Model and Reality

Gary McGath

Revised edition, 2011

The original edition was self-published in 1988.

Contents

Introduction to the 2011 edition **4**

Acknowledgements, 1988 ..... **5**

Introduction **6**

I. Models and Black Boxes **9**

II. Modeling and Understanding **21**

"Scientific" Understanding 25

III. Mathematics and Logic **28**

The Role of Logic 29

Logic in Reality 32

The Basis of Logic 34

IV. Probability **37**

V. Gödel and the Unknowable **46**

Gödel's Theorem 48

The Barber's Razor 50

VI. The Quantum Universe **56**

The Permanently Unknowable 61

Common Sense and Quantum Reality 64

VII. That Which is not Seen **66**

Model Without Reality 70

VIII. The Faculty of Awareness **72**

IX. The Issue of Free Will **78**

Choice vs. Indeterminacy 80

Choice and Causality 81

X. Information-Based Models of the Mind **86**

Knowledge and Information....86
Minds and Souls....88
Turing on Consciousness....90
Can Science Deal with the Mind? 91
Reasons for Distinguishing Minds and Models....92

XI. The Turing Test....96
Machines that Imitate People....96
Turing's Cat....97
The Contrary Views....99
Making Machines Learn....103
Final Comments....104

XII. Imaginary Dialogues....106

XIII. Artificial Intelligence....111
Frames and Scripts....114
Top-Down Descriptions vs. Conceptual Knowledge....115
Mind and Metaphor....116
Language Issues....118
Strengths and Weaknesses of AI....120

XIV. Can Minds Be Modeled? 123
Hardware Suitability....129
The Assumptions of AI....129
Conclusions....132

XV. Science and Mysticism....134
The Indeterminate Cat....137
Science vs. Experience? 138
Healing the Breach....139

Bibliography....141

Introduction to the 2011 edition

This is a cleaned-up electronic edition of a work I wrote and self-published in 1988. I haven't tried to do any serious rewriting; there are many issues in this book which I haven't given detailed thought in a long time, so whatever flaws there may be in what I wrote then, I couldn't improve it now without a lot of work. I've simply gone through the text and corrected the most obvious errors, typographical and otherwise, and made some stylistic fixes. Once or twice I've found something that made no sense and tried to fix it up. Some people have asked from time to time to make the book available again, and I think it's still of some value.

This edition wouldn't have been possible without Martin Johansen, who prodded me to make it available and scanned the original book so I could edit out the typos.

Acknowledgements, 1988

I would like to thank the following people for their help in preparing this book:

Steve Wright, for helping me to make sense out of quantum physics, and for a long association in which he has helped me to develop a disciplined writing style and to understand many of the ideas that led to the writing of this book.

Ann Temlak, for being a friend when I needed one most.

Leonard Peikoff, whom I have never met, but without whose taped lecture courses in philosophy I could never have started this work.

Greg Swann, for many kilobytes of discussion, comment, and encouragement to "press on regardless."

David Kelley, for his valuable criticism of my manuscript.

Thanks are also due to Ellen Spertus, Lisa Jungherr, Warren Ross, Erich Veyhl, and Anna Franco, for comments on the manuscript as well as help in formulating the ideas that went into it. Many other people connected with Ergo and MIT's Objectivist Study Group also helped to create the kind of stimulating intellectual atmosphere which encouraged me to develop these ideas.

Also, without retracting a word of the verbal fusillade I have launched on Joseph Weizenbaum in this book, I must in justice acknowledge his characterization of the concept of "model" as one of the indispensable bases of this volume.

Finally, none of this book would have been possible without the accomplishments of one of the most brilliant and most maligned philosophers of all time, whose work in epistemology forms the indispensable base of my writing here: the late Ayn Rand.

Introduction

Science and philosophy go hand in hand—or at least they should. Metaphysics tells us what is true of existence as a whole; it tells us what we can expect of everything we encounter, no matter what its particulars. Science then applies these general principles in order to learn what the particulars are.

Unfortunately, events have created a breach between philosophy and science, between knowledge of the whole and knowledge of the parts. Scientists are often disdainful of metaphysics; some philosophers disparage the world of mere reason, while others lose sight of the broad abstractions in an effort to be "analytic" and compatible with science.

This breach began in ancient Greece, when Plato presented the parable of the Cave. Imagine, he said, that people are bound in a cave so that they can only see one wall of it. On this wall, they see the shadows of objects behind them. Knowing nothing else, they assume these shadows to be reality; they do not realize that the true reality is the objects which they cannot see. This, Plato proposed, is the situation in which we all live; the objects which we see in the world are but a shadow or imperfect image of true reality. What is this reality? It is the world of abstractions, or forms. Specific people, for instance, are but flawed reflections of the Form of Man, which is the ideal archetype. Abstractions, Plato held, are not formed from specifics as we would normally suppose; rather, specific entities are generated from their abstract archetype.

People can, if they are so inclined, go around measuring and studying these reflections; but in the Platonic view, the highest purpose a mind can seek is study of true reality, of the Forms. This is done not through physical instruments, which are themselves merely reflections, but through contemplation. Thus began the breach between the philosopher and the scientist; while the scientist got his hands dirty and measured things which weren't really real, the philosopher ascended into his ivory tower to contemplate a true reality that had nothing to do with everyday life.

Centuries later, Kant strengthened the breach by presenting his own version of the two worlds. In his terms, they are the "noumenal" world of true reality, and the "phenomenal" world we observe. Kant's true reality is not even as accessible as Plato's; we can grasp only a few bits of it on faith. The phenomenal world is created out of the noumenal world through the categories our senses impose on it. We see things as having shape, size, color, and so on, not because they really are that way, but because we are built to see them that way.

Once again, the roles of the scientist and the philosopher become divorced. The scientist—the person who relies on observation, measurement, and systematic study—is relegated to the role of studying unreal appearances. The philosopher takes a leap of faith into reality and tells us about the few snatches of it that can be discovered.

