Physica D 42 (1990) 335–346
North-Holland

THE SYMBOL GROUNDING PROBLEM

Stevan HARNAD
Department of Psychology, Princeton University, Princeton, NJ 08544, USA

There has been much discussion recently about the scope and limits of purely symbolic models of the mind and about the proper role of connectionism in cognitive modeling. This paper describes the "symbol grounding problem": How can the semantic interpretation of a formal symbol system be made *intrinsic* to the system, rather than just parasitic on the meanings in our heads? How can the meanings of the meaningless symbol tokens, manipulated solely on the basis of their (arbitrary) shapes, be grounded in anything but other meaningless symbols? The problem is analogous to trying to learn Chinese from a Chinese/Chinese dictionary alone. A candidate solution is sketched: Symbolic representations must be grounded bottom-up in nonsymbolic representations of two kinds: (1) *iconic representations*, which are analogs of the proximal sensory projections of distal objects and events, and (2) *categorical representations*, which are learned and innate feature detectors that pick out the invariant features of object and event categories from their sensory projections. Elementary symbols are the names of these object and event categories, assigned on the basis of their (nonsymbolic) categorical representations. Higher-order (3) *symbolic representations*, grounded in these elementary symbols, consist of symbol strings describing category membership relations (e.g. "An X is a Y that is Z").

Connectionism is one natural candidate for the mechanism that learns the invariant features underlying categorical representations, thereby connecting names to the proximal projections of the distal objects they stand for. In this way connectionism can be seen as a complementary component in a hybrid nonsymbolic/symbolic model of the mind, rather than a rival to purely symbolic modeling. Such a hybrid model would not have an autonomous symbolic "module," however; the symbolic functions would emerge as an intrinsically "dedicated" symbol system as a consequence of the bottom-up grounding of categories' names in their sensory representations. Symbol manipulation would be governed not just by the arbitrary shapes of the symbol tokens, but by the nonarbitrary shapes of the icons and category invariants in which they are grounded.

1. Modeling the mind

1.1. From behaviorism to cognitivism

For many years the only empirical approach in psychology was behaviorism, its only explanatory tools input/input and input/output associations (in the case of classical conditioning [42]) and the reward/punishment history that "shaped" behavior (in the case of operant conditioning [1]). In a reaction against the subjectivity of armchair introspectionism, behaviorism had declared that it was just as illicit to theorize about what went on in the *head* of the organism to generate its behavior as to theorize about what went on in its *mind*. Only *observables* were to be the subject matter of psychology; and, apparently, these were expected to explain themselves.

Psychology became more like an empirical science when, with the gradual advent of cognitivism [17, 25, 29], it became acceptable to make inferences about the *unobservable* processes underlying behavior. Unfortunately, cognitivism let mentalism in again by the back door too, for the hypothetical internal processes came embellished with subjective interpretations. In fact, semantic interpretability (meaningfulness), as we shall see, was one of the defining features of the most prominent contender vying to become the theoretical vocabulary of cognitivism, the "language of thought" [6], which became the prevailing view in cognitive theory for several decades in the form of the

0167-2789/90/$03.50

"symbolic" model of the mind: The mind is a symbol system and cognition is symbol manipulation. The possibility of generating complex behavior through symbol manipulation was empirically demonstrated by successes in the field of artificial intelligence (AI).

1.2. Symbol systems

What is a symbol system? From Newell [28], Pylyshyn [33], Fodor [6] and the classical work by von Neumann, Turing, Gödel, Church, etc. (see ref. [18]) on the foundations of computation, we can reconstruct the following definition:

A symbol system is:

(1) a set of arbitrary *physical tokens* (scratches on paper, holes on a tape, events in a digital computer, etc.) that are

(2) manipulated on the basis of *explicit rules* that are

(3) likewise physical tokens and *strings* of tokens. The rule-governed symbol-token manipulation is based

(4) purely on the *shape* of the symbol tokens (not their "meaning"), i.e. it is purely *syntactic*, and consists of

(5) *rulefully combining* and recombining symbol tokens. There are

(6) primitive *atomic* symbol tokens and

(7) *composite* symbol-token strings. The entire system and all its parts – the atomic tokens, the composite tokens, the syntactic manipulations (both actual and possible) and the rules – are all

(8) *semantically interpretable*: The syntax can be *systematically* assigned a meaning (e.g. as standing for objects, as describing states of affairs).

According to proponents of the symbolic model of mind such as Fodor [7] and Pylyshyn [32, 33], symbol strings of this sort capture what mental phenomena such as thoughts and beliefs are. Symbolists emphasize that the symbolic level (for them, the mental level) is a natural functional level of its own, with ruleful regularities that are independent of their specific physical realizations. For symbolists, this implementation independence is the critical difference between cognitive phenomena and ordinary physical phenomena and their respective explanations. This concept of an autonomous symbolic level also conforms to general foundational principles in the theory of computation and applies to all the work being done in symbolic AI, the field of research that has so far been the most successful in generating (hence explaining) intelligent behavior.

