear from this discussion that strict psychofunctionalism cannot provide the sufficient conditions for intentionality and hence intelligence. This just means that functioning in intelligent ways does not imply intelligence because functioning can occur without any appeal to meaningful entities. This is one of the motivations for the AI camp to argue that while functions alone are not sufficient, if the functions are performed using meaningful symbols, representations, then this will be sufficient for intelligence. With this I turn to an examination of representationality.

III. Representations and Symbols

We have already seen that performance alone is not sufficient for genuine intelligence.

This leads us to the second important aspect of the Physical Symbol System Hypothesis--the symbol part. The Physical Symbol System Hypothesis contains the basis for the distinction between computationalism and functionalism. The distinction lies in the fact that computations are functions that involve symbols, and these symbols are representational. The idea is that we can avoid the as-if functional ascriptions we've already seen if we add the condition that the functions must be carried out over meaningful symbols. The symbolic AI theorist thus argues for the Information Processing Theory of intelligence that holds that the symbols processed qualify as information about the world, and thus intelligence is nothing more than information processing. But accounting for the representational or informational character of computational symbols does not turn out to be as simple as the computationalists would hope. If the computationalists want representations to explain intelligence, they are first going to need an account of how to determine what computational symbols represent in a way that can explain intelligence. I argue that a computational approach cannot succeed in providing a useful account of the representational character of a symbol or computational state.

Before I begin, there is a distinction that needs to be drawn. Discussions about mental representation usually involve mentioning or implying one of two opposing views, internalism or externalism. These concepts are well expressed by Goschke and Koppelberg (1991):

> ...we will now consider the two main philosophical research programs for a naturalized theory of semantic content. Both approaches characterize representations relationally, but they differ on the question of what kinds of relations are to be taken into account. The basic idea of the first research program, known as *externalism*, is to explicate representations in terms of a lawlike dependency relation between a representing event and the external event it represents. Within this approach we find the different versions of correlational theories. The basic idea of the second research program, often called *internalism*, is to explicate representation in terms of dynamic relations between events

> internal to a representing system. This approach subsumes many different versions of conceptual role or functional role semantics (Goschke & Koppelberg, 1991, p. 132).

So there are actually two different routes for the computationalist to take.[11] The symbolic AI theorists are generally trying to establish some form of internalism, whereby the physical symbols are internally meaningful to the system. The connectionists, on the other hand, are trying to establish some form of externalism, whereby the system succeeds in representing external phenomena. Before I address the symbolic and connectionist approaches to representation, I will first investigate the potential of these more general programs.

The symbolic internalist approach has long been under fire. John Haugeland (1981a) provides a good summary of the general objection:

> The idea is that a semantic engine's[12] tokens only have meaning because we give it to them; their intentionality, like that of smoke symbols and writing, is essentially borrowed, hence *derivative*. To put it bluntly: computers themselves don't mean anything by their tokens (any more than books do)--they only mean what we say they do. Genuine understanding on the other hand, is intentional "in its own right" and not derivatively from something else (Haugeland, 1981a, p. 32-33).

The argument that computers don't understand the symbols they process has had wide appeal and the most famous argument in this vein is John Searle's (1980) Chinese Room example. This is

[11] I take it that the computational approach is seeking extrinsic, relational characterizations rather than intrinsic characterizations. It is important to note that I follow Goschke and Koppelberg in using "internal" and "external" to distinguish two kinds of extrinsic content. I have never seen a computationalist attempt to characterize content intrinsically. Accordingly, when I talk about symbols or tokens, I mean for these to be intrinsically "empty."

[12] Semantic engines are defined by Daniel Dennett (1981) as automatic formal systems which process over semantically meaningful tokens. This is the same as an information processing system.

the classic story of a man locked in a room with a list of English commands for manipulating Chinese characters. He succeeds in answering questions in Chinese, but only by following the English rules. He does not himself understand Chinese, nor does his rule-following qualify as giving him an understanding of Chinese. The strongest and most popular (internalist) proposal for getting around the Chinese Room problem is to argue that if the man were given the complete rules of the Chinese language, then he would thereby understand Chinese. This amounts to saying that the symbol's meaning is a matter of the system on the whole; the man doesn't understand Chinese because he doesn't have the rules for Chinese, just a small set of specialized cases.

The notion that we can make symbols meaningful by appealing to an individual symbol's relationships to other symbols in the system is held by various versions of holistic or conceptual role semantics. Conceptual role semantics holds that the semantic content of a representation is determined and exhausted by the relations it bears to other representations in the system. As Laurence BonJour (1991) puts it:

> Thus, although the individual representations..., taken apart from their actual functioning in the person's cognitive processes, could have meant many things other than what they actually mean, the claim is that when these representations are taken together in the context of the overall inferential pattern, they can each have only the specific content that they actually have (BonJour, 1991, p. 341).

