Representation and Mind
Hilary Putnam and Ned Block, editors

Representation and Reality, *by Hilary Putnam*
Explaining Behavior: Reasons in a World of Causes, *by Fred Dretske*

Representation and Reality

Hilary Putnam

A Bradford Book
The MIT Press
Cambridge, Massachusetts
London, England

with an appropriate part of its physical environment as analogous to a computer, and seek to describe functional relations within this larger system? Why not seek to characterize reference, in particular, as a functional relation between representations used by organisms and things which may be either inside or outside those organisms?" Although the discussion which followed focused on what I have just described as the "epistemological" difficulties, there is a remarkable "ontological" presupposition contained in the very statement of the project. *The project simply assumes from the outset that there is a single system ("the organisms and their physical environment") which contains all the objects that anyone could refer to.* The picture is that there is a certain domain of entities such that all ways of using words referentially are just different ways of singling out one or more of those entities. In short, the picture is that what an "object" of reference is is fixed once and for all at the start, and that the totality of objects in some scientific theory or other will turn out to coincide with the totality of All The Objects There Are.

But, from my "internal realist" perspective at least, there is no such totality as All The Objects There Are, inside or outside science. "Object" itself has many uses, and as we creatively invent new uses of words, we find that we can speak of "objects" that were not "values of any variable" in any language we previously spoke. (The invention of "set theory" by Cantor is a good example of this.) What looked like an innocent formulation of the problem—"Here are the objects to be referred to. Here are the speakers using words. How can we describe the relation between the speakers and the objects?"—becomes far from innocent when what is wanted is not a "natural-language processor" that works in some restricted context, but a "theory of reference." From an internal realist point of view, the very problem is nonsensical.

Of course, from my point of view the "epistemological" and the "ontological" are intimately related. Truth and reference are intimately connected with epistemic notions; the open texture of the notion of an object, the open texture of the notion of reference, the open texture of the notion of meaning, and the open texture of reason itself are all interconnected. It is from these interconnections that serious philosophical work on these notions must proceed.

Appendix

Theorem. Every ordinary open system is a realization of every abstract finite automaton.

Physical Principles. The proof I shall give requires the following two physical principles (which hold in classical physics when (1) the fields have no sources except particles; and (2) the number of point particles is at most denumerably infinite):

Principle of Continuity. The electromagnetic and gravitational fields are continuous, except possibly at a finite or denumerably infinite set of points. (Since we assume that the only sources of fields are particles, and that there are singularities only at point particles, this has the status of a physical law.)

Principle of Noncyclical Behavior. The system S is in different maximal states at different times. This principle will hold true of all systems that can "see" (are not shielded from electromagnetic and gravitational signals from) a clock. Since there are natural clocks from which no ordinary open system is shielded, all such systems satisfy this principle. (N.B.: It is not assumed that *this* principle has the status of a physical law; it is simply assumed that it is in fact true of all ordinary macroscopic open systems.)

In the sequel, we shall make use of the fact that this principle holds true both on the boundary of any ordinary open system (i.e., the state of the boundary of such a system is not the same at two different times) and a little way *inside* the boundary as well.

Lemma. If we form a system S' with the same spatial boundaries as S by stipulating that the conditions *inside* the boundary are to be the conditions that obtained inside S at time t while the conditions *on* the boundary are to be the ones that obtained on the boundary of S at time t', where $t \neq t'$ [note that this will be possible only if the spatial boundary assigned to the system S is the same at t and at t'], then the resulting system will violate the Principle of Continuity.

Proof (of the lemma): Every ordinary open system is exposed to signals from many clocks C (say, from the solar system, or from things

which contain atoms undergoing radioactive decay, or from the system itself if it contains such radioactive material—in which latter case the system S itself coincides with the clock C). In fact, according to physics, there are signals from C from which it is not possible to shield S (for example, gravitational signals). These signals from C may be thought of, without loss of generality, as forming an "image" of C on the surface of S. For the same reason, there are also "images" of C *inside* the boundary of S. The "image" of C at, say, $t' = 12$ may be thought of as showing a "hand at the 12 position"; while the "image" of C at, say, $t = 11$ shows a "hand at the 11 position." Thus, for these values of t and t', the system S' would have a "12 image" on its boundary and an "11 image" at an arbitrary small distance inside its boundary; but this is to say that the fields which constitute the "images" would have a discontinuity along an entire continuous area, and hence at nondenumerably many points.

Proof of the Theorem. (I have stated the theorem in terms of finite automata, but the technique is easily adapted to other formalisms.) A finite automaton is characterized by a table which specifies the states and the required state-transitions. Without loss of generality, let us suppose the table calls for the automaton to go through the following sequence of states in the interval (in terms of "machine time") that we wish to simulate in real time: *ABABABA*. Let us suppose we are given a physical system S whose spatial boundary we have exactly defined, at least during the real-time interval we are interested in (say, a given 7-minute interval, e.g., from 12:00 to 12:07). We wish to find physical states A and B such that during the time interval we are interested in the system S "obeys" this table by going through the sequence of states *ABABABA*, and such that given just the laws of physics (including the Principle of Continuity) and the boundary conditions of S, a Laplacian supermind could predict the next state of the system (e.g., that S will be in state B from 12:03 to 12:04) given the previous state (given that S was in state A from 12:02 to 12:03). This will show that S "realizes" the given table during the interval specified. Since the technique of proof applies to *any* such table, we will have proved that S can be ascribed any machine table at all, and the description will be a "correct" one, in the sense that there really are physical states with respect to which S is a realization of the table ascribed.