All eight of the properties listed above seem to be critical to this definition of symbolic[#1]. Many phenomena have some of the properties, but that does not entail that they are symbolic in this explicit, technical sense. It is not enough, for example, for a phenomenon to be *interpretable* as rule-governed, for just about anything can be interpreted as rule-governed. A thermostat may be interpreted as following the rule: Turn on the furnace if the temperature goes below 70°F and turn it off if it goes above 70°F, yet nowhere in the thermostat is that rule explicitly represented. Wittgenstein [45] emphasized the difference between *explicit* and *implicit* rules: It is not the same thing to "follow" a rule (explicitly) and merely to behave "in accordance with" a rule (implicitly)[#2]. The critical difference is in the compositness (7) and systematicity (8) criteria. The explicitly represented symbolic rule is part of a formal system, it is decomposable (unless primitive), its application and manipulation is purely formal (syntactic, shape dependent), and the entire system must be semantically interpretable,

[#1] Paul Kube (personal communication) has suggested that (2) and (3) may be too strong, excluding some kinds of Turing machine and perhaps even leading to an infinite regress on levels of explicitness and systematicity.

[#2] Similar considerations apply to Chomsky's [2] concept of "psychological reality" (i.e. whether Chomskian rules are really physically represented in the brain or whether they merely "fit" our performance regularities, without being what actually governs them). Another version of the distinction concerns explicitly represented rules versus hard-wired physical constraints [40]. In each case, an explicit representation consisting of elements that can be recombined in systematic ways would be symbolic whereas an implicit physical constraint would not, although *both* would be semantically "interpretable" as a "rule" if construed in isolation rather than as part of a system.

not just the chunk in question. An isolated ("modular") chunk cannot be symbolic; being symbolic is a systematic property.

So the mere fact that a behavior is "interpretable" as ruleful does not mean that it is really governed by a symbolic rule[#3]. Semantic interpretability must be coupled with explicit representation (2), syntactic manipulability (4), and systematicity (8) in order to be symbolic. None of these criteria is arbitrary, and, as far as I can tell, if you weaken them, you lose the grip on what looks like a natural category and you sever the links with the formal theory of computation, leaving a sense of "symbolic" that is merely unexplicated metaphor (and probably differs from speaker to speaker). Hence it is only this formal sense of "symbolic" and "symbol system" that will be considered in this discussion of the grounding of symbol systems.

1.3. Connectionist systems

An early rival to the symbolic model of mind appeared [36], was overcome by symbolic AI [27] and has recently re-appeared in a stronger form that is currently vying with AI to be the general theory of cognition and behavior [23, 39]. Variously described as "neural networks", "parallel distributed processing" and "connectionism", this approach has a multiple agenda, which includes providing a theory of brain function. Now, much can be said for and against studying behavioral and brain function independently, but in this paper it will be assumed that, first and foremost, a cognitive theory must stand on its own merits, which depend on how well it explains our observable behavioral capacity. Whether or not it does so in a sufficiently brainlike way is another matter, and a downstream one, in the course of theory development. Very little is known of the brain's structure and its "lower" (vegetative) functions so far; and the nature of "higher" brain function is itself a theoretical matter. To "constrain" a cognitive theory to account for behavior in a brainlike way is hence premature in two respects: (1) It is far from clear yet what "brainlike" means, and (2) we are far from having accounted for a lifesize chunk of behavior yet, even without added constraints. Moreover, the formal principles underlying connectionism seem to be based on the associative and statistical structure of the causal interactions in certain dynamical systems; a neural network is merely one possible implementation of such a dynamical system[#4].

Connectionism will accordingly only be considered here as a cognitive theory. As such, it has lately challenged the symbolic approach to modeling the mind. According to connectionism, cognition is not symbol manipulation but dynamic patterns of activity in a multilayered network of nodes or units with weighted positive and negative interconnections. The patterns change according to internal network constraints governing how the activations and connection strengths are adjusted on the basis of new inputs (e.g. the generalized "delta rule", or "backpropagation" [23]). The result is a system that learns, recognizes patterns, solves problems, and can even exhibit motor skills.

1.4. Scope and limits of symbols and nets

It is far from clear what the actual capabilities and limitations of either symbolic AI or connectionism are. The former seems better at formal and language-like tasks, the latter at sensory, motor and learning tasks, but there is considerable overlap and neither has gone much beyond the stage of "toy" tasks toward lifesize behavioral capacity. Moreover, there has been some disagree-

[#3]Analogously, the mere fact that a behavior is *interpretable* as purposeful or conscious or meaningful does not mean that it really is purposeful or conscious. (For arguments to the contrary, see ref. [5].)

[#4]It is not even clear yet that a "neural network" needs to be implemented as a net (i.e. a parallel system of interconnected units) in order to do what it can do; if symbolic simulations of nets have the same functional capacity as real nets, then a connectionist model is just a special kind of symbolic model, and connectionism is just a special family of symbolic algorithms.

ment as to whether or not connectionism itself is symbolic. We will adopt the position here that it is not, because connectionist networks fail to meet several of the criteria for being symbol systems, as Fodor and Pylyshyn [10] have argued recently. In particular, although, like everything else, their behavior and internal states can be given isolated semantic interpretations, nets fail to meet the compositeness (7) and systematicity (8) criteria listed earlier: The patterns of interconnections do not decompose, combine and recombine according to a formal syntax that can be given a systematic semantic interpretation[#5]. Instead, nets seem to do what they do *non*-symbolically. According to Fodor and Pylyshyn, this is a severe limitation, because many of our behavioral capacities appear to be symbolic, and hence the most natural hypothesis about the underlying cognitive processes that generate them would be that they too must be symbolic. Our linguistic capacities are the primary examples here, but many of the other skills we have – logical reasoning, mathematics, chess playing, perhaps even our higher-level perceptual and motor skills – also seem to be symbolic. In any case, when we interpret our sentences, mathematical formulas, and chess moves (and perhaps some of our perceptual judgments and motor strategies) as having a systematic *meaning* or *content*, we know at first hand that this is literally true, and not just a figure of speech. Connectionism hence seems to be at a disadvantage in attempting to model these cognitive capacities.