The problem with conceptual roles in accounting for representational content is that the only property available to internal representations is their relational patterns. BonJour goes on to argue that even if the relational pattern that bears upon an individual symbol were sufficient to uniquely distinguish it from the other symbols in the system, the pattern will not be sufficient to make the symbol meaningful. Basically, the pattern does not itself constitute meaning for we, at least, do not understand our thoughts in this way:

> ...it is at least clear that my awareness of what I am thinking is not an explicit awareness of the inferential/causal pattern as such, as shown by the fact that I would find it very hard in any particular case to say what that pattern is (BonJour, 1991, p. 344).

Insofar as thoughts are meaningful or representational to the system itself, they are not merely extrinsically representational. Thoughts are not understood in terms of their relations; they are understood in themselves--intrinsically. I find this to be devastating for any internalist account of symbolic meaning.

The other approach, externalism, seeks to relate the symbols to external phenomena as a means to impart them with representationality. Robert Cummins (1989) provides what I have found to be the most thorough and straightforward presentation of the various externalist accounts of mental representation. All externalist accounts of how computational states or symbols achieve a representational content ultimately come down to a version of covariance[13] theory (Cummins, 1989). The idea here is quite simple: a mental state or token represents that which regularly causes it. Thus, my being presented with frog stimuli will cause me to create "frog" representations. The problem for all theories of covariance is to explain misrepresentations. What happens if a toad causes me to have a "frog" representation? It does not matter how clever our causal story about how toads can imitate frogs to our intelligent system gets, the issue is what "frog" actually represents in that system.

[13] Also called correlational theories.

The "frog" representation, whether it is instantiated as a binary numeral, a linguistic token or a pattern of activation, has no intrinsic or *a priori* content. This is just to say that all tokens are equal; we could call telephones "phones," "dogs," "xorpts," or "Truman Capotes." Social conventions fix the representations of our words, but how are we to fix our mental representations (i.e. make sure our "frog" thoughts are about frogs and not toads or telephones)? A token only achieves the representational status it has by its causal relations. So long as a token is representational purely in virtue of its being part of a formal causal system, its representational content will be exclusively determined by that system. If the system produces "frog" when presented with frogs or toads, then "frog" means |frog or toad|.[14] And we could imagine a strange enough causal scene in which a telephone caused a "frog" (it was dark and the telephone made a croaking noise) and we get "frog" meaning |frog or toad or telephone|.

Cummins points out that all covariance theories have acknowledged this problem and attempt to solve it in essentially the same way--idealization (Cummins, 1989). Idealization tells us that the token's meaning is fixed by what it represents under ideal or optimal causal conditions. So our "frog" won't mean |frog or toad or telephone| because telephones will only cause "frog" in less than ideal conditions. But as Cummins further points out, there is no way to spell out precisely what those conditions are, especially from the intelligent system's position. Idealization is ultimately circular:

> We're going to have covariance only when the epistemological conditions are right. Good epistemological conditions are ones that are going to get you correct (or at least rational)

[14] For the remaining discussions, I will use quotes for tokenings, absolute value signs for what the tokens represent and nothing for the "in-the-world" causes of tokenings.

> results. Conditions like that are bound to require semantic specification (Cummins, 1989, p. 55).

We can't specify ideal conditions until we have fixed the representation (otherwise, what is it are the conditions ideal for producing?) but we can't fix the representation until we've specified ideal conditions. Neither Cummins nor I see any way around this problem.

There have, however, been several attempts to get around the idealization problem, all of which Cummins has clearly explained and definitively discredited. Among them I find only two worth mentioning again. The first of these is Adaptational Role semantics which contends that the covariance could be reliably established by evolutionary history.[15] The first obvious problem with this approach is that a determination of a token's meaning could require a complete evolutionary history of that system, perhaps all the way back to pre-organic history or even the Big Bang. The organism itself would never have this sort of access so it is not clear how the organism might understand this kind of representation, but we could suppose that this process has fixed the token's meaning objectively.

The decisive problem turns out to be that adaptational role semantics is inconsistent with computationalism due to computationalism's insistence on functional equivalencies.[16] If we

[15] Cummins points out that adaptational role semantics was not designed as an answer to the idealization problem *per se*. His arguments are designed to show that this approach will not work for a computationalist inclined to use it in this way. It is thus not an argument against adaptational role semantics, just against using it in a computationalist theory.

[16] Functional equivalence requires only the same input/output relations for each system. Thus we could instantiate functionally equivalent systems in a digital computer, a Turing machine, a biological organism, or using gears and springs.

produce a system functionally equivalent to one which has adaptationally determined representations, are we going to say that its representations are meaningful or not? If we say that they are not meaningful because this system did not evolve correctly, then we have denied computationalism in general--it is no longer a token's functional significance or computational role which determines its meaning. In order to maintain computationalism one must say that if the evolved system is meaningful, any systems functionally equivalent to it are also meaningful.[17] But if we do this, it seems that we end up ignoring the evolutionary history--if we can make a system functionally equivalent to an evolutionarily determined system, why can't we make a system for which there is a possible but not actual evolutionary history? If we want to stick to computationalism, appeals to evolutionary history aren't going to help us determine representation.