Scientists have thus been faced with an unpleasant, and in fact unreal, choice: to accept reality or to accept what they see. Their answer has been the only one that a normal human being, with a mind not crippled by academic madness, could make: that reality is what we see. But in doing so, they have rejected philosophy, which means rejecting the idea that reality as such can be understood. The scientist studies the behavior of specific aspects of reality, but

distrusts explanations of what they are aspects of. He studies causal relationships, but is left without an account of causality.

I am not saying that philosophers have ignored these issues, but for the most part they have gone in one of two directions: they have studied reality as something apart from observation, or they have pursued a methodology for science without grounding it metaphysically in reality. (The word "metaphysics" is, perhaps, slightly unfortunate in that it can suggest a world beyond the physical; what it means is simply the principles which apply to all of reality as such.) Modern trends in the treatment of mathematics and logic are particularly strong indicators of this split; it is commonplace today to regard these two fields, which are the basis of so much of science, as arbitrary constructs not based on any factual considerations.

The result is that scientists have not fully grounded their work in reality. It is a philosophical statement to say that the objects and events we observe are real and not mere appearances, that the reality we observe goes on even when we aren't watching, that scientific study is a process of discovering reality and not one of creating it. Normally, we can say these things simply as a matter of common sense. But science has advanced at a tremendous rate in the twentieth century, and some of its discoveries have created "crises of confidence" in reality.

The best known of these crises is the one created by quantum physics. To the best of today's scientific knowledge, events on the scale defined by Planck's constant happen in a random way, with no possibility of precision of measurement or prediction beyond a certain point. This state of affairs has raised the question: *Are* these events anything in particular, or are they blurs in reality as well as in our measurement? Is there a reality which is independent of our observations, or does "reality" simply mean that which we see? And if the latter is true, is it our perceptions which create reality?

Another crisis, though one which has not usually been identified in these terms, arises from the creation of computers of ever-growing power. These machines can perform not only "computations" in the old-fashioned, arithmetical sense of the word, but can manipulate symbols, engage in logical deduction, and converse with people. In brief, they can engage in actions which, people had always assumed before, require thinking.

The concept of "thinking" itself occupies a vulnerable position in a world that deals only with observables. Thinking, the fact of consciousness, is something which cannot be laid open to observation by all. A person cannot observe others thinking, yet we generally conclude that others think. How do we do so? Largely by observing what they do. But if computers can do what people do, doesn't that suggest that they are also thinking? And if they are, does that mean that thinking is simply a process of appropriate manipulations of data? Does the concept of consciousness have any place in science, or must it be discarded for a mechanistic concept of thought which is built out of elements that we can all observe?

I had originally planned to write a book addressing only these last questions regarding the nature of thought and the possibility or impossibility of thinking computers. But in considering the subject, I realized that the error which lay behind the belief in mechanistic brains was only one aspect of a broader error, and that an adequate treatment of the fallacy of a thinking computer required a discussion of the way an essential tool of scientific methodology is so often misapplied.

We have to describe reality in some set of terms. We select some attributes of an object, process, or system to measure, and we devise appropriate ways to measure them. Whenever we do this, we have to disregard other attributes temporarily; there is a limit to how much complexity any mind or computer can handle at a time. We can characterize a human being, for example, as a unitary mass, as a system of organs, as a collection of interacting subatomic particles, as a processor of information, or as a being with thoughts and feelings. Each of these approaches is a true description of reality, and none contradicts the others; but each neglects some facts which the others take into account. Each of these characterizations lends itself to a scientific model of man, with an appropriate set of measurements taken as accurately as possible. Still other models may be possible, based on information which we do not yet know how to acquire.

Because each of these models is itself very complex when fully treated, it can be tempting to take one of them as a full description and neglect the facts which it does not deal with. In physics, this leads to treating the uncertainty in the quantum model as if it must be an irreducible fact of reality. In mathematics, it leads to regarding symbols as if they have a life of their own independent of the facts they were devised to describe. In computer science, it leads to treating a simulation of thought as if it were the real thing.

In its worst form, this error can lead to its own kind of supernaturalism, as descriptions become completely detached from they reality they are supposed to describe. Some students of quantum physics have proposed that it is the act of observation that collapses the wave function and thereby makes reality something in particular. This brings us full circle to a world of magic, in which the mind does not observe reality, but creates it. The scientist in this world joins the philosopher in the ivory tower and leaves the real world unattended. In doing so, he repeats Plato's error of regarding abstractions as more "real" than reality.

It is to point out and uproot the error of letting descriptions float free from facts, models be separated from reality, that I have written this book. It is not my purpose to reach any new scientific conclusions here, but to warn of dead ends that will make it impossible to look at the world in new ways and discover new facts. Science divorced from reality can generate a new wave of superstitions, one which can gain as much credence in the modern world as superstitions generated by priests did in an older world; the result in each case is a blow against rational thought which we must strive to avoid.

Other people will, I hope, expand on the areas where my knowledge has been incomplete, and will find that this book helps them to gain a new perspective on the issues in science which they are facing. If I can provide people with the groundwork from which to avoid errors which are currently common, and to see possibilities which those errors have obscured, then this book will have accomplished its purpose.

I. Models and Black Boxes

It is immensely satisfying to see things first-hand, with our own eyes. Nothing gives the same sense of reality, the sense of actually knowing that something is what we believe it to be. If we couldn't have any experiences of our own, if everything were presented to us through reports, studies, simulations, and recordings, we wouldn't have the least idea whether to believe them or not. We can't know if a source is reliable unless we have some way, however indirect, to check its accuracy; and ultimately this means checking against our own experience. If we were told, without having seen it, that a vehicle weighing tons could lift itself into the air and fly across the country, or that a machine weighing a few pounds could play a good game of chess, even the most convincing explanations would leave us in doubt; yet because we see these things happening, we accept them without question. That is the power of experience.

