COGNITIVE SCIENCE **18**, 355–360 (1994)

REPLY

Reply to Touretzky and Pomerleau: Reconstructing Physical Symbol Systems

ALONSO H. VERA AND HERBERT A. SIMON
Carnegie Mellon University

We agree with much of Touretzky and Pomerleau's (1994) critique of situated action theory, but reject their narrow construal of the term "symbol." Contrary to their claim, defining symbols as patterns that possess denotations does not strip the concept of meaning or deprive the physical symbol system hypothesis (PSSH) of significance as a theory of intelligence.

Hence, our disagreement with Touretzky and Pomerleau (1994) boils down to the definition of "symbol." Definitions are not matters of fact, but of convention; definitions should be chosen for convenience and clarity. In our paper (Vera & Simon, 1993) we used the term "symbol" in its conventional sense, which we believe to be a convenient usage.

Our usage is not a neologism, for *Webster's New International Dictionary* (1924) carries as the first definition of "symbol": "that which refers to or represents something else." Reference to the philosophical literature (e.g., Russell & Whitehead, Russell's *Inquiry into Meaning and Truth,* Carnap, Church's *Introduction to Mathematical Logic,* Kleene) reveals that the term is used nontechnically and infrequently, and then in a sense consistent with ours. In logic (e.g., textbooks by Mendelsohn, Schoenfeld, Rosenbloom), it frequently has a rather special use: symbols are the primitive elements of a logic; strings of symbols (symbol structures) are the well-formed expressions (which by no means fit the requirement of Touretzky and Pomerleau, 1994, that they be "arbitrary"). Turing speaks of symbols only in association with a particular machine, and his symbols are not arbitrary, for you cannot change them without changing the operation of the machine. In sum, we find nothing in the literature prior to *Human problem solving* (Newell & Simon, 1972) that forecloses our definition of symbols as patterns that denote, and Harnad's (1992) narrower definition to which they refer must be regarded as a departure from common usage.

But we do not rest our case simply on conventional usage. Touretzky and Pomerleau's (1994) concern that our definition of symbol is so broad as to be vacuous is most simply refuted by pointing to the many examples of patterns that do *not* have denotations, hence are not, by our definition, symbols.

We can begin with snowflakes, and proceed through the whole vast collection of nondenotative patterns exhibited by natural (and artificial) systems: The wind ripples on lakes, the shapes of birds, the masses of clouds, crystals, the grain in wood, a Jackson Pollock painting, the performance of a Mozart sonata, a thunderclap—where does the list stop?

Therefore, to say that a particular pattern denotes, hence is a symbol, distinguishes it from most other patterns in the world, and says something quite strong about its functioning. In particular, a symbol in a physical symbol system (PSS) can be manipulated to obtain information and to draw inferences about the thing it denotes. The property of denoting has strong consequences: the consequences that characterize PSSs.

Touretzky and Pomerleau (1994) argue that "if everything from the retina on up were a symbol system, that is, if reflexive object tracking were symbol processing, then 'symbol' would be synonymous with 'signal'." (p. 348) They wish to require that symbols have arbitrary shapes unrelated to their meanings, and that symbol structures be recursively composable by rule. They fail to see that these two requirements are effectively incompatible, for composed symbol structures are not arbitrary. Symbols can be arbitrary with respect to the objects they designate, but they do not necessarily need to be so. In our definition of symbols, as patterns that denote, we of course include recursively composed symbol structures, for these are also patterns that denote. This inclusion is explicit in the usual definition of a PSS (e.g., Newell & Simon, 1972).

Symbols include some types of signals, such as those on the retina, but they are not restricted to being nothing but signals. They can also refer to patterns with nonarbitrary relations to their referents. As long as the abstracted pattern or representation encodes information that will be interpreted (and which has the functional role of carrying interpretable information) to generate action, then it is a symbol. Hemoglobin in the circulatory system actually carries oxygen; it does not just represent it, nor do the recipient cells interpret it. Antibodies and DNA, on the other hand, lie in a fuzzier area with respect to the symbol–signal distinction. These biological processes involve physical patterns that are interpreted, albeit in a low-level mechanical way. In this respect, these latter systems are more like thermostats than like hemoglobin. They interpret a physical signal pattern according to a predetermined symbol system (e.g., a sequence of animo acids or a scale of voltage values) to generate a representation (be it of a protein molecule or of current room temperature). This is the essence of the PSSH.

