Projected Meaning, Grounded Meaning and Intrinsic Meaning

C. Franklin Boyle
CDEC
Carnegie Mellon University
Pittsburgh, PA 15213
fb0m@andrew.cmu.edu

Abstract

It is proposed that the fundamental difference between representations whose constituent symbols have intrinsic meaning (e.g. mental representations) and those whose symbols have meanings we consider "projected" (e.g. computational representations) is causal. More specifically, this distinction depends on differences in *how* physical change is brought about, or what we call "causal mechanisms". These mechanisms serve to physically ground our intuitive notions about syntax and semantics.

Introduction

One of the defining characteristics of mind is that the contents of mental states, i.e. concepts and conceptual relationships, are intrinsically referential; they refer to things in the world without requiring external agency to realize this capacity — their meanings are intrinsic. Moreover, it seems likely that this quality of mind is responsible for the kind of understanding we experience when reading or when listening to spoken language.

It has been argued that formal symbols like those instantiated in digital computers do not have intrinsic meaning; that is, formal symbol manipulation is not sufficient for semantics (Searle, 1980; 1990). Any meaning such symbols purportedly have is projected onto them by us (Harnad, 1990). Yet it is often pointed out that "at bottom" everything is just syntax or, as Haugeland (1989) cogently observes, "meanings do not exert mechanical forces". Unless we are willing to believe in the existence of some sort of non-physical "mindstuff", we either have to agree with the major tenets of this latter, functionalist view or accept the burden of proving that there is some fundamental physical difference between mental states and computational states that might explain differences in their referential capacities.

Of course, not all the potentially supporting evidence for functionalism has been gathered and may not be for quite some time; computer systems are far from being robust enough to match the functional complexity of mind and, therefore, to test the "multiple instantiations" hypothesis (Thagard, 1986) which asserts that it is only the causal *relationships* between mental states that are physically relevant to mind, not particular substrates or architectures. There are plenty of reasons for believing that this hypothesis will never be proved: the system complexity required to test it, the operationally ill-defined nature of Turing-type tests, the "other minds" problem, etc. If, however, such causal associations are the only physical requirement on which having a mind depends (whether or not this can actually be demonstrated), it would mean that semantics is very likely the product of a system's functioning, implying that computational systems could understand the way we do. Searle's counter to this position is an intuitive argument about the nature of our understanding and a claim that brains, not computers, have the right "causal property" to produce intentional states (Searle, 1980). Though he says nothing further about this causal property, we might infer that he does not believe it to be just the set of causal relationships between mental states, since that would render it indistinguishable from functionalism.

Instead of waiting for the set of causal relationships between mental states to be instantiated computationally (if, in fact, that is possible) or relying on intuitive arguments about understanding, our strategy here is to advance Searle's position by 1) introducing a set of *causal mechanisms* which determine the kinds of physical changes that can occur when physical objects interact, and 2) arguing that differences in these causal mechanisms give rise to different ways representing entities mean. Since Searle does not explain what "causal property" is, we take the liberty of equating it with causal mechanism and then identify different causal mechanisms as causal properties of different kinds of information processing systems. This provides us with a physically-principled basis for arguing that digital computers (indeed, all pattern-matching systems, as we will see) differ from brains in how their respective

representing entities mean. This latter difference strongly suggests that understanding, insofar as it depends on how representing entities mean, will *always* be different for computers and brains; that having a set of causal relations between states is not equivalent to having a particular causal mechanism that enables those relations, i.e. that transforms one state into the next. In other words, there are special physical processes that are the basis for mental representations having intrinsic meaning.

Three Types of Meaning

We begin by listing three ways representing entities can mean, that is, three ways they come to be about the things they purportedly represent:

> Projected Meaning — representing entities have meaning by virtue of our projecting it onto them. The association between representation and referent is arbitrary; that is, inputs are encoded by us or by procedures we construct.
>
> Grounded Meaning — representing entities have meaning by virtue of their being grounded in the analog projections of sensory stimuli -- the relationship between stimulus and internal representation is non-arbitrary (Harnad 1990).
>
> Intrinsic Meaning — representing entities have meaning by virtue of their being both grounded in analog projections *and* causal by the same kind of structure-preserving process that underlies their grounding (Boyle 1991).

We will argue that differences in these are due to differences in a specific aspect of physical object interactions that has so far been overlooked as being fundamental to our understanding of so-called information-processing systems — *how* physical objects are causal (Boyle, 1991), to be distinguished from *what* changes they cause. Our claim is that this is the only *principled* criterion for distinguishing the above ways representing entities can mean, not correlation or form similarity which are customarily used to reason about the nature of meaning in representational systems.