On the other hand, we would be severely handicapped if we could know *only* what we experienced directly. What makes us distinctively human is our ability to deal with higher levels of knowledge. When we read the following passage, we turn the symbols on paper into images in our minds, and we can visualize Mark Twain's White Elephant and even know a bit about its character, without ever having seen it in fact:

> Height, 19 feet; length from apex of forehead to insertion of tail, 26 feet; length of trunk, 16 feet; length of tail, 6 feet; total length, including trunk and tail, 48 feet; length of tusks, 9 1/2 feet; ears in keeping with these dimensions; footprint resembles the mark left when one upends a barrel in the snow; color of the elephant, a dull white; has a hole the size of a plate in each ear for the insertion of jewelry, and possesses the habit in a remarkable degree of squirting water upon spectators and of maltreating with his trunk not only such persons as he is acquainted with, but even entire strangers; limps slightly with his right hind leg, and has a small scar in his left armpit caused by a former boil; had on, when stolen, a castle containing seats for fifteen persons, and a gold-cloth saddle-blanket the size of an ordinary carpet.

Descriptions, abstractions, concepts, generalizations, theories—in brief, the power not just to perceive, but to think—these are what give power to our minds. Perception provides the raw material without which we would not have anything to think about; but conceptual thought is necessary in order to form new ideas, to communicate, to consider alternatives, and to discover general relationships.

Direct perception lets us deal with one object at a time. But when we try to think about more than one object, our abilities are rapidly overloaded unless we use a different approach. Most of us, for instance, deal with many people: family members, friends, employers, co-workers, store clerks, librarians, and on and on. If the only way we could know anything about them was to remember the particulars of each person as a separate image, we would soon be swamped. Any encounter with a stranger or any unusual situation with a familiar person, without the knowledge we have about people as a class, would force us to discover everything anew. Our knowledge of language, of commerce, of medicine all come from our ability to deal in abstractions.

The power of abstraction is most obviously necessary when we are dealing not with static objects, which we can reduce to images, but with processes of change. The division of a living cell, the life and death of an ant, the construction of a building, the political organization of a city, the pattern of the seasons, the geological changes of a continent; these are all patterns of development that result from complex combinations of factors. To understand what is going on, we must separate out the particulars, identify the combinations that each one makes, and establish the relationships among them.

Take a relatively simple example, such as the burning of a candle. On a simple view, there are three major elements in this process: the wax, the wick, and the oxygen in the atmosphere. The wax and the wick both combine with the oxygen to sustain the flame; the wick provides a second function as well, as a carrier for the melted wax to ride up and expose more of itself to the hot oxygen around it. On this level, we can see in principle how a candle works; take away or modify any of the three contributing elements, and the results would be different.

"But say with what degree of heat," counsels Robert Frost. "Talk Fahrenheit, talk Centigrade." To understand the flame well enough to predict its temperature and longevity, we must take other factors into account as well. The flame consumes oxygen and produces carbon dioxide, threatening to exhaust its own reaction. The increased heat expands the gases around the flame and generates an updraft, further affecting the interaction of the components. By the time we can answer Frost's question, we must have obtained a much more detailed understanding of the elements involved.

Even when we do that, we are taking a great deal for granted. Why does the wax interact with oxygen at high temperatures to produce even more heat? To answer this question, we must go to the atomic level and below, discovering the nature of the components in tremendously fine detail.

How do we do all this? How can we keep track of the bewilderingly complex body of knowledge which is required to understand fully something as simple as a candle? The answer to this question is the subject of the science of epistemology. Epistemology is, quite simply, the study of knowledge; it is the branch of philosophy that answers the question, "How do you know anything?"

Some of the best answers to this question have been provided by philosopher Ayn Rand. Rand is well known for her novels and her ethical and political theories; she is not so well known for her work in epistemology, but her contributions there are extremely valuable, often cutting through academic Gordian knots with startling insight. This is not to say, of course, that she did everything herself; her work comes from a tradition that we can trace back to Aristotle. Yet her formulations are among the best yet offered, and her ability to bring together seemingly unrelated issues is astonishing. In the course of this book, I will be referring to her work many times.

For Rand, the key issue in epistemology is concept formation. Concepts are the means by which we tie together more knowledge than we can hold in the form of images, establish general principles, and understand not just how particular things work, but how the world works.

A concept, by Rand's definition, is "a mental integration of two or more units possessing the same distinguishing characteristic(s), with their particular measurements omitted." The

concept of "candle," for instance, includes all burning devices consisting of a waxy substance and a wick, regardless of their particular color, shape, and size.

Concepts allow us to think of classes in much the same way that we think of single items. In thinking of candles, we can start by discovering the behavior of one candle, then comparing it to the behavior of candles made of different materials to discover what is in common to all of them. We see that the tip of the wick burns off as the wax fails to reach it, that the melted wax accumulates in a pool with some runoff down the sides, and so on. This gives us a general understanding that we can apply to any particular candle.

A concept is not just a definition; everything that one knows about the class is implicitly contained in the concept. Nor is it an amorphous cluster of referents that just happen to be called by the same name; while the edges of a concept may blur in a person's mind, concepts would serve no purpose if people could not identify facts about the entire class identified by a concept. Often several closely related concepts are identified by the same word, leading to confusion; but mental chaos would be the result if people's concepts were just mental grab bags. (This is verified by considering the condition of people whose concepts *are* grab bags.)

Omission of measurements is a key element in Rand's characterization of concepts. Philosophers of the nominalist persuasion argue that elements of a class really have no common characteristics. No two people, for instance, have exactly the same weight, height, intelligence, or amount of almost any other characteristic. This view leads to the conclusion that definition by essential characteristics is really impossible, that the best which is possible is to treat concepts as clusters of more or less similar particulars. The nominalist view is very troublesome to science; if a "mammal" or a "star" is merely something which is similar to other things called mammals or stars, what confidence can we place in universal statements about them?