I shall use the symbolic expression $St\ (S, t)$ to denote the maximal state of S at t (in classical physics this would be the value of all the field parameters at all the points inside the boundary of S at t). Let the beginnings of the intervals during which S is to be in one of its

stages A or B be $t_1, t_2, \ldots, t_n$ (in the example given, $n = 7$, and the times in question are $t_1 = 12{:}00$, $t_2 = 12{:}01$, $t_3 = 12{:}02$, $t_4 = 12{:}03$, $t_5 = 12{:}04$, $t_6 = 12{:}05$, $t_7 = 12{:}06$). The end of the real-time interval during which we wish S to "obey" this table we call t_{n+1} ($= t_8 = 12{:}07$, in our example). For each of the intervals t_i to t_{i+1}, $i = 1,2, \ldots, n$, define a (nonmaximal) *interval state* s_i which is the "region" in phase space consisting of all the maximal states $St\ (S, t)$ with $t_i \leq t < t_{i+1}$. (I.e., S is in s_i just in case S is in one of the maximal states in this "region.") Note that the system S is in s_1 from t_1 to t_2, in s_2 from t_2 to t_3, . . . , in s_n from t_n to t_{n+1}. (Left endpoint included in all cases but not the right—this is a convention to ensure the "machine" is in exactly one of the s_i at a given time.) The disjointness of the states s_i is guaranteed by the Principle of Noncyclical Behavior.

Define $A = s_1 \vee s_3 \vee s_5 \vee s_7$; $B = s_2 \vee s_4 \vee s_6$.

Then, as is easily checked, S is in state A from t_1 to t_2, from t_3 to t_4, from t_5 to t_6, and from t_7 to t_8, and in state B at all other times between t_1 and t_8. So S "has" the table we specified, with the states A,B we just defined as the "realizations" of the states A,B described by the table.

To show that being in state A at times t with $t_1 \leq t < t_2$ "caused" S to go into state B during the interval $t_2 \leq t < t_3$ (and similarly for the other state transitions called for by the table), we argue as follows: Given that S is in state A at a time t ($t_1 \leq t < t_2$), and letting the maximal state of the boundary of S at that time t be B_t, it follows from the lemma that $St\ (S, t)$ is the only maximal state in any of the "regions" (nonmaximal states) $s_1, s_2, \ldots, s_7$ that a system S under the boundary condition B_t could be in without violating the Principle of Continuity. (If the shape, size, or location of S changes with time, then unless S resumes the boundary it had at t at least once, the boundary of S at t will be the only boundary associated with any maximal state in the union of these regions which fits the boundary condition B_t, and the lemma is unnecessary.) *A fortiori*, $St\ (S, t)$ is the only maximal state in A compatible with B_t. Hence, given the information that the system was in state A at t, and given the information that the boundary condition at t was B_t, a mathematically omniscient being can determine from the Principle of Continuity that the system S must have been in $St\ (S, t)$, and can further determine, given the boundary conditions at subsequent times and the other laws of nature, how S evolves in the whole time interval under consideration. Q.E.D.

Discussion. When we model cognitive functions, we do not, of course, model them by means of automata without inputs and outputs.

Rather, we imagine that the "automaton" is connected with input devices—sensors, such as eyes or ears (or, in the simplest case, a "paper tape" on which the operator can print messages in a specified alphabet); and also connected with output devices—motor organs, speech organs, etc. (or, in the case originally imagined by Turing, another "paper tape" on which the automaton can print messages in another specified alphabet). These inputs and outputs have specified realizations, or at least their realizations must be of certain constrained kinds depending on our purposes; usually we are not allowed to simply *pick* physical states to serve as their "realizations," as we are allowed to do with the so-called "logical states" of the automaton.

If a physical object does not have motor organs or sensors of the specified kind, then, of course, it cannot be a model of a description which refers to a kind of automaton which, *ex hypothesi*, possesses motor organs and sensors of that kind. And even if it does possess such "inputs" and "outputs," it may behave in a way which violates predictions which follow from the description (e.g., print two "1"s in a row when it is a theorem that the machine with the given description never does this). So there is no hope that the theorem just proved will also hold, unchanged, for automata which have inputs and outputs which have been specified (or at least constrained) in physical terms.

Imagine, however, that an object S which takes strings of "1"s as inputs and prints such strings as outputs behaves from 12:00 to 12:07 exactly as *if* it had a certain description D. That is, S receives a certain string, say "111111," at 12:00 and prints a certain string, say "11," at 12:07, and there "exists" (mathematically speaking) a machine with description D which does this (by being in the appropriate state at each of the specified intervals, say 12:00 to 12:01, 12:01 to 12:02, . . . , and printing or erasing what it is supposed to print or erase when it is in a given state and scanning a given symbol). In this case, S too can be *interpreted* as being in these same logical states $A,B,C, \ldots$ at the very same times and following the very same transition rules; that is to say, we can find *physical* states $A,B,C, \ldots$ which S possesses at the appropriate times and which stand in the appropriate causal relations to one another and to the inputs and the outputs. The method of proof is exactly the same as in the theorem just proved (the unconstrained case). Thus we obtain that *the assumption that something is a "realization" of a given automaton description (possesses a specified "functional organization") is equivalent to the statement that it behaves as if it had that description.* In short, "functionalism," if it were

correct, would imply behaviorism! If it is true that to possess given mental states is simply to possess a certain "functional organization," then it is also true that to possess given mental states is simply to possess certain behavior dispositions!