MATTER, MIND AND MODELS

Marvin Minsky
Massachusetts Institute of Technology
Cambridge, Massachusetts

INTRODUCTION

This paper 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 (1) because the question is ultimately meaningless or at least unanswerable, but it may also be (2) because an adequate answer requires a powerful analytical apparatus. My view is that many important questions about the relation between mind and brain are of this latter 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. In this paper we suggest a theory of why introspection does not give clear answers to these questions. The paper does not go very far toward finding technical solutions to the questions, but there is probably some value in finding at least a clear explanation of why we are confused.

KNOWLEDGE AND MODELS

If a creature can answer a question about a hypothetical experiment, without actually performing that experiment, then he 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 the encoded description of the world situation described by the question.

We use a term *model* in the sense:

> *To an observer* **B**, *and object* $\mathbf{A}^*$ *is a model of an object* **A** *to the extent that* **B** *can use* $\mathbf{A}^*$ *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 $\mathbf{A}^*$. If **A** is the world, questions for **A** are experiments. $\mathbf{A}^*$ is a *good* model of **A**, in **B**'s view, to the extent that $\mathbf{A}^*$'s answers agree with **A**'s, on the whole, over those 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, $\mathbf{W}^*$, inside of **M**. It would be most convenient if we could discern physically within **M** two separate regions $\mathbf{W}^*$ and $\mathbf{M} - \mathbf{W}^*$ such that $\mathbf{W}^*$ *really contains the knowledge* and $\mathbf{M} - \mathbf{W}^*$ contains only general-purpose machinery for coding questions, decoding answers, and general administrative work. However, one cannot really 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. For our purposes what is important 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.

MODELS OF MODELS

Questions about things in the world are answered by making statements about properties of corresponding structures in one's model $\mathbf{W}^*$ of the world. For simple mechanical, physical or geometric matters one can imagine, as did Craik[1],

machinery that does symbolic calculation but - when read through proper codings - has an apparently analog character. But what about broader questions about the nature of the world? These have to be treated (by **M**) not as questions to be answered *by* $\mathbf{W}^*$, but as questions to be answered by making general statements *about* $\mathbf{W}^*$. It $\mathbf{W}^*$ contains a model $\mathbf{M}^*$ of **M**, then $\mathbf{W}^{**}$ may contain a model $\mathbf{M}^{**}$ of $\mathbf{M}^*$. Indeed, this must be the case if **M** is to answer gefneral questions about himself. Ordinary questions about himself, e.g., how tall is he, are answered by $\mathbf{M}^*$, but very broad questions about his nature - what kind of a thing is he, etc., are answered, if at all, by descriptive statements made by $\mathbf{M}^{**}$ about $\mathbf{M}^*$.

The reader may be anxious, at this point, for more details about the relation between $\mathbf{W}^*$ and $\mathbf{W}^{**}$. How can he tell, for example, when a question is of the kind that requires reference to $\mathbf{W}^{**}$ rather than to $\mathbf{W}^*$. Is $\mathbf{W}^{**}$ a part of $\mathbf{W}^*$? (Certainly $\mathbf{W}^*$, like everything else, is part of **W**.) Unfortunately, I cannot supply these details yet, and expect serious problems in eventually clarifying them. I think we must envision $\mathbf{W}^{**}$ as including an interpretative mechanism that can make reference to $\mathbf{W}^*$ - using it as a sort of computer-program subroutine - to a certain depth of recursion. In this sense $\mathbf{W}^{**}$ must contain $\mathbf{W}^*$, but in another more straightforward sense $\mathbf{W}^*$ can contain $\mathbf{W}^{**}$. This suggests (1) that the notion *contained in* is not sufficiently sophisticated to describe the kinds of relations between parts of program-like processes and (2) 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 intermodel 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.

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. The other part is associated with things like goals, meanings, social matters and the like. This division of $\mathbf{W}^*$ carries through the representations of many things in $\mathbf{W}^*$, especially to **M** itself. Hence a man's model of himself is distinctly bipartite, one part concerning his body as a physical object, the other accounting for his social and psychological experience. When we see and object we account for its mechanical support and coherence (we are amazed at levitations) and we also account, in difference terms, for its teleology - who put it there for what purpose. When something moves we find either a simple force or a purpose - rarely both - in ordinary common-sense explanation; the kind that concerns us here.