Rand avoids this difficulty by regarding distinguishing characteristics as a matter of the *kind* of attributes which a member of a class possesses, apart from its particular measurement. This is the way which people use concepts in daily life. A piano with eighty keys rather than eighty-eight would still be a piano, in spite of the quantitative difference; but an instrument which plucked its strings rather than striking them with hammers would be a harpsichord or some other instrument. This is a common-sense point, but Rand's identification of the role of omitting measurements is vital to understanding why concepts work.

Only the fundamental characteristics of the members of a class (that is, those on which the greatest number of other characteristics depend) playa role in its definition; but all the derivative characteristics playa role in the concept itself. The possession of fingernails, for instance, is not one of the defining characteristics of the human species, but it is a consequence of our essential biology; thus, we are justified in expecting a human being to have fingernails unless extraordinary circumstances intervene. Fingernails form part of the *concept* of a human being, though not part of the concept's *definition*. Recognizing this fact is important to recognizing the value of concepts; a concept has much more content than just its definition.

The theory of concepts forms the basis for understanding the use of models in science. The key matter to bridge in going to models is the fact that we do want to deal with exact measurements, not just with kinds of measurements.

Returning to the example of the candle, suppose we want to learn "with what degree of heat" a particular candle burns. We are no longer omitting all the measurements; we now need to know exactly what kind of wax and wick we are dealing with. Does this mean we have to give up concepts and go back to dealing with one image at a time? Not at all. We can put the needed measurements back into the concept; but we put them back not as specific values, but as *variables* that take on specific values for anyone candle, while serving to account for all candles. When we do this, we have a description that accounts for all candles, yet can be tailored to show just how anyone candle will behave. We have a *model*.

A "model," as the term is used in this book, means a description of a set of relationships in mathematical or quantitative terms, such that variables within the model correspond to varying events or conditions in the process being described, or variable relationships among the characteristics of the entities being modeled. Models are tremendously useful tools of understanding. Once created, an accurate model provides even those who have not studied the process with a means of predicting its course and, perhaps, the results of alterations to that course.

Paradoxically, a large part of a model's usefulness to understanding is that it is possible to apply a model without understanding it. Concepts, the basic elements of thought, are difficult things to work with; there is no quick and automatic way to convey a concept from one mind to another. Definitions can be passed on, but conveying the full content of a concept is much more difficult. A model, on the other hand, can be passed on to anyone who can handle its mathematics. When using concepts, we omit all specifics, but stand ready to put the specifics back in when they are needed. With a model, we treat some specifics as constants, others as variables, and the rest as matters of indifference. Changing those selections implies changing the model. A concept could be regarded as active thought, and a model as frozen thought.

However—and this must be kept in mind to resolve the preceding paradox—doing anything more with a model than applying it by rote requires reconstructing the thought process which was frozen into the model. Even knowing what the model actually describes requires the amount of thought necessary to understand the process being described. When I say that applying a model "does not require any understanding," I am referring strictly to using it to churn out answers.

This lack of any need to understand is also the limitation to models. If a model fails to emulate the behavior of its referent process, it will not contain any explanation of why it went wrong. To correct it, the model's author must go back to the original concepts, and the particulars contained therein, and discover what measurements have been neglected or treated incorrectly.

A model symbolically describes a process without entailing any actual knowledge of the process. The process can be reconstructed only by an act of thought that goes beyond the model.

A model usually treats a whole class of processes, assigning different values to variables in order to cover the differences among them. But some measurements remain omitted in the model; perhaps these are normally insignificant but affect the process significantly in certain cases. (For instance, we can normally ignore the temperature of billiard balls in considering how they will bounce; but if they get so hot that they start to melt, the temperature variable becomes significant.) Also, the model may be valid only over certain ranges of the variables that are considered. Newton's laws of motion provide the basis for a model of force and

acceleration which is essentially accurate for everyday experiences; but when the velocity variable begins to approach the speed of light, the model breaks down.

This is not to say that models aren't useful for understanding; for complex processes, they can be essential. Try thinking about a computer in terms of the flow of electrons through its chips, or imagine building a house without an architectural plan. But understanding comes from relating the elements of the model to the process. What the model lets us do is consider the characteristics of one piece at a time. The simplification is essential to the process. As Weizenbaum notes, "The aim of a model is, of course, precisely not to reproduce reality in all its complexity. It is rather to capture in a vivid, often formal, way what is essential to understanding some aspect of its structure or behavior."[1]

We all use informal models in ordinary life. A very simple example of modeling is tracing a planned route on a map. (A map isn't itself a model in the sense the term is used here, since it is a purely static description. However, the use of a map to trace out and measure a trip is an act of modeling.) The model applies to all the applicable trips that might be made, whether by car, bus, or truck, whether today or next month. By assigning values to certain variables, such as speed and fuel consumption, we can tell how long a trip will take, and how much gas it will consume. If the model is wrong—for instance, if the road is closed for construction—then the predictions will be wrong, and nothing in the model itself will let us know that fact.

A more complex example of modeling is a board wargame. Fighting units are modeled by cardboard counters described by a few numbers; the advance of troops is modeled by pushing these counters over a grid according to the rules of the game. To the extent that the model is accurate, and the players follow the same strategies as the generals who actually directed the battle, the game will tend to follow the course of the battle or follow alternative paths that could easily have happened with minor shifts of luck. There are, of course, very obvious deficiencies in a game's description of the components of the battle, yet statistically, the description of the battle in terms of formal rules may come close to showing the actual course of events.