In what follows, we first discuss correlation and form and their shortcomings with respect to determining how symbols mean. We then investigate the causal aspects of representations and explain how the above types of meaning depend on what we identify as "causal mechanisms".

Correlation

As the designers of cognitive models, *we* determine what meanings the constituent symbols and symbolic expressions have simply by designating their referents. We then proceed to make these designations (which we store in our heads) consistent with the effects those symbols and symbolic expressions have on system behavior via procedures that associate the symbols and symbolic expressions with actions. Such actions are "grounded" in the interaction of system and environment — behavioral grounding — so that the meanings of the symbols behaviorally correlate with what we initially intended them to be about.

Clearly, the meanings of the symbols and symbolic expressions *before* we integrate them into a functioning system are *projected*, just as any object in the world could be interpreted as representing something else. But does making this pre-implementational, projected meaning consistent with system behavior change it from projected to intrinsic? In other words, is consistency based on correlation sufficient for semantics? According to functionalism it should be; if a computational system's behavior is indistinguishable from our own, then from the multiple-instantiations hypothesis, so are its "computational" states indistinguishable from our mental states to the extent that the relevant features are causal relations between mental states. Thus, its constituent symbols and symbolic expressions must mean in the same way the contents of our mental states mean, which we consider to be intrinsic. This sort of reasoning seems to be invoked in the so-called "systems reply" to Searle's Chinese Room argument — if the system is behaviorally indistinguishable from a native Chinese speaker, then it must understand the input (words) it processes in a manner similar to the way we understand language.

There are, however, two issues which suggest that this hypothesis about meaning in such systems, based as it is on correlation, is not empirically testable. The first is a practical one; because verification depends on behavior, if we fail to actually build such a system, we may be unable to determine if our failure was due to the omission of certain internal state relationships (e.g., state X causes state Y) or because computer systems lack some physical property that prevents us from successfully implementing all such relationships. The second issue is a reminder of the limitations inherent in making inferences about the nature of a system's internal characteristics based on its behavior. *Any* system whose internal representation of the world affects its behavior requires some (presumably high) degree of consistency between its symbols and their referents (what Haugeland (1989) sees as a strong constraint on the number of possible interpretations

of its symbols) if that system is to behave in a manner we would call rational. But since this should be the case for any coherent, representing system, whether brain or machine, it implies that behavior-based correlation is not adequate for distinguishing between systems with intrinsic meaning, like brains, and those whose meanings may only be projected, like computational models of cognition.

Functionalists might respond by pointing out that completeness is also necessary; that a system will have intrinsic meaning only when it is as behaviorally robust as the brain. But why would a more complete system, which differs from one that is less complete only in the number of state relationships it instantiates, necessarily be more consistent except to the extent that there is simply more of it to be consistent? The only plausible answer, one which avoids the implication that the symbols in *any* consistent program, no matter how simple, have intrinsic meaning, is that cognitive properties might emerge when system complexity (in terms of the number of rules, for instance) is increased beyond some threshold. But until we determine what might cause this sort of emergence, if indeed it could actually happen, we must depend on consistency.

Thus, it seems clear that using behavioral criteria to explain how representing entities mean leaves too many questions unanswered. Since we believe there is something quite specific that gives rise to intrinsic meaning and since we reject explanations based on any sort of non-physical Cartesian mind-stuff or emergence, the only alternative at this point is to look for *physical* differences that depend on the structural characteristics of representing entities, independent of the particular medium, and on how these physically affect system behavior. For example, what are the medium-independent physical differences between mental representations of trees and their computational counterparts?

There appear to be two kinds of physical differences. The first is associated with the similarity between a representing entity's form and that of its referent. The second is based on *how* symbols in various systems bring about change; *how* they are causal. As we noted above, with respect to meaning, the latter is fundamental.

Form

If the physical forms of symbols are unlike those of their referents, which is the case for formal symbols in computers, then how can they represent what it is they are purportedly about unless we say they do, that is, unless their meanings are projected? On the other hand, if the structures of symbol and referent are similar or nearly so, can we say the meanings of such symbols are intrinsic? In the Chinese Room, for example, it could be argued that understanding is very different from our own because there are no forms accompanying the structurally-arbitrary input symbols that are isomorphic to the forms of the referents of those symbols. *We* acquire this kind of "form information" visually and associate it with words in our language, presumably to understand them, so should not computers require the same to understand language? Perhaps, but using form similarity to determine whether the meaning of a representation is intrinsic or not, and, hence, whether a computer's understanding is like our own, is problematic for two reasons.