I think it is important to mention connectionism here because it might seem that there is a sort of evolutionary process involved in the creation of distributed representations. Supervised training in neural networks might produce highly reliable and fault-tolerant distributed representations, but the representation is still going to be computational. The connectionist is still going to want to say that it is the connections and weights of the network which are responsible for creating the proper patterns of activation, and it does not matter how these came to be. The correctness of the representation is still a matter of the network making the proper responses to

[17] History does not matter to the computationalists as Cummins indicates, "It is which data structures you have, not how you got them, that counts. Without this assumption, AI makes no sense at all" (Cummins, 1989, p. 84).

the given inputs. Supervised training is a method for getting connection weights which are more likely to work; it is not a process for bestowing representational status upon patterns of activation.

Another attempt to get around the Idealization problem is Asymmetrical Dependence. Asymmetrical dependence, put forth by Jerry Fodor (1989) in his *Psychosemantics*, maintains that misrepresentations are dependent on correct representations in a way that can account for misrepresentation while avoiding the problems previously pointed out. This is most easily explained by example. Cummins (1989) describes this as it pertains to shrews causing mouse representations, a case of misrepresentation that should be explained by asymmetrical dependence:

> The fact that shrews sometimes cause |mouse|s[18] in me depends on the fact that mice cause |mouse|s in me. On the other hand, the fact that mice cause |mouse|s in me doesn't depend on the fact that shrews sometimes cause |mouse|s in me. Mice look mousey to me and that mousey look causes a |mouse|. But it is only because shrews also look mousey to me that shrews cause |mouse|s. Thus, if mice didn't cause |mouse|s, shrews wouldn't either. But it needn't work the other way; I could learn to distinguish shrews from mice, in which case mice would cause |mouse|s even though shrews would not (Cummins, 1989, p. 58).

The problem for covariance is not just to explain why shrews can sometimes cause mouse representations, but how mouse representations can be about mice. This is what covariance is ultimately incapable of doing:

> We are told that representation rests on a covariance between representation and representandum--between cats and |cats|, for example. Covariance, in turn, is grounded in a mechanism that, under the right conditions, will produce a |cat| from a cat. According to the CTC [computational theory of cognition], the mechanism in question can be understood only by appeal to inner representations, for the mechanism in question is

[18] Cummins uses the absolute value signs around representations to distinguish them from actual instances of an environmental condition.

> one of inference from stored *knowledge. It follows that in order to understand the mechanism that the CTC invokes to explain the covariance between cats and |cat|s we must already understand representation and the explanatory role it plays in mental mechanisms. And that, by my lights, is enough to undermine the power of the covariance theories to help us to understand the nature of representation in the CTC.
>
> The problem, of course, is that it isn't enough to avoid intentional/semantic vocabulary; you must do it in a way that explains what representation is. It becomes obvious that just avoiding intentional/semantic vocabulary isn't enough when you see how easy it is. The problem, remember, was to say under what conditions cats are sufficient for |cat|s, and to do it in naturalistic vocabulary (Cummins, 1989, p. 65).

In the end, covariance turns out to be trivial. We simply find a case where a cat causes a |cat|, and then specify the mechanism that produced it and the situational conditions. This is trivial because representationality is just whatever makes something representational, not at all an interesting assertion.

The real problem for computational accounts of mental representation is inherent in our very notions of mentality and representation, and was alluded to in the latter quote from Cummins. Computationalism (Cummins' CTC) requires that representations derive their content exclusively from the relations they have to other representations or the environment. This means one of two things: either the representations result from relations to other representations, or the representations result from relations to the environment. Our intuition, and my assumption, is that representational content must be internally accessible if it going to help us explain intelligence; but neither computational approach to representation will give us this. The first, internalist content, gives us individuated relational patterns which are void of genuine content. This is rather like having a papyrus of Egyptian hieroglyphics before the Rosetta Stone was discovered--it could be a rich, complex, meaningful system or a child's scribblings and the system won't care either way. The second, externalist content, if successful, might establish law-like relations between internal events and events in the environment, but these relations will be necessarily external and internally inaccessible. This is akin to being given a doll and told it is a voodoo doll; we can stick needles in our doll but we can never conceive of the person it is supposed to represent, we don't know who it represents, and we can't be sure that it is even is a representation of something--that it is really a voodoo doll at all. In each case there might be a representational content to the thoughts, but in neither is it accessible to the system nor is it significant in determining behavior. Either way the computationalist is left with no representations that are useful in explaining the system's or general intelligence.