This example points out an important characteristic of models: they are generally applicable only on one level. This level is defined by which measurements are included in the model and which are left out. The variables selected for a particular model may be useless for describing the fine details of a process. A wargame normally (and fortunately) does not deal with the grisly details of war, with the count of broken legs, shattered spines, and so on. It presents casualties as being taken in large chunks, rather than one at a time as they actually occur. Following a route on a map does not say anything about the potholes that may be encountered, beyond a general notation that some roads may be in better condition than others.

At broader levels than a model is intended to cover, its detail may be sufficient, but its scope may be too limited. This can be an issue of choosing variables that are impractical to calculate over such a large scale; it can also be an issue of the range over which the characterization of those variables applies. If a model describes traffic flow in a city, it is likely to be of little applicability to a rural highway, and hence to describing traffic flow in a whole state or nation. A model of the human body on the level of individual cells would be impractical, or at best very inefficient, for characterizing the process of walking.

A model may even be completely wrong in its underlying assumptions about the process involved, yet be accurate in describing and predicting results. The ancient Greeks, without Newton's laws to guide them, built up a complex model of the heavens based on spheres

within spheres. This model had nothing to do with the causal factor of gravitational attraction that actually controls the planets' motions, but it did describe their motions with considerable accuracy. Assuming that a model that produces correct results is necessarily correct as a description is a dangerous practice. For one thing, its predictions might not be correct next time! (Stock market modelers, take note.) In William James's terms, such a model is "instrumentally true," since it leads to satisfactory results, but "instrumental truth" can turn into "instrumental falsehood" without warning, when the mistaken assumptions behind the model break down.

This point has to be emphasized; the accuracy of a model in terms of predictive value and its correctness in terms of having elements that correspond to causal factors in the process being modeled are two different issues. We might call these, respectively, "predictive accuracy" and "structural correspondence." They are certainly related; the fact that a model predicts accurately suggests that its structure does correspond to the process, and modeling a process with due attention to the factors that actually shape the process is likely to produce accurate results. But the two don't always go together. It may be that the measurements used in a model are derivative or parasitic effects, which are closely associated with the underlying causes. In this case, the model may prove to be accurate for a certain range of cases, yet fail later on when the factors being modeled cease to correspond to the more basic causes. In this case, the model is predictively accurate (at least for a while), but deficient in its structural correspondence. It is also possible for the measurements which have been left out of a model to be important to the behavior of the process, so that in spite of a broad structural correspondence, it does not predict accurately. Since any model has to omit some factors, this danger always exists until the model is actually tested.

In a day in which pragmatism is so highly valued, not much has to be said to point out the need for a model to make correct predictions. The importance of correspondence between the causal factors of a process and the elements of a model, though, needs a little comment. Isn't a model perfectly satisfactory, one might ask, as long as it gives adequate predictions? One reason this correspondence is important is that without it, we don't know whether a model will continue to be accurate for any length of time. If we can't say why a model produces accurate results, we really can't have any confidence in it.

A second concern is that we don't use models just to get predictions, but also to enhance our understanding. A model provides a simplified picture of a complex phenomenon, and its study can help us to understand the basic operations of that phenomenon. If the way the model works doesn't have a close correspondence to the way the process works, the people studying it are likely to develop mistaken or confused ideas about the process itself.

Both factors come into play if we want to change some aspect of a model. Imagine a model which describes the financial dynamics of a business; an executive might want to change some aspects of it to find out what would happen if the company changed its organization. Suppose one of the factors the model relies on is the number of bathrooms the company owns. This number is certainly likely to correlate with the company's growth. However, if the executive uses that number as a predictor and considers the consequences of building a plant with fewer but larger bathrooms, the model would lead him to unwarranted conclusions. Moreover, he might decide from studying the model that the way to make the company successful is to build lots of bathrooms.

If this seems far-fetched, consider the fact that an economic model which correlates national economic strength with consumer spending has had a major impact on economic

policy. Under this model, national leaders have decided that saving was bad, since saving more means spending less, and have actively sought to discourage investment and savings. The effects of an ill-conceived model can be far-reaching.

Lack of structural correspondence is not necessarily bad, but it does impose limits on the model that have to be kept in mind. For instance, one of the simplest possible models of the growth of a population is an exponential curve. This model doesn't have any structure to speak of, yet it can give reasonable predictions for the growth of a population under certain special conditions (effectively unlimited room for growth and food supply, and no significant variations in external influences). Change any of the conditions, though, and the model offers no guidance.

These considerations can be summed up briefly: Predictive accuracy and structural correspondence in a model are correlated but distinct elements. If a model is predictively accurate, that suggests that its structure is valid, but does not guarantee it; in such a case, the predictive accuracy may disappear unexpectedly when circumstances change. A structurally accurate model will tend to be accurate in prediction, but if it omits factors which are important, its predictions may be wildly off. Neither theories unchecked by experience nor superficial descriptions of results will result in a reliable model.

There are two basic kinds of models. The models this chapter has discusses so far have mostly been models based upon a process that existed before the model was formulated. But a model can also be a description of a process that is to be implemented based on the model. In this case, the model represents the intended behavior of the process. We might call the first kind of model, the one which reduces an existing process to mathematics, a reductive model, and the second kind, the one from which a process is to be constructed, a constructive model.

A reductive model can be the basis for constructing another process, which imitates or simulates the original process. The new process might even be valuable in its own right. Computer "clones" are a noteworthy example. Computer manufacturers, wanting to take advantage of software for popular computers, reduce those computers to a model in an abstract enough way to avoid legal problems, then they build their own computers based on those models. The new computer is a simulation of the old computer; yet it has obvious utility beyond finding out what the original computer would do. In such a case, a reductive model has turned into a constructive model.

But is the clone really the equivalent of the original, or even in the same class with it? In the case of computers, this question can be stated in fairly definite terms: Will the clone run all the programs the original will run? Will it run them as fast? Is it equivalent in visual and tactile qualities? In other cases, though, the question could be more complex.

