Matter, Mind and Models
Marvin L. Minsky

Published in Semantic Information Processing, MIT Press, 1968
(This version has some editorial simplifications made in 1995)

1 Introduction

        This chapter attempts to explain why people become confused by
questions about the relation between mental and physical events.
When a question leads to confused, inconsistent answers, this may
be because the question is ultimately meaningless or at least
unanswerable, but it may also be because an adequate answer
requires a powerful analytical apparatus. It is the author's view that
many important questions about the relation between mind and brain
are of that second kind, and that some of the necessary technical and
conceptual tools are becoming available as a result of work on the
problems of making computer programs behave intelligently. We
shall suggest a theory to explain why introspection does not give
clear answers to these questions. Technical solutions to the questions
will not be attempted, but there is probably some value in finding at
least a clear explanation of why we are confused.

2 Knowledge and Models

        If a creature can answer a question about a hypothetical
experiment without actually performing it, then it has demonstrated
some knowledge about the world. For, his answer to the question
must be an encoded description of the behavior (inside the creature)
of some sub-machine or "model" responding to an encoded
description of the world situation described by the question We use
the term "model" in the following sense: To an observer B, an object
A* is a model of an object A to the extent that B can use A* to
answer questions that interest him about A. The model relation is
inherently ternary. Any attempt to suppress the role of the intentions
of the investigator B leads to circular definitions or to ambiguities
about "essential features" and the like. It is understood that B's use
of a model entails the use of encodings for input and output, both for
A and for A*. If A is the world, questions for A are experiments. A*
is a good mode of A, in B's view, to the extent that A*'s answers
agree with those of A's, on the whole, with respect to the questions
important to B. When a man M answers questions about the world,
then (taking on ourselves the role of B) we attribute this ability to
some internal mechanism W* inside M. It would be most convenient
if we could discern physically within M two separate regions, W*
and M-W*, such that W* "really contains the knowledge" and M-W*
contains only general-purpose machinery for coding questions,
decoding answers, or administering the thinking process. However,
one cannot ready expect to find, in an intelligent machine, a clear
separation between coding and knowledge structures, either
anatomically or functionally, because (for example) some
"knowledge" is likely to be used in the encoding and interpreting
processes. What is important for our purposes is the intuitive notion
of a model, not the technical ability to delineate a model's boundaries
Indeed, part of our argument hinges on the inherent difficulty of
discerning such boundaries.

3 Models of Models

Questions about things in the world are answered by making statements about the behavior of corresponding structures in one's model W* of the world. For simple mechanical, physical, or geometric matters one can imagine, as did Craik (1), some machinery that does symbolic calculation but when read through proper codings has an apparently analogue character. But what about broader question about the nature of the world? These have to be treated (by M) not as questions to be answered by W*, but as questions to be answered by making general statements about W*. If W** contains a model M* of M then M* can contain a model W** of W*; and, going one step further, W** may contain a model M** of M*. Indeed, this must be the case if M is to answer general questions about himself. Ordinary questions about himself, e.g., how tall he is, are answered by M*, but very broad questions about his nature, e.g., what kind of a thing he is, etc., are answered, if at all, by descriptive statements made by M** about M*.

The reader may be anxious, at this point, for more details about the relation between W* and W**. How can he tell, for example, when a question is of the kind that requires reference to W** rather than to W*. Is W** a part of W? (Certainly W*, like everything else, is part of W.) Unfortunately, I cannot supply these details yet, and I expect serious problems in eventually clarifying them. We must envision W** as including an interpretative mechanism that can make references to W*, using it as a sort of computer-program subroutine, to a certain depth of recursion. In this sense W** must contain W*, but in another, more straightforward, sense W* can contain W**. This suggests first that the notion "contained in" is not sufficiently sophisticated to describe the kinds of relations between parts of program-like processes and second that the intuitive notion of "model" used herein is likewise too unsophisticated to support developing the theory in technical detail. It is clear that in this area one cannot describe inter-model relationships in terms of models as simple physical substructures. An adequate analysis will need much more advanced ideas about symbolic representation of information-processing structures.

4 Dimorphism of our World Models

A man's model of the world has a distinctly bipartite structure: One part is concerned with matters of mechanical, geometrical, physical character, while the other is associated with things like goals, meanings, social matters, and the like. This division of W* carries through the representations of many things in W*, especially to M itself. Hence, a man's model of himself is bipartite, one part concerning his body as a physical object and the other accounting for his social and psychological experience. When we see an object, we account for its mechanical support and coherence (we are amazed at levitations) and we also account, in different terms, for its teleology (who put it there and for what purpose). [When we see an object start to move we expect to find either a physical force or a psychological purpose in the kind of ordinary common-sense explanation that concerns us here.]