So it might seem there is no way to account for representation on a computationalist account. This is not really the case; the computationalists are just looking for the wrong kind of representation. In introducing his own theory of representation Cummins tells us:

> Interpretational Semantics is an account of representation in the CTC, whereas most philosophical discussion of mental representation has to do with Intentionality--i.e., with the contents of thoughts--rather than with the contents of the representations of a computational system. ...The kind of meaning required by the CTC is, I think, not Intentional Content anymore than entropy is history (Cummins, 1989, p. 88).

What the computationalists are trying to do is define the representational content of a computational system such that representational content can be identified with intentional content. It is this backdoor approach which has produced so many problems for their accounts of representation.

But for most computations, it doesn't seem that we have any trouble interpreting the symbols involved if we are given a "table of assignments" or a coherent interpretation. The problem is that the symbols *require* interpretation; they do not represent anything intrinsically and interpretation ultimately requires an intelligent system. Interpretation is the ascribing of meaning to a token, and this cannot be achieved computationally. Essentially, even if we were to build a computational "interpreter" for our system, we would need to explain the "interpreter's" intelligence which would in turn require another "interpreter". This is a homunculur *reductio ad absurdum* where our homunculus is a computational "interpreter."

It would certainly be unpalatable to think that there was no representationality, or that psychology has nothing to study and AI nothing to model. The consequence of my examination and arguments is that the having of functions and symbolic representations is not sufficient conditions for intelligence. The upshot of all this is that there is a way to conceive of cognition and representation which doesn't have the philosophical problems just presented, and there is accordingly a domain of study for cognitive psychology and AI. This is because while instantiating functions is not a sufficient condition for intelligence, we can still give a functional *description* of intelligent systems. And so I turn to a discussion of the prospects for psychology in light of the arguments thus far given.

IV. Cognitive Psychology and a Science of the Mind

I think the previous sections have shown that computationalist psychology does not and cannot study beliefs, intentions or thoughts. The objective methods of empirical science are simply ill-equipped to explain subjective experience or meaning. So what has computational psychology been studying for so long? To this I would reply that the behaviorists studied behavior and the functionalists studied functioning, but neither studied the mind. Behavioral studies are perfectly valid research strategies even though behaviorism is incomplete. Similarly, functional descriptions can be valid even though functionalism is an incomplete theory of intelligence. Behavior and functioning are correct for certain levels of explanation, but they are not the fundamental level of explanation for thought--people will always behave and their minds will function in certain ways, but neither of these activities necessarily constitute thinking. And likewise, the computationalists derive functions and compute with them, but they are not getting at the heart of thought.

The only defensible computational psychology is one that does not try to make any claims about thought content or subjective experience. But how can psychologists talk about *thinking* without talking about *thoughts*? What is required is a psychology of thought *processes*, not of thought *contents*. Fodor calls for as much in his clarification of Methodological Solipsism in a footnote to *Psychosemantics*:

> More precisely, methodological solipsism is a doctrine not about individuation in psychology at large but about individuation in aid of the psychology of mental processes. Methodological solipsism constrains the ways mental processes can specify their ranges and domains: They can't apply differently to mental states just in virtue of the truth or falsity of the propositions that the mental states express. And they can't apply differently to concepts depending on whether or not the concepts denote (Fodor, 1987, p. 158).

Fodor is concerned here with drawing a distinction between methodological *individualism* and methodological *solipsism*. The former is the notion that, "psychological states are individuated *with respect to their causal powers*," while the latter holds that, "psychological states are individuated *without respect to their semantic*[19] *evaluation*" (Fodor, 1987, p. 42). Thus, individualism is functionalism *par excellence*--semantics collapses into syntax and we are confronted with covariance again. But methodological solipsism is an empirical claim about process, and explicitly avoids making any claims about content.[20] Methodological solipsism, thus conceived, is not functionalism but merely objective thought individuation--an objective analysis of processes which are essentially subjective in content.

The conflict that needs to be resolved is between theory and methodology. We know we cannot produce a theory of mental content or general intelligence based on causal or functional accounts no matter how ingeniously we devise our system. And yet we know that there is an objective and observable aspect of thought which can and should be studied by psychology, namely thought process. This is just to say that we can always give an as-if intentional account of a system. As Haugeland (1981b) puts it:

> If one can systematically explain how an [intentional system] works, without "de-interpreting" it, it is an *information processing system* (an IPS). By "without de-interpreting," I mean explaining its input/output ability in terms of how it would be characterized under the intentional interpretation, regardless of whatever other descriptions might be available for the same input and output behavior (Haugeland, 1981b, p. 258).

[19] Fodor uses semantic here to mean such aspects of belief as actual truth or falsity and reference, he does not include meaning among these. Fodor argues that meaning can be determined via causal significances, but I see no way of doing this which makes meaning anything different than syntax.