In particular: When we implement a model of thinking, is the implementation an instance of thought, or just a simulation of thought? Does it depend on how good the model is? For now I simply leave these questions for consideration. But the conclusion is not foregone. The principle here is that processes with a common model are not necessarily the same kind of process. A simulation is not the same kind of activity as the process it simulates, even if it is just as complicated. To address the issue of thinking, we will have to consider what kind of process thinking is.

When discussing implementation of models, the computer inevitably comes to mind. If a model is frozen thought, a computer is the perfect embodiment of frozen thought. Models

abound in computer programs. A spreadsheet models the effect of changes in quantities of money or other units that concern the user. A chess program models the movement of pieces on a chessboard. A statistical package can be used to model any number of real world processes. A word processor or a typography program models the layout of text on a printed page.

Inside the program, even more models are to be found. Programs with "smart" output behavior model the appearance of their printouts or screen displays in order to justify text and neatly arrange lines. Driver programs—the programs that control the detailed behavior of peripheral devices—model the devices in terms of registers, data paths, tracks, and sectors. Peripheral devices in turn often have their own built-in firmware programs, which model the physical actions of the device.

In another sense, all programming is an act of constructive modeling. A programmer writes down instructions in Pascal, Cobol, or assembly language; the instructions are the description of a model of how the computer will act when it runs the program. If the program is written in a language which can run on several different machines, then the single model (program) characterizes several different processes, which are implementations of the model.

The reason there are so many different programming languages is that people have so many different ideas about how to model a computer's operations. Assembly language allows the creation of a very detailed model, which corresponds closely to the physical operations of the machine. A higher-level language, such as C or Pascal, provides a more abstract model, which omits specifics on the machine level but deals with issues which assembly language leaves to the applications programmer. Such a language provides two benefits at the cost of efficiency and detailed control over the machine: it makes programs independent of any particular machine and lets the programmer ignore the details of implementation.

Most programming languages of today model the machine's activity as a series of sequential operations on linearly addressed data. This is the Von Neumann model, which reflects the way most computers to date have operated. Other models, though, may be better suited for many programming tasks. Object-oriented languages, such as Smalltalk, provide a model of a set of data objects that process messages which are passed from object to object. Logic-oriented languages, such as Prolog, provide a model of a set of logical rules from which inferences can be made. On the hardware level, multiple processors and associative memory are advantageous for many tasks.

Every model is based on some assumption about the underlying methodology of its implementation. We could call this basic set of abilities, which will be used to interpret the model, its *interpretive context*. When a model is explained to a person the description can be informal, because people can be expected to understand much of what is not made explicit and can ask questions about what they do not understand.

Programming languages provide explicit interpretive contexts for models, as well as being models themselves for the operation of the computer. Some interpretive contexts are better suited for certain kinds of models than other contexts are. Modeling the act of speaking English is possible in Fortran, but hardly desirable if there are alternatives.

The status of languages as both models in themselves and interpretive contexts for other models is an example of the layering of models, a practice which saves huge amounts of effort. Any model of a general type of process can be part of the context for interpreting a

model of a process which depends on or is generated by the general process. The idea of subroutines in programming languages, as well as that of microcoded instruction sets in computer hardware, illustrate the way in which related models can be combined and can interact with one another.

Modeling and computers can hardly be separated in our minds today. Yet for all their versatility, computers do not add anything to the concept of modeling; they only extend its practical range. In principle, though, the most powerful computer cannot do anything that could not be done with paper and pencil, or with a very simple computer with a sufficient amount of mass storage.

Establishing this fact leads us into a new and intriguing convolution: computers, the most powerful tools for modeling, can themselves be modeled. Computer pioneer Alan M. Turing discovered a remarkably simple model, which contains the full capabilities of any computer from a Sinclair ZX-80 to a Cray-l. This model is known today as the Turing machine.

A Turing machine is a device that operates on a tape of infinite length. The tape is divided into cells, each of which can hold one bit (a binary digit, either a 1 or a 0). The machine is capable of reading or writing the bit which is currently under the tape head, and of moving the tape in either direction. The machine's program is defined in terms of states which the machine can be in. In a given state, the machine will perform one set of actions if it reads a "1" bit from the tape, and another set of actions if it reads a "0" bit. An action consists of one or more of the following: writing a new value to the bit under the head, moving the tape forward or backward by one cell, and going to a new state. The number of states in a given program is required to be finite.

Turing showed that no realistic extension to this machine would extend its power. Adding more tapes or "processors," or allowing more information to be written in one cell, could reduce the number of states necessary to implement a program, but it did not allow any new programs to be introduced that were not equivalent to programs in the simpler machine.

This model, simple as it is, actually presents a theoretical upper limit on the power of a computer. There is no such thing as an infinite tape, nor do we have infinite time to wait for a slow machine to perform complex operations. What the model tells us is that if something can be done on a Turing machine, it can eventually be done on a real computer with enough storage, and that if something can't be done on a Turing machine, we might as well give up trying to do it on a computer. (This leads to the fascinating question of whether there are problems which can't be solved by a Turing machine but can be solved by other means; but that is a subject for later discussion.) Although building a quasi-Turing machine (with a finite storage device) is a popular lab project for students, such machines are more often described on paper than built. In fact, the operations of any computer program can be carried out on paper, given enough paper and time.

There is nothing distinctive about a particular computer's capabilities, other than differences of speed and capacity in performing equivalent operations; what is distinctive in kind is the software it runs. A model, therefore, can be considered as a candidate for computer implementation without dealing with the specific characteristics of a given computer. (The form of the user interface may depend on the particular computer; for instance, one machine may have graphics hardware that another lacks. However, this is only a difference in methods of presenting the same information—something which may be very important to the user, but

is not relevant for the current discussion.) The implementation of a model neither adds to nor detracts from its validity, and any two correct implementations of a model are equivalent.