Yet it is not clear whether connectionism should for this reason aspire to be symbolic, for the symbolic approach turns out to suffer from a severe handicap, one that may be responsible for the limited extent of its success to date (especially in modeling human-scale capacities) as well as the uninteresting and ad hoc nature of the symbolic "knowledge" it attributes to the "mind" of the symbol system. The handicap has been noticed in various forms since the advent of computing; I have dubbed a recent manifestation of it the "symbol grounding problem" [14].

[#5] There is some misunderstanding of this point because it is often conflated with a mere implementational issue: Connectionist networks can be simulated using symbol systems, and symbol systems can be implemented using a connectionist architecture, but that is independent of the question of what each can do *qua* symbol system or connectionist network, respectively. By way of analogy, silicon can be used to build a computer, and a computer can simulate the properties of silicon, but the functional properties of silicon are not those of computation, and the functional properties of computation are not those of silicon.

2. The symbol grounding problem

2.1. The Chinese room

Before defining the symbol grounding problem I will give two examples of it. The first comes from Searle's [37] celebrated "Chinese room argument", in which the symbol grounding problem is referred to as the problem of intrinsic meaning (or "intentionality"): Searle challenges the core assumption of symbolic AI that a symbol system capable of generating behavior indistinguishable from that of a person must have a mind. More specifically, according to the symbolic theory of mind, if a computer could pass the Turing test [43] in Chinese – i.e. if it could respond to all Chinese symbol strings it receives as input with Chinese symbol strings that are indistinguishable from the replies a real Chinese speaker would make (even if we keep testing for a lifetime) – then the computer would understand the meaning of Chinese symbols in the same sense that I understand the meaning of English symbols.

Searle's simple demonstration that this cannot be so consists of imagining himself doing everything the computer does – receiving the Chinese input symbols, manipulating them purely on the basis of their shape (in accordance with (1) to (8) above), and finally returning the Chinese output symbols. It is evident that Searle (who knows no Chinese) would not be understanding Chinese under those conditions – hence neither could the computer. The symbols and the symbol manipulation, being all based on shape rather than mean-

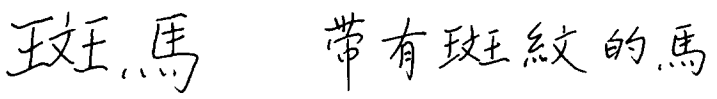

Fig. 1. Chinese dictionary entry. For translation, see footnote 17.

ing, are systematically *interpretable* as having meaning – that, after all, is what it is to be a symbol system, according to our definition. But the interpretation will not be *intrinsic* to the symbol system itself: It will be parasitic on the fact that the symbols have meaning for *us*, in exactly the same way that the meanings of the symbols in a book are not intrinsic, but derive from the meanings in our heads. Hence, if the meanings of symbols in a symbol system are extrinsic, rather than intrinsic like the meanings in our heads, then they are not a viable model for the meanings in our heads: Cognition cannot be just symbol manipulation.

2.2. *The Chinese / Chinese dictionary-go-round*

My own example of the symbol grounding problem has two versions, one difficult, and one, I think, impossible. The difficult version is: Suppose you had to learn Chinese as a second language and the only source of information you had was a Chinese/Chinese dictionary. The trip through the dictionary would amount to a merry-go-round, passing endlessly from one meaningless symbol or symbol-string (the definiens) to another (the definiendum), never coming to a halt on what anything meant[#6]. (See fig. 1.)

The only reason cryptologists of ancient languages and secret codes seem to be able to successfully accomplish something very like this is that their efforts are *grounded* in a first language and in real world experience and knowledge[#7]. The second variant of the dictionary-go-round, however, goes far beyond the conceivable resources of cryptology: Suppose you had to learn Chinese as a *first* language and the only source of information you had was a Chinese/Chinese dict-

[#6]Symbolic AI abounds with symptoms of the symbol grounding problem. One well-known (though misdiagnosed) manifestation of it is the so-called "frame" problem [22, 24, 26, 34]: It is a frustrating but familiar experience in writing "knowledge-based" programs that a system apparently behaving perfectly intelligently for a while can be foiled by an unexpected case that demonstrates its utter stupidity: A "scene-understanding" program will blithely describe the goings-on in a visual scene and answer questions demonstrating its comprehension (who did what, where, why?) and then suddenly reveal that it does not "know" that hanging up the phone and leaving the room does not make the phone disappear, or something like that. (It is important to note that these are not the kinds of lapses and gaps in knowledge that people are prone to; rather, they are such howlers as to cast serious doubt on whether the system has anything like "knowledge" at all.) The "frame" problem has been optimistically defined as the problem of formally specifying ("framing") what varies and what stays constant in a particular "knowledge domain," but in reality it is the problem of second-guessing all the contingencies the programmer has not anticipated in symbolizing the knowledge he is attempting to symbolize. These contingencies are probably unbounded, for practical purposes, because purely symbolic "knowledge" is ungrounded. Merely adding on more symbolic contingencies is like taking a few more turns in the Chinese/Chinese dictionary-go-round. There is in reality no ground in sight: merely enough "intelligent" symbol manipulation to lull the programmer into losing sight of the fact that its meaningfulness is just parasitic on the meanings he is projecting onto it from the grounded meanings in his own head. (I have called this effect the "hermeneutic hall of mirrors" [16]; it is the reverse side of the symbol grounding problem.) Yet parasitism it is, as the next "frame problem" lurking around the corner is ready to confirm. (A similar form of over-interpretation has occurred in the ape "language" experiments [41]. Perhaps both apes and computers should be trained using Chinese code, to immunize their experimenters and programmers against spurious over-interpretations. But since the actual behavioral tasks in both domains are still so trivial, there is probably no way to prevent their being decrypted. In fact, there seems to be an irresistible tendency to overinterpret toy task performance itself, preemptively extrapolating and "scaling it up" conceptually to lifesize without any justification in practice.)