Why this division, so richly represented in language and thought? We recognize that a person's $\mathbf{W}^*$ is not really two clearly disjoint parts but must have many overlapping, indistinctly-bounded models. The bipartite structure proposed here is only an approximation and we do not really want to suggest that the argument depends at all on a clear division into any particular number of parts.

The distinction between energetic explanations and informational (or symbolic) explanations is another aspect of the same general dimorphism. In one sphere, mechanical-geometric constraints are powerful - impenetrability in the arrangement of physical objects, conservation in the transformation, for instance. 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, e.g., in the forms of limitation 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 above sentence. But this involvement reflects not as much any synthesis of the two kinds of explanation, I am afraid, as it reflects the poverty of either the model for description of complicated situation.

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

THE CENTRAL ARGUMENT: BELIEF IN DUALISM

When a man is asked a general question about his own nature, he tries to give a general description of his model of himself. That is, the question will be answered by $\mathbf{M}^{**}$. To the extent that (1) $\mathbf{M}^*$ is divided as we have supposed and (2) that the man

has discovered this-that is, this fact is represented in $\mathbf{M}^{**}$, his reply will show this.

> *His statement (his belief) that he has a mind as well as 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 $\mathbf{M}^*$ 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.

A condense version of this argument was presented in Minsky[2].

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 all reach their goals. Still, such a success might have little effect on the overall form of out personal world-models. I will 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 the intermediate concepts are 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 accounting, *directly*, for the mental events of everyday life.

Thus synthesis by direct theoretical reduction is unlikely to have a large effect on the overall form of $\mathbf{W}^*$. The heuristic need for approximately self-contained subtheories is too strong to resist, in practical life and thought. Now one might hope for another kind of unity - parallel rather than hierarchical - in which the quasi-separate models are converted to basically similar structures and then merged by removal of redundancy, with coding for those differences that remain significant. It is doubtful that much can be done in this direction. The use of psychological explanations for physical processes runs exactly counter to the directions that have led to scientific progress. 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-procession models of mode 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, the incremental complexity is small because the model is already there as part of the *physical* part of $\mathbf{W}^*$.) It would be hard to give up such metaphors, even though they probably interfere with out 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 informational in character. Mental processes resemble more the kinds of processes found in computer programs - arbitrary symbol-associations, tree-like 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 $\mathbf{W}^*$.

Finally we should note that in a creature with high intelligence one can expect to find a well-developer 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 allocating shrewdly 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.

INTERPRETERS

The notion of *part* is more complicated for things like computer programs that for ordinary physical objects. A single conditional branch makes it possible for a program to behave, functionally, like two very different machines in difference 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 the 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 expanding all its structure on its description. In particular, the machine can contain an *interpretative* program which can use the internal description to calculate what the machine would do under some hypothetical circumstance. Similarly, while it is impossible for a machine or mind to analyze, from moment to moment precise 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 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 operations play an important role in out 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 interpretatively rather than executing compiled programs. While interpreting, one has the opportunity of examining the next step in the task before doing it.

FREE WILL

If one thoroughly understands a machine or a program one finds no urge to attribute *volition* to it. If one does not understand it so well, one must supply a complete 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 $\mathbf{M}^{**}$? If one asks how one's mind works, one notices areas where it is (perhaps incorrectly) understood - that is, where one recognizes rules. One sees other areas where one lacks rules. One can fill in by postulating chance or random activity. But this too, by another route, exposes the self to the indignity of remote control. We resolve this unpleasant form of $\mathbf{M}^{**}$ 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. This is because whenever a regularity is observed, its representation is transferred to the deterministic rule regions. The will-model is thus not formed so much from a 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.

CONCLUSION

When intelligent machines are constructed, we should not be surprised to find them as confused and as stubborn as men on 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 world, or about the man, except for what it tells us about his model of himself.

The gross divisions of our models probably have much heuristic value to us. Indeed we identify (in children) some stages in delineating the distinctions

between these models as associated with growth of intelligence. The distinctions could be abandoned only at great cost - in everyday practice. That is why, even if one accepts the conclusions of this essay, he is unlikely to note any serious effect on his way of thinking about most things.

I am indebted to S. Papert for several ideas in this essay.

References

[1] K.J.W. Craik, *The Nature of Explanation, Cambridge, 1952*

[2] M.L.Minsky, *Steps Toward Artificial Intelligence*