This may seem to contradict the earlier principle that processes with the same model are not necessarily the same kind of process. But the two principles take two different points of reference. All implementations are equivalent from the point of view of the model; that is, any differences among them that do not violate the model make no difference to their correctness as implementations. On the other hand, if the whole process is what is important, the fact that two different processes have a common model does not wipe out their differences or make them insignificant.

The greatest danger in using models as a tool for thinking is to consider only the model's viewpoint and start to think that everything which is true of the model is true of the process. One of the worst forms of this error is to suppose that artificial limitations in the model are also limitations in the actual process. This error may be called the "black box fallacy."

The black box fallacy, in general terms, is the treatment of a particular mode of understanding as the equivalent of the entity or process which one is attempting to understand. Any given approach to understanding something entails neglecting some of its particulars, either because of lack of knowledge or lack of capacity for detail. Failing to recognize this fact can lead to treating the object of understanding as a black box—as a locus of behavior which is completely explained by a particular description of its behavior.

Any model, or in general any method of understanding, results from observations and generalizations arising in a given context. The black box fallacy neglects the fact of context and assumes that there is nothing more to be understood. At the same time, it neglects the fact that a given method of understanding may involve cognitive artifacts which are not actually a part of the object under consideration.

An example of the difference between models and the reality behind them is found in the fact no model can have "free will." Random numbers might be a part of it, but not free choice originating in the model itself. Hence, if we construct a model of human behavior, we cannot include free will in the model. But it would be a serious mistake to conclude from that fact that people do not have free will.

On the other side of the coin, a model might include elements which are indeterminate and must be generated by random numbers, simply because there is no known relationship between the process being modeled and the values of these elements. It would be an equally serious mistake to conclude by looking at a model of this kind that these elements are forever undetermined and random, that they do have free will. In either case, the basic causes that underlie the model fall outside the model itself. A reductive model is not in general equivalent to the process it models; limitations which apply to the model do not necessarily apply to the process.

An even worse error is to suppose that a model, or an arbitrary implementation of it, is an instance of the process from which it was reduced. The most blatant example of this error is the voodoo practice of sticking pins into a doll as if that would inflict injury on an enemy. The field of artificial intelligence frequently makes this error in supposing that an adequate implementation of a model of thought would be an instance of a model of thought. This instance of the general fallacy will be the subject of later chapters in this book; but first there is a great deal of groundwork to do.

There is another danger to watch for as well. Blind rebellion against an error often leads to acceptance of a converse form of the same error; and this can apply to the black box fallacy. Some people have come to realize that no model is a full description of that which it represents, and have reached one or both of the following conclusions: (1) that models are anti-reality, and (2) that models shouldn't be expected to have any relationship to reality. The first of these leads to blind empiricism, to insistence on direct knowledge without the aid which models can provide; the second leads to blind rationalism, to dealing with mathematical constructs that are not expected to describe anything real.

Because models are so important in scientific and engineering thought, questions about whether they are being used properly or fallaciously are vital ones. The fallacies involved may not be obvious—educated people are unlikely to confuse open voodoo with science—but a fallacy which goes unchallenged becomes harder to recognize as a fallacy as its use comes to seem more familiar. Hence, errors can accumulate where they are not corrected. By examining the basic issues involved, we can identify those errors and avoid them, at the same time gaining more confidence in the proper use of modeling.

* * *

Where do we go from here? First, a closer look at the relationship between modeling and understanding is necessary in order to avoid the mental traps that so often result from thinking in terms of models alone. After this, there are three broad areas to cover.

The first of these areas, covered in Chapters 3 through 5, is the use of formalisms in mathematics and logic. Formal systems can be regarded as pure black boxes, as models which are completely purged of any reference to actual phenomena. The use of formalisms has grown very popular in the twentieth century, and logic is often regarded as little more than a formalism -that is to say, as having little or nothing to do with reality. This approach is very dangerous in its implications, since it splits our most important tools of thought apart from the things that we think about. In particular, Gödel's theorem tells us that any formal system of sufficient power must be either incomplete or inconsistent, and this conclusion has often been taken as a limitation on human understanding. Whether this conclusion is valid, though, requires a closer look into the nature and significance of the formalisms involved.

The second area is one which I approach with trepidation, but which cannot be avoided in considering the issue of models: the study of quantum physics. In the Copenhagen interpretation of quantum mechanics, the positivist principle that only the observed is real leads, paradoxically, to the conclusion that mathematical formalisms which have no apparent interpretation in the physical world are, nonetheless, explanations of the fundamental nature of that world. The key to the paradox is that positivism does not admit an underlying, "metaphysical" reality which exists when one's back is turned; the consequence is that only mathematical descriptions of the relationships among observations are admitted as scientific principles. Schrödinger's Cat provides the reductio ad absurdum of this approach in a literal black box, as a cat whose condition of life or death cannot be observed or deduced is regarded as occupying a twilight zone encompassing both states. This state of affairs, and some general thoughts on how to avoid it, are covered in Chapters 6 and 7.

Chapters 8 through 14 deal with a question which is of great interest today: whether minds comparable to those of humans can be created by computational processes. The issue here is, first of all, just what the mind is, and second, what the relationship is between a mind and a model of a mind. My contention is that a mind is not the same as its model, and that a

computational model of thought, no matter how effective it may be in answering questions and conducting conversations, is nonetheless a simulation of thought and not itself a thought process. I will give special consideration to Alan Turing's arguments in favor of saying that computers will someday think (or—an important distinction—be generally regarded as thinking), and Hubert Dreyfus's contention that human thought is not susceptible to computer simulation.

Having covered all of this ground, I will look in Chapter 15 at the dangers that the black box fallacy, and other fallacies in the applications of models, pose to scientific thinking. Misinterpretations of Gödel's theorem or of quantum physics can lead to the conclusion that reality is unknowable or unreal; the belief that computers can think can come full circle to the conclusion that people only manipulate symbols, thus leaving reality out of the picture. Hence, by the end of this book, I hope to have shown that avoiding the fallacies related to models is a matter of vital concern to every scientist, and indeed to every person.