[#7]Cryptologists also use statistical information about word frequencies, inferences about what an ancient culture or an enemy government are likely to be writing about, decryption algorithms, etc.

ionary[#8]! This is more like the actual task faced by a purely symbolic model of the mind: How can you ever get off the symbol/symbol merry-go-round? How is symbol meaning to be grounded in something other than just more meaningless symbols[#9]? This is the symbol grounding problem[#10]

2.3. *Connecting to the world*

The standard reply of the symbolist (e.g. Fodor [7, 8]) is that the meaning of the symbols comes from connecting the symbol system to the world "in the right way". But it seems apparent that the problem of connecting up with the world in the right way is virtually coextensive with the problem of cognition itself. If each definiens in a Chinese/Chinese dictionary were somehow connected to the world in the right way, we would hardly need the definienda! Many symbolists believe that cognition, being symbol manipulation, is an autonomous functional module that need only be hooked up to peripheral devices in order to "see" the world of objects to which its symbols refer (or, rather, to which they can be systematically interpreted as referring)[#11]. Unfortunately, this radically underestimates the difficulty of picking out the objects, events and states of affairs in the world that symbols refer to, i.e. it trivializes the symbol grounding problem.

It is one possible candidate for a solution to this problem, confronted directly, that will now be sketched: What will be proposed is a hybrid nonsymbolic/symbolic system, a "dedicated" one, in which the elementary symbols are grounded in two kinds of nonsymbolic representations that pick out, from their proximal sensory projections, the distal object categories to which the elementary symbols refer. Most of the components of which the model is made up (analog projections and transformations, discretization, invariance detection, connectionism, symbol manipulation) have also been proposed in various configurations by others, but they will be put together in a specific bottom-up way here that has not, to my knowledge, been previously suggested, and it is on this specific configuration that the potential success of the grounding scheme critically depends.

Table 1 summarizes the relative strengths and weaknesses of connectionism and symbolism, the two current rival candidates for explaining *all* of cognition single-handedly. Their respective strengths will be put to cooperative rather than competing use in our hybrid model, thereby also remedying some of their respective weaknesses. Let us now look more closely at the behavioral capacities such a cognitive model must generate.

3. Human behavioral capacity

Since the advent of cognitivism, psychologists have continued to gather behavioral data, although to a large extent the relevant evidence is already in: We already know what human beings

[#8] There is of course no need to restrict the symbolic resources to a dictionary; the task would be just as impossible if one had access to the entire body of Chinese-language literature, including all of its computer programs and anything else that can be codified in symbols.

[#9] Even mathematicians, whether Platonist or formalist, point out that symbol manipulation (computation) itself cannot capture the notion of the intended interpretation of the symbols [31]. The fact that formal symbol systems and their interpretations are not the same thing is hence evident independently of the Church–Turing thesis [18] or the Gödel results [3, 4], which have been zealously misapplied to the problem of mind-modeling (e.g. by Lucas [21]) – to which they are largely irrelevant, in my view.

[#10] Note that, strictly speaking, symbol grounding is a problem only for cognitive modeling, not for AI in general. If symbol systems alone succeed in generating all the intelligent machine performance pure AI is interested in – e.g. an automated dictionary – then there is no reason whatsoever to demand that their symbols have intrinsic meaning. On the other hand, the fact that our own symbols do have intrinsic meaning whereas the computer's do not, and the fact that we can do things that the computer so far cannot, may be indications that even in AI there are performance gains to be made (especially in robotics and machine vision) from endeavouring to ground symbol systems.

[#11] The homuncular viewpoint inherent in this belief is quite apparent, as is the effect of the "hermeneutic hall of mirrors" [16].

Table 1
Connectionism versus symbol systems.