[20] This point can be very easy to miss since methodological solipsism as first presented by Fodor (1981) is clearly applied to narrow content and intentionality. Even if we accept this later distinction, Fodor's larger project in *Psychosemantics* is to provide a complete account of both methodological individualism and solipsism. I think methodological solipsism as now clarified could only justifiably be fleshed out to look like Cummins' Interpretational Semantics--functionalism with a little "f."

This just means that we can give a cognitive description of a system even though there may be other equally valid descriptions, such as digital electronic or neurophysiological ones, available. What is called for is a theory of representation which can be used by a cognitive psychology without overstepping its bounds and becoming a theory of content.

Cummins, in providing his own account of mental representation, gives us a straightforward approach to a coherent psychology. I think it amounts to a complete theory of representation for psychology which honors methodological solipsism without carrying the burden of intentionality Fodor and Block are so eager to bear. Cummins calls his psychological method Interpretational Semantics. It begins by distinguishing cognition from thought. Quite simply:

> Some cognitive systems are not minds (not, at least, as we know minds ostensively), and many aspects of mentality are not cognitive. Cognitive science is founded on the empirical assumption that cognition (hence the study of cognitive systems) is a natural and relatively autonomous domain of inquiry (Cummins, 1989, p. 18-19).

Accordingly, Cummins' Computational Theory of Cognition (CTC) holds that many of our cognitive processes might be describable in computational terms, or specified as functions, but these descriptions don't purport to specify the actual contents of any thoughts.

Cummins' proposal is that an essential aspect of representation is interpretation, and this should be admitted by our theory. By distinguishing cognition from thought and intelligence in general, we do not have to worry that our semantics could be applied to intelligent and non-intelligent systems. Unlike as-if intentionality, we are not ascribing intentionality to non-intelligent systems, just cognitive functions. A system can be said to instantiate a function if its inputs and outputs are interpretable as proper for the function. In this way, we can instantiate the addition function:

> To get a physical system to add, you have to get it to satisfy a function interpretable as addition. And that means you have to design it so that getting it to represent a pair of addends causes it to represent their sum (Cummins, 1989, p. 90).

Thus, to say something "instantiates a function" is merely to say that it can be interpreted as having instantiated that function.[21] There are, of course, criteria upon which to judge the validity and reasonability of a particular interpretation. Cummins' metaphor is the Tower Bridge. Basically, the Tower Bridge has two spans and two supports. The bottom span is the physical system which moves from the first support, the input, to the second support, the output. The top span is the interpreted function, which travels from an interpretation of the inputs to an interpretation of the outputs. If the function "fits" the system for its inputs and outputs, it is a valid interpretation. Whether it is reasonable depends on many factors including the environmental context and purposes of the function.

This approach to psychological description does not differ much from that taken by computational psychologists. David Marr (1982) distinguishes three levels of description available for describing psychological phenomena. The highest level is the *computational theory*, which is the abstract computational function which governs the process. At the middle level are the *representations* and *algorithms*. This is the symbolic and logical architecture used to satisfy the function. The lowest level is the *hardware implementation* level, the electronic wires or neurons which actually carry out the computation. Cummins' point is simply that when

[21] For a thorough discussion of intentional and representational interpretation of function-instantiating systems see Haugeland (1981b).

investigating a process, we can never be certain that we have interpreted the system as having the representations which it actually has. There is nothing in the hardware level which determines our representational scheme, it can only constrain our interpretations. It is ultimately the computational theory which will determine the representations and algorithms. And since we can have multiple, equally plausible theories we cannot verify our representational analysis of a process.

This conservative analysis of representation results in what Cummins calls s-representation, where the "s" is for simulation. S-representation is representation according to an analytic or theoretical scheme; the representation we find depends on what we are looking for. An s-representation is in this way relative to the simulation under observation. This is a way of including our theory as part of our observation. A consequence of this is that a system which simulates or instantiates one function will also simulate many other functions (all those functions with an isomorphic range and domain), leaving open the possibility for many interpretations. The question is not which interpretations are correct but which are useful or interesting. And since we aren't trying to get at the subjective content, it no longer matters that we cannot describe or determine subjective content.

Cummins is quite clear about the limitations of his proposal:

> It seems to follow that the CTC, like mathematical science, will work only for the cognition of autonomously law-governed domains. The precondition for success is the same in both cases: There must be a well-defined upper span to the Tower Bridge. Special science isn't always possible. If cognition is possible where special science is not--in the cases of clothing and faces of conspecifics, for example--then the CTC's Tower Bridge picture of cognition can't be the whole story.
>
> If, as seems increasingly likely, the specification problem for cognition should prove to be intractable, or to be tractable only in special ways, where will that leave us?

> I think it will leave us with a kind of biological chauvinism (Block, 1978). Cognition will simply be identified ostensively, and hence extrinsically, as *what humans do when they solve problems, find their way home, etc* (Cummins, 1989, p. 113).