There can be systems that are not PSSs and, nevertheless, show some level of intelligent behavior (as Newell suggested in the passage Touretzky & Pomerleau, 1994, quoted on p. 348). However, we do not believe that ALVINN or the creatures are examples of such a system. The circulatory system, which carries out a variety of complex tasks, is such a system. If sym-

bols are to include analog representations and patterns with nonarbitrary shapes (as we suggest, and Touretzky & Pomerleau appear to accept), then systems such as ALVINN and Brooks's creatures make good examples of PSSs. If Touretzky and Pomerleau do not accept this, then NAVLAB's "symbolic map component" (p. 348) would have to be considered nonsymbolic, for maps are excellent examples of analog representation. Maps are, without question, patterns that hold nonarbitrary relations to their geographical referents.

Because recursively composed symbol structures do not have arbitrary shapes (for the rules of composition are not arbitrary) they would not qualify as symbols in the Touretzky and Pomerleau terminology, a very inconvenient result. In particular, Touretzky and Pomerleau explicitly rule out analogs as symbol structures. But analogs, because of their rich representational structure, have been, in science, fruitful sources of powerful symbolic representations. We see no problem in regarding the voltage level in a thermostat as a symbol that denotes the temperature. (Moreover, the relation of symbol to denotation is quite arbitrary in this case because it would be easy to design the device so that a higher voltage would denote a lower, instead of a higher, temperature. Would this device be more symbolic than the usual one?)

Nor do we understand why Touretzky and Pomerleau think analogies lack combinatorial power. We only need to look at Bohr's great contribution to quantum theory in 1913, to see how a classical picture of a planetary system (Rutherford's analogical picture of the atom) was modified and combined with (analogized with) Planck's notion of the energy quantum to produce the first model of the hydrogen atom. The whole history of physics is full of analogical symbol structures transported, with modification and combination, from one application to another: the combined wave-particle model of modern quantum mechanics being one of the more recent examples.

Psychologists have long been aware that these sorts of analogical processes are a central part of our reasoning abilities (see Medin, Goldstone, & Gentner, 1993; and Holyoak & Thagard, 1993, for reviews of this area). Research on analogy has focused on understanding what information, from specific primitive features to broad causal relations, is transferred when we reason by direct comparison. Analogy, as a cognitive activity, is a decomposable phenomenon that most psychologists would view as compatible with the PSSH.

Touretzky and Pomerleau (1994) claim that the sufficiency of PSSH is "a given, unless one is a dualist or some other sort of mystic" (p. 348). This is not a logical consequence of the definitions. If true, it is an empirical result of the fact that PSSs can be made that behave intelligently over a wide range of tasks; and it is widely disbelieved by those committed to situated action. But if the PSSH is accepted, why, as Touretzky and Pomerleau

claim, are there in biology "practical reasons for performing information-processing functions using 'direct,' nonsymbolic means where possible" (p. 348)? What are these practical reasons, and what are the means that do not fit our definition of symbolic?

While discussing levels of processing, Touretzky and Pomerleau assert that "a representation becomes a symbol system at the point where it acquires combinatorial power" (p. 348). This condition clearly is met by the retinal arousal (in ALVINN or in people) produced by patterns of light, and the resulting retinal patterns are interpreted and acted upon in other parts of the brain. Surely Touretzky and Pomerleau do not wish to claim that the arousal of each rod or cone is interpreted independently of the arousal of the others. In actual fact, these arousals constitute a pattern that is analogical along many dimensions to the arrangement of objects that were the sources of the stimulating radiation. The retinal pattern denotes this arrangement, hence is a symbol.

In particular, the hidden-layer activity patterns in ALVINN are clearly combinatorial. As Touretzky and Pomerleau themselves point out, "analysis of [these patterns'] input weights and response properties reveals that they are responding to objects such as road edges and lane stripes" (p. 348): They implement the (analogical) symbol, road-edge or lane-stripe, and denote the corresponding properties of the external scene.

That our disagreements with Touretzky and Pomerleau are largely definitional is evident in the last three sections of their comment: on nonarbitrary symbols, symbols versus concepts, and automaticity.

As to nonarbitrary symbols, they begin by admitting that "the requirement that symbols have purely arbitrary shapes unrelated to their meanings is probably not met fully in humans, not because a pure physical symbol system would be inadequate to the task of cognition, but because of the process by which humans were constructed" (p. 349). But they nowhere tell us in what respects evolution found symbol systems (in our sense of the term) wanting.

Their characterization of many biological processes as nonsymbolic also derives from a too-narrow definition of symbol. When people burn their fingers and a subcortical reaction causes them to remove their hand quickly, the single (evolutionarily derived) function of the nerve impulse is to indicate that bodily damage is occurring. The nerve impulse does not carry the burn in any sense; it communicates its occurrence. This symbol denoting the event is transmitted to the spinal cord, which then transmits a symbol to the muscles, which contract to pull the hand away. Although subcortical, this process is a good example of a PSS at work, much like a thermostat.