Why is this division so richly represented in language and thought? We recognize that a person's W* is not really two clearly disjoint parts but usually has more than one overlapping, indistinctly

bounded models. The two-part structure proposed here is only an approximation, and we do not really want to suggest that the argument depends at all on [any particular "dualistic" structure. However, human thought seems largely based on many such "dumb-bell" distinctions, such as between information versus energy, physical versus psychological, or analog versus digital]. In one sphere, mechanical-geometric constraints are powerful, e.g., impenetrability in the arrangements of physical objects, or (Piaget-style) conservations in their transformation. In the other sphere, one finds symbolic constraints of (substantially) equal power. The two domains overlap in many complicated ways: a child discovers mechanical obstacles (in the form, e.g., of limitations of reach, mobility, strength, and precision) to its psychological goals; it discovers emotional symbols in the geometric arrangements of facial expressions, and intentions in postural attitudes. In explanations of complicated things the two models become inextricably involved‹ viz. the imagery of the preceding sentences. These complications reflect the inadequacy of either model for description of complicated situations.

As for the genesis of such partitions, I suppose that they grow apart rather than together, on the whole. That is not to say that infantile, primitive models are more unitary, but rather that they are simply too indistinct to admit approximate boundaries. An infant is not a monist: It simply hasn't enough structure in its M* to be a dualist yet; it can hardly be said to have a position on the mind-body problem.

5 The Central Argument Belief in Dualism

When a man is asked a general question about his own nature, he will try to give a general description of his model of himself. That is, the question will be answered by M**. To the extent that M* is divided as we have supposed and that the man has discovered this (that is, this fact is now represented in M**), his reply will show this. His statement (his belief) that he has a mind as well a body is the conventional way to express the roughly bipartite appearance of his model of himself.

Because the separation of the two parts of M* is so indistinct and their interconnections are so complicated and difficult to describe, the man's further attempts to elaborate on the nature of this "mind-body" distinction are bound to be confused and unsatisfactory.

6 Heuristic Value of Quasi-Separate Models

From a scientific point of view, it is desirable to obtain a unitary model of the world comprising both mechanical and psychological phenomena. Such a theory would become available, for example, if the workers in artificial intelligence, cybernetics, and neurophysiology would all reach their goals. Still such a success might have little effect on the overall form of our personal world models. I maintain that for practical, heuristic reasons, these would still retain their form of quasi-separate parts. Even when a discipline is grossly transformed in techniques, bases, and concepts, it can maintain its identity if its problems and concerns remain grouped together for practical reasons. For example, Chemistry survives today as a science because the primitives of the quantum theory are a little too remote for direct application to practical problems; a

hierarchy of intermediate concepts is necessary to apply the theory to everyday problems. The primitive notions of physics, or even of neurophysiology, will be far too remote to be useful in commonsense explanations of the mental events of everyday life.

Thus synthesis by direct theoretical reduction is unlikely to have much effect on the overall form of W*. The practical need for approximately self- contained subtheories is too strong to resist in everyday life and thought. [One might search for another kind of unification in which the two quasi-separate models are described in similar ways and then merged by removing redundancy, with coding for those differences that remain significant.] It is doubtful that much can be done in this direction, because using psychological explanations for physical processes runs exactly counter to the directions that have led to scientific process. Similarly, there have long been available plenty of "reductions" of psychological explanations to analogies with simple physical systems, but these are recognized as inadequate and are giving way to information-processing models of more abstract character.

In everyday practical thought, physical analogy metaphors play a large role, presumably because one gets a large payoff for a model of apparently small complexity. (Actually, only the incremental complexity is small because most of the model is already there as part of the "physical" part of W*.) It would be hard to give up such metaphors, even though they probably interfere with our further development, just because of this apparent high value-to-cost ratio. We cannot expect to get much more by extending the mechanical analogies, because they are so inflexible in character. Mental processes resemble more the kinds of processes found in computer programs: arbitrary symbol- associations, treelike storage schemes, conditional transfers, and the like. In short, we can expect the simpler useful mechanical analogies to survive, but it seems doubtful that they can grow to bring us usable ideas for the parallel unification of W*.