This means that psychology is limited to studying cognitive behavior only for certain specifiable processes. It may sound strict, but this is where cognitive psychology has proven successful anyway. Cognitive psychologists have made great strides in demonstrating how humans solve logic puzzles, perform "mental" rotations, decipher written language, and interpret visual fields. These domains are easy to specify, the solutions are clear and the mistakes can be clearly interpreted. But cognitive psychology cannot become a science of the mind, and it cannot describe subjective experience.

Whether empirical science will ever concern itself with questions of subjective experience is another matter, for it seems that we could investigate introspective data, or try to document subjective experiences. This is indeed what William James did and suggested that others do. In a paper entitled "Subjective Effects of Nitrous Oxide," James intimates his views towards a scientific discovery of subjective experience:

> Some observations of the effects of nitrous-oxide-gas-intoxication..have made me understand better than ever before both the strength and the weakness of Hegel's philosophy. I strongly urge others to repeat the experiment, which with pure gas is short and harmless enough. The effects will of course vary with the individual, just as they vary in the same individual from time to time; but it is probable that in the former case, as in the latter, a generic resemblance will obtain (James, 1882, p. 186).

Introspective investigation and personal experimentation may not be the next step empirical psychology wishes to take, however, as there is a strong phobia regarding introspective data in psychology. And perhaps such investigations are better left to philosophy and literature.

V. Connections vs. Symbols

Having provided the sufficient background to the issues involved, I return now to the debate raging between orthodox AI and connectionism. Goschke and Koppelberg point out that this debate rests largely on the debate between externalism and internalism over the nature of representation itself (Goschke & Koppelberg, 1991). The orthodox AI approach is one of symbolic representation wherein the symbols employed by a system are representational based on the relations they bear to other symbols in the system. The symbolic approach is thus an internalist approach to mental representation. The connectionist approach is to correlate patterns of activation, internal states, to external conditions. It is thus an externalist approach to representation.

The methodologies of each modelling approach should be sufficient to demonstrate their presupposed representational schemes. The representational approach taken in orthodox AI is clear from such AI programming languages as LISP and PROLOG. These languages are generally used to structure tokens or variables into large hierarchical, relational structures. Characteristics are ascribed to tokens, e.g. |dogs| have |fur|, and tokens are classified according to group membership, e.g. |dogs| are |mammals|. The idea is that specifying all of the classifications and characteristics of |dogs| will allow the system to give a strict definition of |dog| and the proper control program will allow the system to use this token appropriately. This knowledge-base approach in AI is clearly an application of conceptual-role semantics. As this approach does in no way depend upon environmental relations for its representational achievements, it is internalist to the core. The connectionists, on the other hand, take a radically different approach to representation. They avoid completely explicit rules or facts in the architecture of their models. Rather, they judge the representational character of the internal states of the model purely by means of performance. If the model responds correctly to the set of stimuli it is given, it is described as having properly encoded the necessary representations for its task. In a more obvious manner, a researcher might try to correlate the activity of a particular node with the characteristics of the input. In this way, the representationality of a node or the activation patterns of the system on the whole is not judged by internal consistency but with external correspondence. It is most clearly an applied version of externalism. Thus, the debate between connections and symbols is really a debate between the externalist and internalist representational schemes.

The question of which approach is right or better thus depends upon what we think about representation. I have shown in the previous sections how the internalist and externalist projects fail, each in their own right, to account for thought or intelligence with respect to meaning or content. The result of those arguments was the Interpretational Semantics of Cummins. This approach to semantics helps to clarify the debate, but does not solve it. In each approach one is applying a representational interpretation to the functions instantiated. The only difference between them lies in the functions: for the symbolic theorist the functions and terms are explicit and known where for the connectionist the functions are implicit and discovered.

The consequence of this is that neither approach will give us true Artificial Intelligence. What both approaches will give us are models of cognitive processes. The debate between them then is not which is correct, but which gives us the better model. This determination is ultimately dependent on what we intend to model. The symbolic approach is more useful if we are testing a theoretical cognitive function; we model that function and see if it works. The connectionist approach is more valuable if we are looking for the function. We design a system which instantiates a process and try to derive the function from its behavior; it is thus valuable in isolating processes and functions.