Strengths of connectionism
(1) *Nonsymbolic function*: As long as it does not aspire to be a symbol system, a connectionist network has the advantage of not being subject to the symbol grounding problem.
(2) *Generality*: connectionism applies the same small family of algorithms to many problems, whereas symbolism, being a methodology rather than an algorithm, relies on endless problem-specific symbolic rules.
(3) "*Neurosimilitude*": Connectionist architecture seems more brain-like than a Turing machine or a digital computer.
(4) *Pattern learning*: Connectionist networks are especially suited to the learning of patterns from data.
Weaknesses of connectionism
(1) *Nonsymbolic function*: Connectionist networks, because they are not symbol systems, do not have the systematic semantic properties that many cognitive phenomena appear to have.
(2) *Generality*: Not every problem amounts to pattern learning. Some cognitive tasks may call for problem-specific rules, symbol manipulation, and standard computation.
(3) "*Neurosimilitude*": Connectionism's brain-likeness may be superficial and may (like toy models) camouflage deeper performance limitations.
Strengths of symbol systems
(1) *Symbolic function*: Symbols have the computing power of Turing machines and the systematic properties of a formal syntax that is semantically interpretable.
(2) *Generality*: All computable functions (including all cognitive functions) are equivalent to a computational state in a Turing machine.
(3) *Practical successes*: Symbol systems' ability to generate intelligent behavior is demonstrated by the successes of Artificial Intelligence.
Weaknesses of symbol systems
(1) *Symbolic function*: Symbol systems are subject to the symbol grounding problem.
(2) *Generality*: Turing power is too general. The solutions to AI's many toy problems do not give rise to common principles of cognition but to a vast variety of ad hoc symbolic strategies.

are able to do. They can (1) *discriminate*, (2) *manipulate*[#12], (3) *identify* and (4) *describe* the objects, events and states of affairs in the world they live in, and they can also (5) *produce descriptions* and (6) *respond to descriptions* of those objects, events and states of affairs. Cognitive theory's burden is now to explain *how* human beings (or any other devices) do all this[#13].

3.1. Discrimination and identification

Let us first look more closely at discrimination and identification. To be able to *discriminate* is to be able to judge whether two inputs are the same or different, and, if different, *how* different they are. Discrimination is a relative judgment, based on our capacity to tell things apart and discern their degree of similarity. To be able to *identify* is to be able to assign a unique (usually arbitrary) response – a "name" – to a class of inputs, treating them all as equivalent or invariant in some respect. Identification is an absolute judgment, based on our capacity to tell whether or not a given input is a member of a particular *category*.

Consider the symbol "horse". We are able, in viewing different horses (or the same horse in different positions, or at different times) to tell

[#12]Although they are no doubt as important as perceptual skills, motor skills will not be explicitly considered here. It is assumed that the relevant features of the sensory story (e.g. iconicity) will generalize to the motor story (e.g. in motor analogs [20]). In addition, large parts of the motor story may not be cognitive, drawing instead upon innate motor patterns and sensorimotor feedback. Gibson's [11] concept of "affordances" – the invariant stimulus features that are detected by the motor possibilities they "afford" – is relevant here too, though Gibson underestimates the processing problems involved in finding such invariants [44]. In any case, motor and sensory-motor grounding will no doubt be as important as the sensory grounding that is being focused on here.

[#13]If a candidate model were to exhibit all these behavioral capacities, both *linguistic* (5)–(6) and *robotic* (i.e. sensorimotor) (1)–(3), it would pass the "total Turing test" [15]. The standard Turing test [43] calls for linguistic performance capacity only: symbols in and symbols out. This makes it equivocal about the status, scope and limits of pure symbol manipulation, and hence subject to the symbol grounding problem. A model that could pass the total Turing test, however, would be grounded in the world.

them apart and to judge which of them are more alike, and even how alike they are. This is discrimination. In addition, in viewing a horse, we can reliably call it a horse, rather than, say, a mule or a donkey (or a giraffe, or a stone). This is identification. What sort of internal representation would be needed in order to generate these two kinds of performance?

3.2. *Iconic and categorical representations*

According to the model being proposed here, our ability to discriminate inputs depends on our forming *iconic representations* of them [14]. These are internal analog transforms of the projections of distal objects on our sensory surfaces [38]. In the case of horses (and vision), they would be analogs of the many shapes that horses cast on our retinas[#14]. Same/different judgments would be based on the sameness or difference of these iconic representations, and similarity judgments would be based on their degree of congruity. No homunculus is involved here; simply a process of superimposing icons and registering their degree of disparity. Nor are there memory problems, since the inputs are either simultaneously present or available in rapid enough succession to draw upon their persisting sensory icons.

So we need horse icons to discriminate horses, but what about identifying them? Discrimination is independent of identification. I could be discriminating things without knowing what they were. Will the icon allow me to identify horses? Although there are theorists who believe it would ([30]), I have tried to show why it could not [12, 14]. In a world where there were bold, easily detected natural discontinuities between all the categories we would ever have to (or choose to) sort and identify – a world in which the members of one category could not be confused with the members of any another category – icons might be sufficient for identification. But in our underdetermined world, with its infinity of confusable potential categories, icons are useless for identification because there are too many of them and because they blend continuously[#15] into one another, making it an independent problem to *identify* which of them are icons of members of the category and which are not! Icons of sensory projections are too unselective. For identification, icons must be selectively reduced to those *invariant features* of the sensory projection that will reliably distinguish a member of a category from any nonmembers with which it could be confused. Let us call the output of this category-specific feature detector the *categorical representation*. In some cases these representations may be innate, but since evolution could hardly anticipate all of the categories we may ever need or choose to identify, most of these features must be learned from experience. In particular, our categorical representation of a horse is probably a learned one. (I will defer till section 4 the problem of how the invariant features underlying identification might be learned.)