First, similarity between two shapes or structures is a matter of degree, whereas meaning is *either* intrinsic *or* it is not. Otherwise, we might end up with a representation in which the meanings of some symbols, or even parts of symbols, are intrinsic while others are not, implying that somehow specific objects in the world give rise to different types of meaning, or that the system understands different objects differently, both of which seem highly unlikely. A second problem with form similarity concerns the issue of what physically makes a particular representation similar in form to its referent, and to whom. At first glance, the answer to the first part seems obvious. After all, the bitmap of a tree is clearly similar in structure to its referent. However, this may only be a similarity *to us;* digital computers probably do not "see" it that way. For them it is just another pattern to be matched, no different than any other bitmap or arbitrary combination of symbols because it is only the presence of a matcher which "fits" the pattern that is relevant to the pattern's effect on system behavior, not its *particular* form, i.e., not its *appearance*. Hence, such structures might be characterized as "intrinsically meaningless" to digital computers because they are not causal according to appearance. This will become clearer after we introduce causality as the basis for distinguishing different types of meaning.

Form, therefore, is really a criterion for distinguishing different kinds of representations (at least for us) such as extrinsic (e.g. propositional) and intrinsic (e.g. iconic) representations (Palmer, 1978), not meanings. That is, form has to do with how a representation encodes what it represents rather than how it means.

Causal Criteria

Having argued that behavior-based symbol-referent correlation and form similarity are not adequate for distinguishing how representing entities mean, we now turn to causality, but causality considered in a

non-standard way. Typically, causality is expressed in terms of *why* something happens (cause) or *what* happens (effect). These are combined to form cause-and-effect pairs which associate a particular entity, a symbolic expression for example, with the effects it brings about — a highly functional characterization of physical change akin to "if-then" rules. The physical processes which actually produce the effects get buried in the structureless connection between the antecedent and consequent of such forms. In other words, there is no sense in which the associative link conveys *how* the cause actually brings about the effect, only that it does.

Here, however, we consider causality *deterministically*. That is, we determine *how* particular effects could be brought about when physical objects interact. The different ways effects are physically brought about we refer to as "causal mechanisms".

Three Causal Mechanisms

There are only three causal mechanisms for bringing about change in physical interactions: *nomologically-determined change, pattern matching* and *structure-preserving superposition* (Boyle, 1991). Each mechanism depends on a particular aspect of physical objects that is responsible for the resulting changes. These are *measured attributes, form* and *appearance*, respectively. Though physical objects have only two *physical* aspects — measured attributes and extended structure — we describe the latter as form or as appearance depending on whether the causal mechanism is pattern matching or structure-preserving superposition, respectively.[1]

1). *Nomologically-determined change* is the causal mechanism that underlies most physical interactions. Exemplified by what is customarily described in the literature as "billiard ball collisions", the effects of such interactions are determined according to nomological relationships between measured attributes (e.g. momentum) of the colliding objects. When two billiard balls collide, the outcome of the interaction is determined by the law of conservation of momentum along with constraint relationships that depend on structural aspects of the particular situation, such as the angle of closest approach. Thus, the changes that result from an interaction depend only on the *values of measured attributes* of the colliding objects, which determine the magnitude and direction of the forces that bring about those changes.

Informationally (i.e. if measured attributes are taken to represent), the changes are not indicative of the particular *objects* which interacted, only of the values of their measured attributes. Certainly initial measured-attribute values of the particular objects will cause specific value changes, but these are situation-specific, not object-specific — they do not identify particular objects — since a) there are, in essence, an infinite number of configurations for two objects to be in when they collide and, therefore, an infinite number of different values to describe them, and b) this is true for *any* two objects. Analog computers are exemplary of systems that utilize nomologically-determined interactions informationally; specific measured attributes of their component parts are taken to represent quantities in mathematical and physical models of different phenomena and the interactions of these parts are engineered to produce value changes which correspond to value changes of the represented quantities in the models.

2). *Pattern matching* is the *physical* process underlying many biomolecular interactions, such as enzyme catalysis, as well as computational changes in digital computers. Unlike nomologically-determined change, pattern matching physically depends on the *forms* of interacting objects because a successful pattern match can only occur if the pattern and matcher structurally "fit". The values of measured attributes of pattern and matcher are not relevant to the pattern matching process except insofar as they physically enable it to happen. That is, structure fitting involves forces like any other physical interaction. Indeed, if the interacting structures, such as a key and a door lock, do not fit, there is no set of measured-attribute values that could lead to an outcome which would have been produced if they had.