The larger question for AI in general is how to model general intelligence. While AI and connectionism have developed computer models of various cognitive processes that quite often perform better than humans, both of these approaches have reached intractable complexity problems when extrapolated to real-world situations (Dreyfus & Dreyfus, 1990). Dreyfus & Dreyfus provide an excellent brief history of this AI objective from a philosophical perspective. Symbolic AI assumes a world-view in which we can carry out Russell's logical atomism and thereby discover the "ultimate context-free, purpose-free elements" of thought (Dreyfus & Dreyfus, 1990). This is what Wittgenstein expounded in his *Tractus*, but finally concluded was impossible in his later *Philosophical Investigations*:

> It is one of the ironies of intellectual history that Wittgenstein's devastating attack on his own *Tractus*, his *Philosophical Investigations*, was published in 1953, just as AI took over the abstract, atomistic tradition he was attacking. After writing the *Tractus*, Wittgenstein spent years...looking in vain for the atomic facts and basic objects his theory required. He ended by abandoning his *Tractus* and all rationalistic philosophy. He argued that the analysis of everyday situations into facts and rules (which is where most traditional philosophers and AI researchers think theory must begin) is itself only meaningful in some context and for some purpose. Thus, the elements chosen already reflect the goals and purposes for which they are carved out (Dreyfus & Dreyfus, 1990, p. 320).

What Wittgenstein's arguments amount to is a metaphysical conclusion that human thought cannot be specified in mathematically precise enough ways to create a computational model capable of instantiating those functions. He foresaw the complexity problem intrinsic to AI and hints at the Interpretational Semantics of Cummins.

Connectionism also has its intractable complexity problems. While it is relatively simple to model a simple, well defined process, modelling general intelligence is much more difficult. Our intuitions tell us that if we built a large enough neural network and had enough time, we could train it to respond appropriately to every possible input. But just as it is impossible to write out the rules governing every intelligent behavior for the symbolic theorist, it is impossible to create the complete set of inputs that will guarantee the system will produce intelligent behaviors. These are not theoretical impossibilities, but practical impossibilities.[22]

Even if an artificial model did succeed in demonstrating general intelligence (e.g. passed a Turing test), it would not thereby have achieved intelligence. It would merely have succeeded in modelling intelligence, and modelling is indeed a valuable and interesting field of inquiry. I think the best conclusion that can be drawn from everything I have said here is that if we really want to understand the nature of intelligence, thought, and understanding, we must change our approach. Our theories of meaning, content and representation should not be motivated by a desire to justify or bolster our scientific paradigms. Rather, what is needed is a theory of mental content independently derived and defensible, which scientific theories and models strive to satisfy.

[22] The later has also proved computationally impossible. This is just to say that for any given neural network, training for general intelligence is NP-complete. See (Judd, 1990).

References

Barnden, J. A. (1992). Connectionism, generalization, and propositional attitudes: A catalog of challenging issues. In J. Dinsmore (Ed.), The symbolic and connectionist paradigms: Closing the gap (pp. 25-48). Hillsdale, NJ: Lawrence Erlbaum Associates.

Block, N. (1978). Troubles with functionalism. In C. W. Savage (Ed.), Minnesota Studies in the Philosophy of Science, Volume 9, Perception and cognition: Issues in the foundations of psychology. Minneapolis, MN: University of Minnesota Press.

Block, N. (1991). What narrow content is not. In B. Loewer & G. Rey (Eds.), Meaning in mind: Fodor and his critics (pp. 33-64). Oxford, UK: Basil Blackwell.

BonJour, L. (1991). Is thought a symbolic process? Synthese, 89, 331-352.

Bradshaw, D. E. (1991). Connectionism and the specter of representationalism. In T. Horgan & J. Tienson (Eds.), Connectionism and the philosophy of mind (pp. 417-436). Norwell, MA: Kluwer Academic Publishers.

Chalmers, D. J. (1992). Subsymbolic processing and the Chinese room. In J. Dinsmore (Ed.), The symbolic and connectionist paradigms: Closing the gap (pp. 25-48). Hillsdale, NJ: Lawrence Erlbaum Associates.

Churchland, P. M. (1984). Matter and consciousness: A contemporary introduction to the philosophy of mind. Cambridge, MA: MIT Press.

Cummins, R. (1989). Meaning and mental representation. Cambridge, MA: MIT Press.

Cummins, R. (1991). The role of representation in connectionist explanations of cognitive capacities. In W. Ramsey, S. Stich, & D. Rumelhart (Eds.), Philosophy and connectionist theory (pp. 91-114). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cummins, R. & Schwarz, G. (1991) Connectionism, computation, and cognition. In T. Horgan & J. Tienson (Eds.), Connectionism and the philosophy of mind (pp. 60-73). Norwell, MA: Kluwer Academic Publishers.

Cussins, A. (1990). The connectionist construction of concepts. In M. A. Boden (Ed.), The philosophy of Artificial Intelligence (pp. 368-440). Oxford, UK: Oxford University Press.

Dennett, D. (1981). Three kinds of intentional psychology. In R. A. Healy (Ed.), *Reduction, time and reality: Studies in the philosophy of the natural sciences*. Cambridge, UK: Cambridge University Press.

Devitt, M. (1991). Why Fodor can't have it both ways. In B. Loewer & G. Rey (Eds.), *Meaning in mind: Fodor and his critics* (pp. 95-118). Oxford, UK: Basil Blackwell.

Dretske, F. (1988). *Explaining behavior*. Cambridge, MA: MIT Press.