Finally, we should note that in a creature with high intelligence one can expect to find a well-developed special model concerned with the creature's own problem-solving activity. In my view the key to any really advanced problem-solving technique must exploit some mechanism for planning‹for breaking the problem into parts and shrewdly allocating the machine's effort and resources for the work ahead. This means the machine must have facilities for representing and analyzing its own goals and resources. One could hardly expect to find a useful way to merge this structure with that used for analyzing uncomplicated structures in the outer world, nor could one expect that anything much simpler would be of much power in analyzing the behavior of other creatures of the same character.

7 Interpreters

The notion of "part" is more complicated for things like computer programs than for ordinary physical objects. A single conditional branch makes it possible for a program to behave, functionally, like two very different machines in different circumstances, yet using almost (or exactly) the same sets of instructions.

The notion of a machine containing a model of itself is also complicated, and one might suspect potential logical paradoxes. There is no logical problem about the basic idea, for the internal model could be very much simplified, and its internal model could be

vacuous. But, in fact, there is no paradox even in a machine's having a model of itself complete in all detail. For example, it is possible to construct a Turing machine that can print out an entire description of itself and also execute an arbitrarily complicated computation, so that the machine is not expending all its structure on its description. In particular, the machine can contain an "interpretative" program that can use the internal description to calculate what it itself would do under some hypothetical circumstance. Similarly, while it is impossible for a machine or mind to analyze from moment to moment precisely what it is doing at each step (for it would never get past the first step), there seems to be no logical limitation to the possibility of a machine understanding its own basic principles of operation or, given enough memory, examining all the details of its operation in some previously recorded state.

With interpretative operation ability, a program can use itself as its own model, and this can be repeated recursively to as many levels as desired, until the memory records of the state of the process get out of hand. With the possibility of this sort of "introspection," the boundaries between parts, things, and models become very hard to understand.

Does interpreted operation play an important role in our mental function? It is clear that one interprets memorized instructions in certain circumstances One could memorize, for example, the rules for reading musical notation and then actually perform a piece of music, at a very slow tempo, by referring to these rules in executing each note. Eventually, with practice, one plays faster, and it seems clear that one is no longer interpreting the rules for each note, but that one has assembled special mechanisms for the task. This certainly suggests an analogy with the notion of "compiling" a previously interpreted program. Perhaps our level of consciousness is closely related to the extent to which the machine is functioning interpretively rather than executing compiled programs. While interpreting, one has the opportunity of examining the next step in the task before doing it. [I'm using the term "interpretative" here technically, to refer to a process that decodes a high-level programming language while it is running, rather than "compiling" low level code in advance.]

8 Free Will

If one thoroughly understands a machine or a program, he finds no urge to attribute "volition" to it. If one does not understand it so well, he must supply an incomplete model for explanation. Our everyday intuitive models of higher human activity are quite incomplete, and many notions in our informal explanations do not tolerate close examination. Free will or volition is one such notion: people are incapable of explaining how it differs from stochastic caprice but feel strongly that it does. I conjecture that this idea has its genesis in a strong primitive defense mechanism. Briefly, in childhood we learn to recognize various forms of aggression and compulsion and to dislike them, whether we submit or resist. Older, when told that our behavior is "controlled" by such and such a set of laws, we insert this fact in our model (inappropriately) along with other recognizers of compulsion. We resist "compulsion," no matter from "whom." Although resistance is logically futile, the resentment persists and is rationalized by defective explanations, since the alternative is emotionally unacceptable.

How is this reflected in M**? If one asks how one's mind works,

he notices areas where it is (perhaps incorrectly) understood<that is, where one recognizes rules. One sees other areas where he lacks rules. One could fill this in by postulating chance or random activity. But this too, by another route, exposes the self to the same indignity of remote control. We resolve this unpleasant form of M** by postulating a third part, embodying a will or spirit or conscious agent. But there is no structure in this part; one can say nothing meaningful about it, because whenever a regularity is observed, its representation is transferred to the deterministic rule region. The will model is thus not formed from a legitimate need for a place to store definite information about one's self; it has the singular character of being forced into the model, willy-nilly, by formal but essentially content-free ideas of what the model must contain.

9 Conclusion

When intelligent machines are constructed, we should not be surprised to find them as confused and as stubborn as are men in their convictions about mind-matter, consciousness, free will, and the like. For all such questions are pointed at explaining the complicated interactions between parts of the self-model. A man's or a machine's strength of conviction about such things tells us nothing about the man or about the machine except what it tells us about his model of himself.

==========