Note that both iconic and categorical representations are nonsymbolic. The former are analog copies of the sensory projection, preserving its "shape" faithfully; the latter are icons that have been selectively filtered to preserve only some of the features of the shape of the sensory projection: those that reliably distinguish members from nonmembers of a category. But both representations are still sensory and nonsymbolic. There is no

[#14] There are many problems having to do with figure/ground discrimination, smoothing, size constancy, shape constancy, stereopsis, etc., that make the problem of discrimination much more complicated than what is described here, but these do not change the basic fact that iconic representations are a natural candidate substrate for our capacity to discriminate.

[#15] Elsewhere [13, 14] I have tried to show how the phenomenon of "categorical perception" could generate internal discontinuities where there is external continuity. There is evidence that our perceptual system is able to segment a continuum, such as the color spectrum, into relatively discrete, bounded regions or categories. Physical differences of equal magnitude are more discriminable across the boundaries between these categories than within them. This boundary effect, both innate and learned, may play an important role in the representation of the elementary perceptual categories out of which the higher-order ones are built.

problem about their connection to the objects they pick out: It is a purely causal connection, based on the relation between distal objects, proximal sensory projections and the acquired internal changes that result from a history of behavioral interactions with them. Nor is there any problem of semantic interpretation, or of whether the semantic interpretation is justified. Iconic representations no more "mean" the objects of which they are the projections than the image in a camera does. Both icons and camera images can of course be *interpreted* as meaning or standing for something, but the interpretation would clearly be derivative rather than intrinsic[#16].

3.3. Symbolic representations

Nor can categorical representations yet be interpreted as "meaning" anything. It is true that they pick out the class of objects they "name", but the names do not have all the systematic properties of symbols and symbol systems described earlier. They are just an inert taxonomy. For systematicity it must be possible to combine and recombine them rulefully into propositions that can be semantically interpreted. "Horse" is so far just an arbitrary response that is reliably made in the presence of a certain category of objects. There is no justification for interpreting it holophrastically as meaning "This is a [member of the category] horse" when produced in the presence of a horse, because the other expected systematic properties of "this" and "a" and the all-important "is" of predication are not exhibited by mere passive taxonomizing. What would be required to generate these other systematic properties? Merely that the grounded names in the category taxonomy be strung together into *propositions* about further category membership relations. For example:

(1) Suppose the name "horse" is grounded by iconic and categorical representations, learned from experience, that reliably discriminate and identify horses on the basis of their sensory projections.

(2) Suppose "stripes" is similarly grounded.

Now consider that the following category can be constituted out of these elementary categories by a symbolic description of category membership alone:

(3) "Zebra" = "horse" & "stripes"[#17]

What is the representation of zebra? It is just the symbol string "horse & stripes". But because "horse" and "stripes" are grounded in their respective iconic and categorical representations, "zebra" inherits the grounding, through its grounded *symbolic* representation. In principle, someone who had never seen a zebra (but had seen and learned to identify horses and stripes) could identify a zebra on first acquaintance armed with this symbolic representation alone (plus the nonsymbolic – iconic and categorical – representations of horses and stripes that ground it).

Once one has the grounded set of elementary symbols provided by a taxonomy of names (and the iconic and categorical representations that give content to the names and allow them to pick out the objects they identify), the rest of the symbol strings of a natural language can be generated by symbol composition alone[#18], and they will all inherit the intrinsic grounding of the elementary

[#16]On the other hand, the resemblance on which discrimination performance is based – the degree of isomorphism between the icon and the sensory projection, and between the sensory projection and the distal object – seems to be intrinsic, rather than just a matter of interpretation. The resemblance can be objectively characterized as the degree of invertibility of the physical transformation from object to icon [14].

[#17]Fig. 1 is actually the Chinese dictionary entry for "zebra", which is "striped horse". Note that the character for "zebra" actually happens to be the character for "horse" plus the character for "striped." Although Chinese characters are iconic in structure, they function just like arbitrary alphabetic lexigrams at the level of syntax and semantics.

[#18]Some standard logical connectives and quantifiers are needed too, such as not, and, all, etc. (though even some of these may be learned as higher-order categories).

set[#19]. Hence, the ability to discriminate and categorize (and its underlying nonsymbolic representations) have led naturally to the ability to describe and to produce and respond to descriptions through symbolic representations.

4. A complementary role for connectionism

The symbol grounding scheme just described has one prominent gap: No mechanism has been suggested to explain how the all-important categorical representations could be formed: How does the hybrid system find the invariant features of the sensory projection that make it possible to categorize and identify objects correctly[#20]?

Connectionism, with its general pattern learning capability, seems to be one natural candidate (though there may well be others): Icons, paired with feedback indicating their names, could be processed by a connectionist network that learns to identify icons correctly from the sample of confusable alternatives it has encountered by dynamically adjusting the weights of the features and feature combinations that are reliably associated with the names in a way that (provisionally) resolves the confusion, thereby reducing the icons to the *invariant* (confusion-resolving) features of the category to which they are assigned. In effect, the "connection" between the names and the objects that give rise to their sensory projections and their icons would be provided by connectionist networks.