Thus, pattern matching depends on the structural forms of interacting objects. The actual change caused by this kind of interaction, however, is "simple" (Pattee, 1986) in that it does not embody or transmit structural features of the pattern, and, in fact, is generally structureless — e.g. the switching of a computer circuit voltage from "high" to "low" as the output of an electronic comparator. Because the pattern is matched *as a whole*, we say that its *form* causes the change. Informationally, the particular pattern is relevant only to the extent that there is a matcher which fits it. This is the reason symbols in formal symbol systems can have any form as long as they admit of a consistent interpretation.

3). Like pattern matching, the third causal mechanism, *structure-preserving superposition*, or

[1]Because of space limitations, these claims about the existence of only three causal mechanisms and two physical aspects of objects will have to remain unsubstantiated, though we do consider the latter to be self-evident. Objects also have functional and various relational aspects (e.g. part/whole), but these are not physical aspects.

SPS, depends on extended structure, but in a very different way. Whereas pattern matching is based on the existence of two structures which fit, that is, on the forms of *both* pattern and matcher, SPS actually causes a change that is the *transmission* of a pattern, like a stone imprinting its surface structure in a piece of soft clay, so that the *effect* is a structural formation of the specific features of the pattern's extended structure; that is, its *appearance* rather than form. Informationally, the structure of the input is transmitted to the system receiving it, in contrast to pattern-matching systems whose constituent matchers recognize input patterns, but do not transmit them. SPS is "automatic" in that, as a physical process, it can create new structures simply by physically superimposing structures.

To reiterate, the above three causal mechanisms are the only ways physical objects cause change; there are no other ways that one physical object can affect another except by one (or both) of its only two *physical* aspects: measured attributes and extended structure. Insofar as physical objects can be taken to represent, these causal mechanisms explain how what we tend to call information affects the behavior of information processing systems. Thus, they serve as a critical link between information and the physical world. But only in the cases of nomologically-determined change and SPS are the representing entities actually changed. In pattern matching, extended structure is used only to control nomological changes, such as voltage switching.

Causal Mechanisms and Types of Meaning

It was suggested above that form isomorphism between representation and referent is not sufficient for intrinsic meaning; just because a symbol looks *to us* like it represents does not mean it is not arbitrary to the system within which it is embedded. We are not talking here about the kind of arbitrariness that would result from our designating a tree bitmap to represent a cow, for example, but, rather, the arbitrariness that arises when the form of a symbol does not matter to the change it produces, which is the case for pattern matching systems.

In pattern matching systems (which include all artificial information processing systems except analog computers) the matcher and pattern physically fit, so that the forms of symbols and, hence, the structural similarity of symbol and referent, does not matter. *Any* form can be used to trigger a particular effect because form is used strictly for control. For example, the information about tree structure could be encoded as a bitmap (iconic representation) or a textual description (propositional representation), but in both cases matchers that fit the representing forms have to be present in order for that representation to be causal, i.e. for it to affect the system's functioning. The result of a match is a structureless change which *triggers* the next informationally-relevant physical change, such as the execution of a subroutine. Thus, for all pattern matching systems, which includes digital computers as well as current connectionist systems (Boyle, 1991), the meanings of representing entities are *projected* because the physical process of pattern matching eliminates any presumed functional significance from their forms, regardless of how they encode what they purportedly represent. Patterns in such systems seduce us into believing they are inherently meaningful because, in fact, they are to us. But they are not inherently meaningful to the systems precisely because matchers physically fit them, i.e. their appearances are not relevant to their functioning. Only *output* has no matcher, so *its* appearance *does* matter; that is, the appearance of the output determines the interaction of the system with its environment. But "inside" the system there are no such criteria for constraining structure. In effect, structure fitting renders the internal behavior "mindless".

In contrast to the "arbitrary" encodings of referent structures in strict pattern matching systems, arbitrary whether encoded by us or procedures we write for accepting input, *symbol grounding* involves what Harnad (1990) calls the "analog re-presentation" and "analog reduction" of sensory stimuli which generate perceptual category representations that are *not* arbitrary with respect to their referents. Based on his description, we take these analog processes to be examples of SPS. The resulting iconic structures that form perceptual categories are then associated with abstract symbols. The meanings of these symbols are not exactly projected because the relationship between symbol and referent is *physically* grounded in the sensory input; that is, SPS enables the extended structure of the input signal to directly create perceptual representations by transmitting them (in pattern matching systems, the input is not transmitted but *encoded* through a set of matchers).