Touretzky and Pomerleau then provide an interpretation of affordances that coincides almost exactly with ours—except that, because of our disagree-

ment about definition (specifically, whether symbols can be analogical)—they choose to call "nonsymbolic, preconscious processors" (p. 349) what we call symbolic processors.

The distinction that Touretzky and Pomerleau make between symbols and concepts stems from the same definitional dispute. They make the quite unmotivated claim that "cognitive theories deal in concepts not symbols" (p. 350). It is true that some linguists have this belief, but it is not shared by most cognitive psychologists, whose theories deal with primitive features of more complex "concepts" (e.g., the features that make a Roman letter recognizable) quite as much as they deal with the complexes (the letters and the words composed of the letters). The EPAM model of perception and memory is a symbolic theory of exactly this kind (Feigenbaum & Simon, 1984), which handles comfortably the things that Touretzky and Pomerleau call "subsymbolic," "subconceptual," and "primitive symbol tokens" (p. 350).

We do not disagree that symbols are more than purely arbitrary shapes with respect to their referents. This is the accepted view, at least in psychology, and it was our position in the original article (Vera & Simon, 1993). It was this view that led us to include systems such as ALVINN and Brooks's creatures as examples of PSSs. The alternative view of symbols as arbitrary atomic pointers exists only in a few AI communities. We did not use the narrower AI definition suggested, and thus never had a problem reconciling the notion of symbols with that of concepts. Concepts are an important construct in psychology and it has consequently always been important for the PSSH to account for empirical evidence about them. We do not believe that Touretzky and Pomerleau are suggesting that the PSSH has lost explanatory power as psychology has moved from the classical view of concepts to similarity-based ones, and more recently to theory-based views.

Touretzky and Pomerleau's "concepts" is obviously embraced by our "symbol structures." They insist that the primitive symbol tokens of which concepts are composed are indescribable, but do not explain how this prevents these primitives from being denotative. They then further confuse matters by permitting some "patterns" to play the role of symbols and serve as "reduced descriptions" (p. 350). Where now is the arbitrariness of patterns? What they call a "connectionist reformulation of symbol processing" (p. 350) could as readily be described as a "symbol-processing reformulation of connectionism" (p. 350). It is indeed an "intriguing alternative" (p. 350) and it is essentially the alternative that we espouse.

In their discussion of automaticity, Touretzky and Pomerleau now place great emphasis on the conscious–subconscious distinction. But, as they admit earlier, that distinction has nothing to do with whether something is a PSS; symbols may as readily be processed subconsciously as consciously, and are so processed in symbolic models of recognition processes.

We have no differences or quarrels with the Touretzky and Pomerleau description of motor skills, except for our discomfort with their new terminology. Connectionist systems certainly differ in important respects from "classical" simulations of human cognition. Contrary to the view of Touretzky and Pomerleau, symbolic–nonsymbolic is not one of the dimensions of this difference.

REFERENCES

Bohr, N. (1913). On the constitution of atoms and molecules. *Philosophical Magazine,* Sixth series, *26,* 1–25.

Carnap, R. (1937). *The logical syntax of language.* London: Routledge and Kegan Paul.

Church, A. (1944). *Introduction to mathematical logic.* Princeton: Princeton University Press.

Feigenbaum, E.A., & Simon, H.A. (1984). EPAM-like models of recognition and learning. *Cognitive Science, 8,* 305–336.

Harnard, S. (1992). Connecting object to symbol in modeling cognition. In A. Clarke & R. Lutz (Eds.), Connectionism in context. New York: Springer-Verlag.

Holyoak, K.J., & Thagard, P. (1993). *Mental leaps: Analogy in creative thought.* Cambridge, MA: MIT Press.

Kleene, S.C. (1952). *Introduction to metamathematics.* Amsterdam: North-Holland.

Medin, D., Goldstone, R., & Gentner, D. (1993). Respects for similarity. *Psychological Review, 100,* 254–278.

Mendelson, E. (1964). *Introduction to mathematical logic.* Princeton, NJ: Van Nostrand.

Newell, A., & Simon, H.A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice-Hall.

Rosenbloom, P.C. (1950). *The elements of mathematical logic.* New York: Dover Publications.

Russell, B. (1940). *An inquiry into meaning and truth.* London: G. Allen and Unwin.

Shoenfield, J.R. (1967). *Mathematical logic.* Reading, MA: Addison-Wesley.

Touretzky, D.S., & Pomerleau, D.A. (1994). Reconstructing physical symbol systems. *Cognitive Science, 18,* 345–353.

Vera, A.H., & Simon, H.A. (1993). Situated action: A symbolic interpretation. *Cognitive Science, 17,* 7–48.

Whitehead, A.N., & Russell, B. (1925). *Principia mathematica* (2nd ed.). Cambridge: Cambridge University Press.