This circumscribed complementary role for connectionism in a hybrid system seems to remedy the weaknesses of the two current competitors in their respective attempts to model the mind singlehandedly. In a pure symbolic model the crucial connection between the symbols and their referents is missing; an autonomous symbol system, though amenable to a systematic semantic interpretation, is ungrounded. In a pure connectionist model, names are connected to objects through invariant patterns in their sensory projections, learned through exposure and feedback, but the crucial compositional property is missing; a network of names, though grounded, is not yet amenable to a full systematic semantic interpreta-

[#19]Note that it is not being claimed that "horse", "stripes", etc. are actually elementary symbols, with direct sensory grounding; the claim is only that *some* set of symbols must be directly grounded. Most sensory category representations are no doubt hybrid sensory/symbolic; and their features can change by bootstrapping: "Horse" can always be revised, both sensorily and symbolically, even if it was previously elementary. Kripke [19] gives a good example of how "gold" might be baptized on the shiny yellow metal in question, used for trade, decoration and discourse, and then we might discover "fool's gold", which would make all the sensory features we had used until then inadequate, forcing us to find new ones. He points out that it is even possible in principle for "gold" to have been inadvertently baptized on "fool's gold"! Of interest here are not the ontological aspects of this possibility, but the epistemic ones: We could bootstrap successfully to real gold even if every prior case had been fool's gold. "Gold" would still be the right word for what we had been trying to pick out (i.e. what we had "had in mind") all along, and its original provisional features would still have provided a close enough approximation to ground it, even if later information were to pull the ground out from under it, so to speak.

[#20]Although it is beyond the scope of this paper to discuss it at length, it must be mentioned that this question has often been begged in the past, mainly on the grounds of "vanishing intersections". It has been claimed that one cannot find invariant features in the sensory projection because they simply do not exist: The intersection of all the projections of the members of a category such as "horse" is empty. The British empiricists have been criticized for thinking otherwise; for example, Wittgenstein's [45] discussion of "games" and "family resemblances" has been taken to have discredited their view. Current research on human categorization [35] has been interpreted as confirming that intersections vanish and that hence categories are not represented in terms of invariant features. The problem of vanishing intersections (together with Chomsky's [2] "poverty of the stimulus argument") has even been cited by thinkers such as Fodor [8, 9] as a justification for extreme nativism. The present paper is frankly empiricist. In my view, the reason intersections have not been found is that no one has yet looked for them properly. Introspection certainly is not the way to look. and general pattern learning algorithms such as connectionism are relatively new; their inductive power remains to be tested. In addition, a careful distinction has not been made between pure sensory categories (which, I claim, must have invariants, otherwise we could not successfully identify them as we do) and higher-order categories that are *grounded* in sensory categories; these abstract representations may be symbolic rather than sensory, and hence not based directly on sensory invariants. For further discussion of this problem, see ref. [14].

tion. In the hybrid system proposed here, there is no longer any autonomous symbolic level at all; instead, there is an intrinsically *dedicated* symbol system, its elementary symbols (names) connected to nonsymbolic representations that can pick out the objects to which they refer, via connectionist networks that extract the invariant features of their analog sensory projections.

5. Conclusions

The expectation has often been voiced that "top-down" (symbolic) approaches to modeling cognition will somehow meet "bottom-up" (sensory) approaches somewhere in between. If the grounding considerations in this paper are valid, then this expectation is hopelessly modular and there is really only one viable route from sense to symbols: from the ground up. A free-floating symbolic level like the software level of a computer will never be reached by this route (or vice versa) – nor is it clear why we should even try to reach such a level, since it looks as if getting there would just amount to uprooting our symbols from their intrinsic meanings (thereby merely reducing ourselves to the functional equivalent of a programmable computer).

In an intrinsically dedicated symbol system there are more constraints on the symbol tokens than merely syntactic ones. Symbols are manipulated not only on the basis of the arbitrary shape of their tokens, but also on the basis of the decidedly nonarbitrary "shape" of the iconic and categorical representations connected to the grounded elementary symbols out of which the higher-order symbols are composed. Of these two kinds of constraints, the iconic/categorical ones are primary. I am not aware of any formal analysis of such dedicated symbol systems[#21], but this may be because they are unique to cognitive and robotic modeling and their properties will depend on the specific kinds of robotic (i.e. behavioral) capacities they are designed to exhibit.

It is appropriate that the properties of dedicated symbol systems should turn out to depend on behavioral considerations. The present grounding scheme is still in the spirit of behaviorism in that the only tests proposed for whether a semantic interpretation will bear the semantic weight placed on it consist of one formal test (does it meet the eight criteria for being a symbol system?) and one behavioral test (can it discriminate, identify (manipulate) and describe all the objects and states of affairs to which its symbols refer?). If both tests are passed, then the semantic interpretation of its symbols is "fixed" by the behavioral capacity of the dedicated symbol system, as exercised on the objects and states of affairs in the world to which its symbols refer; the symbol meanings are accordingly not just parasitic on the meanings in the head of the interpreter, but intrinsic to the dedicated symbol system itself. This is still no guarantee that our model has captured subjective meaning, of course. But if the system's behavioral capacities are lifesize, it is as close as we can ever hope to get.

[#21] Although mathematicians investigate the formal properties of uninterpreted symbol systems, all of their motivations and intuitions clearly come from the intended interpretations of those systems (see ref. [31]). Perhaps these too are grounded in the iconic and categorical representations in their heads.

References

[1] A.C. Catania and S. Harnad, eds., The Selection of Behavior. The Operant Behaviorism of B.F. Skinner: Comments and Consequences (Cambridge Univ. Press, Cambridge, 1988).

[2] N. Chomsky, Rules and representations, Behav. Brain Sci. 3 (1980) 1–61.

[3] M. Davis, Computability and Unsolvability (McGraw-Hill, New York, 1958).