However, if SPS is involved only in the *formation* of symbols, then we claim that their meaning is not intrinsic because subsequent to their formation, their extended structures are matched. Their grounding may be "fixed", but if they are not causal through SPS, then their appearances are no longer relevant and, therefore, meaningful to the system -- that is, they become causal through pattern matching, in which case they are like symbols in any pattern-matching system. In other words, SPS grounds the structural relationship between symbol and referent, but from then on the symbols behave as *formal* patterns. There is nothing about their particular structures that is necessary for the specific

changes they bring about. Only the presence of identically structured matchers is important. We could have done as much by encoding them ourselves because once the symbols become patterns to be matched, any groundedness they had is superfluous to their effects on system functioning. Nevertheless, we will refer to their meaning as grounded, which, in essence, is projected meaning with a non-arbitrary form relationship between symbol and referent.

According to the present thesis, only if these initially grounded symbols are subsequently causal through SPS would their meanings be intrinsic. To be meaningful to a system, they must cause changes which actually embody their structural features, not be "collapsed" into a formless outcome. Thus we believe SPS to be the physical basis for semantics and, hence, the causal mechanism underlying cognition, which is partly supported by evidence from sensory perception, in the form of retinotopic mappings on the primary visual cortex, for example. Furthermore, as Churchland (1989) notes, "there are many other cortical areas, less well understood as to exactly what they map, but whose topographical representation of distant structure is plain." Thus, it is SPS which we offer here as the fundamental difference between symbols in computers and brains; that the latter are semantic while the former are only syntactic. SPS, we believe, is Searle's hypothesized "causal property".

Summary and Conclusions

Intrinsic meaning is identified here with representations that have a causal capacity to effect physical change through structure-preserving superposition or SPS. Without this causal mechanism underlying physical change in a representing system, any meaning associated with the representation is projected onto it by us. Physical systems exhibiting this latter property are pattern-matching systems, such as digital computers (pattern matching is *their* underlying causal property). Grounded meaning is exhibited by symbols in systems which have a structure-preserving relationship between internal representations and their referents, but whose subsequent effects on system behavior are enabled through pattern matching — their meaning is projected, though they are rooted in a non-arbitrary form relationship with their referents.

In summary, we have tried to show that there is a plausible physical explanation for the apparent differences in meaning and understanding possessed by computers and brains that begins to forge a connection between our intuitions about mind and the physical world. It is based on a previously unexplored analysis of causality; *how* cause effects change. Its implications are that purely syntactic structures are those which are causal through pattern matching, while semantic structures are those which are causal through SPS and, hence, are those whose meanings are intrinsic. This further implies that pattern matching systems may never be able to instantiate the set of causal relationships between mental states and, therefore, may not be capable of simulating mind because the physical process of pattern matching is fundamentally different from SPS. We speculate that this shortcoming will likely manifest itself as a learning deficit.

References

Boyle, C. F. 1991. On the Physical Limitations of Pattern Matching. *Journal of Experimental and Theoretical Artificial Intelligence* 3:191-218.

Churchland, P.M. 1989. *A Neurocomputational Perspective*. Boston, Mass.: MIT Press.

Harnad, S. 1990. The Symbol Grounding Problem, *Physica D* 42:335-346.

Haugeland, J. 1989. Artificial Intelligence and the Western Mind. In J.R. Brink and C.R. Haden (eds), *The Computer and the Brain: Perspectives on Human and Artificial Intelligence*. Elsevier.

Palmer, S. E. 1978. Fundamental Aspects of Cognitive Representation. In E. Rosch and B. Lloyd (eds), *Cognition and Categorization*. Hillsdale, NJ: Lawrence Earlbaum.

Pattee, H.H., 1986. Universal Principles of Language and Measurement Functions. In J.L. Casti and A. Karlqvist (eds), *Complexity, Language and Life: Mathematical Approaches*. New York: Springer-Verlag

Searle, J. 1980. Minds, Brains and Programs, *Behavioral and Brain Sciences* 3:417-457

Searle, J. 1990. Is the Brain's Mind a Computer Program? *Scientific American* 262(1):26-31

Thagard, P. 1986. Parallel Computation and the Mind-Body Problem. *Cognitive Science* 10:301-318