Dretske, F. (1981). *Knowledge and the flow of information*. Cambridge, MA: MIT Press.

Dreyfus, H. L. & Dreyfus, S. E. (1990). Making a mind versus modelling the brain: Artificial Intelligence back at a branch-point. In M. A. Boden (Ed.), *The philosophy of Artificial Intelligence* (pp. 309-333). Oxford, UK: Oxford University Press. (Reprinted from *Artificial Intelligence*, Winter 1988, *117*(1))

Fodor, J. (1981). Methodological solipsism considered as a research strategy in cognitive psychology. In J. Haugeland (Ed.), *Mind design* (pp. 307-338). Cambridge, MA: MIT Press. (Reprinted from *The Behavioral and Brain Sciences*, 1980, *3*, 63-73)

Fodor, J. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*. Cambridge, MA: MIT Press.

Fodor, J. & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A crtitical analysis. *Cognition*, *28*, 3-71.

Goschke, T. & Koppelberg, D. (1991). The concept of representation and the representation of concepts in connectionist models. In W. Ramsey, S. Stich, & D. Rumelhart (Eds.), *Philosophy and connectionist theory* (pp. 129-161). Hillsdale, NJ: Lawrence Erlbaum Associates.

Haugeland, J. (1981a). Semantic engines: An introdusction to mind design. In J. Haugeland (Ed.), *Mind design* (pp. 1-34).

Haugeland, J. (1981b). The nature and plausibility of cognitivism. In J. Haugeland (Ed.), *Mind design* (pp. 243-281). (Reprinted from *The Behavioral and Brain Sciences*, 1978, *1*, 215-226)

Haugeland, J. (1991). Representational genera. In W. Ramsey, S. Stich, & D. Rumelhart (Eds.), *Philosophy and connectionist theory* (pp. 61-89). Hillsdale, NJ: Lawrence Erlbaum Associates.

James, W. (1882). Subjective effects of nitrous oxide. *Mind*, *7*, 186-208.

Judd, J. S. (1990). Neural network design and the complexity of learning. Cambridge, MA: MIT Press.

Lloyd, D. (1991). Leaping to conclusions: Connectionism, consciousness, and the computational mind. In T. Horgan & J. Tienson (Eds.), Connectionism and the philosophy of mind (pp. 444-459). Norwell, MA: Kluwer Academic Publishers.

Loar, B. (1991). Can we explain intentionality? In B. Loewer & G. Rey (Eds.), Meaning in mind: Fodor and his critics (pp. 119-135). Oxford, UK: Basil Blackwell.

Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information. San Francisco, CA: W. H. Freeman and Company.

Newell, A. & Simon, H. (1981). Computer science as empirical inquiry: Symbols and search. In J. Haugeland (Ed.), Mind design (pp. 35-66). Cambridge, MA: MIT Press. (Reprinted from Communications of the Association of Computing Machinery, March 1976, 19, 113-126)

Pinker, S. & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. Cognition, 28, 73-193.

Putnam, H. (1975). The meaning of 'meaning.' In K. Gunderson (Ed.), Minnesota Studies in the Philosophy of Science, Volume 7, Language, Mind and Knowledge. Minneapolis, MN: University of Minnesota Press.

Quine, W. V. (1969). Ontological relativity and other essays. New York, NY: Columbia University Press.

Ramsey, W. (1992). Connectionism and the philosophy of mental representation. In S. Davies (Ed.), Connectionism: Theory and practice (pp. 247-276). Oxford, UK: Oxford University Press.

Schwartz, J. (1992). Who's afraid of multiple realizability?: Functionalism, reductionism, and connectionism. In J. Dinsmore (Ed.), The symbolic and connectionist paradigms: Closing the gap (pp. 25-48). Hillsdale, NJ: Lawrence Erlbaum Associates.

Searle, J. (1969). Speech Acts. Cambridge, UK: Cambridge University Press.

Searle, J. (1980). Minds, brains, and programs. The Behavioral and Brain Sciences, 3, 417-424.

Searle, J. (1983). Intentionality: An essay in the philosophy of mind. Cambridge, UK: Cambridge University Press.

Searle, J. (1987). Indeterminacy, empiricism, and the first person. Journal of Philosophy, 84, 123-146.

Searle, J. (1990). Consciousness, unconsciousness, and intentionality. In C. A. Anderson & J. Owens (Eds.), Propositional attitudes: The role of content in logic, language, and mind (pp. 269-284). Stanford, CA: Center for the Study of Language and Information.

Stalnaker, R. (1990). Narrow content. In C. A. Anderson & J. Owens (Eds.), Propositional attitudes: The role of content in logic, language, and mind. Stanford, CA: Center for the Study of Language and Information.

Stich, S. (1983). From folk psychology to cognitive science: The case against belief. Cambridge, MA: The MIT Press.