[4] M. Davis, The Undecidable (Raven, New York, 1965).

[5] D.C. Dennett, Intentional systems in cognitive ethology, Behav. Brain Sci. 6 (1983) 343–390.

[6] J.A. Fodor, The Language of Thought (Crowell, New York, 1975).

[7] J.A. Fodor, Methodological solipsism considered as a research strategy in cognitive psychology, Behav. Brain Sci. 3 (1980) 63–109.

[8] J.A. Fodor, Précis of the modularity of mind, Behav. Brain Sci. 8 (1985) 1–42.
[9] J.A. Fodor, Psychosemantics (MIT/Bradford, Cambridge, MA, 1987).
[10] J.A. Fodor and Z.W. Pylyshyn, Connectionism and cognitive architecture: A critical appraisal, Cognition 28 (1988) 3–71.
[11] J.J. Gibson, An ecological approach to visual perception (Houghton Mifflin, Boston, 1979).
[12] S. Harnad, Metaphor and mental duality, in: Language, Mind and Brain, eds. T. Simon and R. Scholes (Erlbaum Hillsdale, NJ, 1982).
[13] S. Harnad, Categorical perception: A critical overview, in: Categorical Perception: The Groundwork of Cognition, ed. S. Harnad (Cambridge Univ. Press, Cambridge, 1987).
[14] S. Harnad, Category induction and representation, in: Categorical Perception: The Groundwork of Cognition, ed. S. Harnad (Cambridge Univ. Press, Cambridge, 1987).
[15] S. Harnad, Minds, machines and searle, J. Theor. Exp. Artificial Intelligence 1 (1989) 5–25.
[16] S. Harnad, Computational hermeneutics, Social Epistemology, in press.
[17] J. Haugeland, The nature and plausibility of cognitivism, Behav. Brain Sci. 1 (1978) 215–260.
[18] S.C. Kleene, Formalized Recursive Functionals and Formalized Realizability (Am. Math. Soc., Providence, RI, 1969).
[19] S.A. Kripke, Naming and Necessity (Harvard Univ. Press, Cambridge, MA, 1980).
[20] A.M. Liberman, On the finding that speech is special, Am. Psychologist 37 (1982) 148–167.
[21] J. Lucas, Minds, machines and Gödel, Philosophy 36 (1961) 112–117.
[22] J. McCarthy and P. Hayes, Some philosophical problems from the standpoint of artificial intelligence, in: Machine Intelligence, Vol. 4, eds. B. Meltzer and P. Michie (Edinburgh Univ. Press, Edinburgh, 1969).
[23] J.L. McClelland, D.E. Rumelhart and the PDP Research Group, Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1 (MIT/Bradford, Cambridge, MA, 1986).
[24] D. McDermott, Artificial intelligence meets natural stupidity, SIGART Newsletter 57 (1976) 4–9.
[25] G.A. Miller, The magical number seven, plus or minus two: Some limits on our capacity for processing information, Psychological Rev. 63 (1956) 81–97.
[26] M. Minsky, A framework for representing knowledge, MIT Lab Memo No. 306 (1974).
[27] M. Minsky and S. Papert, Perceptrons: An Introduction to Computational Geometry (MIT Press, Cambridge, MA, 1969) Reissued in an expanded edition (1988).
[28] A. Newell, Physical symbol systems, Cognitive Sci. 4 (1980) 135–183.
[29] U. Neisser, Cognitive Psychology (Appleton-Century-Crofts., New York, 1967).
[30] A. Paivio, Mental Representation: A Dual Coding Approach (Oxford Univ. Press, Oxford, 1986).
[31] R. Penrose, The Emperor's New Mind (Oxford Univ. Press, Oxford, 1989).
[32] Z.W. Pylyshyn, Computation and cognition: Issues in the foundations of cognitive science, Behav. Brain Sci. 3 (1980) 111–169.
[33] Z.W. Pylyshyn, Computation and Cognition (MIT/Bradford, Cambridge, MA, 1984).
[34] Z.W. Pylyshyn, ed., The Robot's Dilemma: The Frame Problem in Artificial Intelligence (Ablex, Norwood, NJ, 1987).
[35] E. Rosch and B.B. Lloyd, Cognition and Categorization (Erlbaum, Hillsdale, NJ, 1978).
[36] F. Rosenblatt, Principles of Neurodynamics (Spartan, New York, 1962).
[37] J. Searle, Minds, brains and programs, Behav. Brain Sci. 3 (1980) 417–457.
[38] R.N. Shepard and L.A. Cooper, Mental Images and Their Transformations (MIT Press/Bradford, Cambridge, 1982).
[39] P. Smolensky, On the proper treatment of connectionism, Behav. Brain Sci. 11 (1988) 1–74.
[40] E.P. Stabler, How are grammars represented? Behav. Brain Sci. 6 (1985) 391–421.
[41] H. Terrace, Nim (Random House, New York, 1979).
[42] J. Turkkan, Classical conditioning: The new hegemony, Behav. and Brain Sci. 12 (1989) 121–179.
[43] A.M. Turing, Computing machinery and intelligence, in: Minds and Machines, ed. A. Anderson (Prentice Hall, Engelwood Cliffs, NJ, 1964).
[44] S. Ullman, Against direct perception, Behav. Brain Sci. 3 (1980) 373–415.
[45] L. Wittgenstein, Philosophical Investigations (Macmillan, New York, 1953).