---

1 Weizenbaum, p. 149.

II. Modeling and Understanding

It is often held that the essence of understanding is modeling. According to Johnson-Laird, "The psychological core of understanding, I shall assume, consists in your having a 'working model' of the phenomenon in your mind."[1] Is this claim valid? Is understanding a process fundamentally a matter of modeling it? Are there other factors which are equally fundamental or more so?

The type of model which is intended by Johnson-Laird and others who hold this theory is not precisely equivalent to models as I have discussed them here; but there is an essential similarity that justifies the use of the same word. The idea of a model in the mind implies that the mind does not directly grasp the outside world, but rather constructs an internal description of external entities, which is always in some way a simplification of their actual nature. The term "working model" is somewhat confusing, since it suggests a functional but simplified prototype or reduction of a device, but this is not what Johnson-Laird is discussing; the term "dynamic image" might be more exact.

Understanding refers to a compounding of knowledge, an ability to grasp relationships and deal with them in various ways. Simply "knowing" can refer to awareness of isolated facts; "understanding" implies an ability to look at the facts in various ways, to see cause-and-effect relationships, to consider alternatives to what is immediately seen.

In many cases, understanding is based on a model in an obvious way. Johnson-Laird cites the models that different persons might have of a television set.

> Your model of a television set may contain only the idea of a box that displays moving pictures with accompanying sound. Alternatively, it may embody the notion of a cathode-ray tube firing electrons at a screen, with the beam scanning across the screen in a raster controlled by a varying electro-magnetic field, and so on ... A person who repairs television sets is likely to have a more comprehensive model of them than someone who can only operate one. A circuit designer is likely to have a still richer model.

In each case, the model is cast in some basic set of terms. The more someone understands a process, the further back he is able to push those basic terms; but no one can fully understand everything. Whether the basic terms are pictures and a dial, or semiconductor junctions and electrons, they must come in as a starting point. Further analysis, though possible in principle, ceases to be personally worthwhile at some point.

Moreover, the same person does not always use the same model. When the circuit designer goes home to watch TV, he does not deal with his own set in terms of its circuitry; the model he uses is the simple user's one, of a box that produces certain pictures and sounds depending on how he sets the controls and what is being transmitted.

However, possessing a model does not by itself constitute understanding. It is possible for a person to manipulate a description of a process and give correct answers about it without knowing what he is talking about. This condition is often found among students who learn from the kind of textbook that gives formula after formula with no explanation. The student's model of gravity may tell him that gravity accelerates all objects equally, regardless of their

mass, yet he may not know that if he drops a two-pound rock and a one-pound rock from the same height at the same time, they will hit the ground at the same time.

A person who thinks in this way has failed to go beyond black boxes; for him, the model is all that there is. When additional considerations appear in the real world, they get in the way of his understanding, since they make it more difficult for him to adapt his model the particular circumstances.

For understanding to exist, another element must be present in the mind prior to using the model; it is necessary to identify, to know, what the referents of the models are, and to be able to use that knowledge in conjunction with the model. A model can be used apart from knowledge of its referents, but such use is simply manipulation, not understanding.

The issue here isn't the extent of a person's knowledge, but the extent to which the model that he uses exceeds his knowledge of its referents. The person who knows nothing of the internals of a TV set can still understand it on the level of a simple model if he knows that his operation of it results in the presentation of certain images and sounds; the person who can calculate all the voltages in a circuit diagram doesn't understand it if he isn't aware that the diagram represents an actual device or doesn't know how to relate the two.

It could be argued that a person can know what a model refers to and yet still not understand it in the sense used here. The student mentioned earlier, for instance, probably knows that rocks are objects on which gravity acts, yet he still does not make the connection between them and the model of gravitational attraction. What is he lacking? Colloquially, we would say that his knowledge is not "real" to him. He can recite statements, but he does not grasp their implications.

More precisely, his knowledge is not *integrated*. He is aware of fact A and fact B, as well as the logical principles which allow a connection between the two, but he has great difficulty making the connection. Once it is pointed out to him, he may kick himself for missing such an obvious relationship, but without assistance he cannot bridge the gap. The ability to integrate one's knowledge is a general skill, but one which is essential to the understanding of any particular process or phenomenon.

Integration of knowledge is not a matter of having a model, or of having a sufficiently rich one, but of being able to use knowledge flexibly and thoroughly. It allows new models to be created, based on new combinations of existing models and previously known facts. It allows models to be checked against one another and against their referents for inconsistencies. In brief, it allows models to be treated as something more than black boxes.

Thus, having a model of a process is not a sufficient condition for understanding it. Nor, as models have been defined here, is it a necessary one. The most basic form of knowledge is not based on models, but on concepts; and in many cases, forming a model is not the most appropriate way to achieve understanding.

In many ordinary situations, we do act on sets of expectations that could be regarded as informal models. For example, in going to a movie, you would expect to give the cashier some money, receive a ticket, enter the theatre, sit in a seat, see a movie, get up, and leave. These common sets of expectations are the basis of the concept of "scripts," as used in artificial intelligence. But understanding occurs only when we understand the significance of each of

these events: that the management wants to make money, that in order to do so it requires people to pay to see the movie, and so on.

Concepts, as discussed in the first chapter, are the mental units which tie awareness of specific things (percepts) together so that we can deal with facts that pertain to whole classes of entities. Any understanding of a new phenomenon requires first knowing what its elements are and what is common to its various instances. The application of concepts already known or creation of new concepts is the step which is needed to identify and tie together these elements. We must know that a candle burns -and what burning is before we can ask with what degree of heat. The child learning about television must learn what it is, in the basic terms appropriate to his needs, in order to formulate even the simplest model of how it works. The concept and model may grow together; the child may learn about a TV set by fiddling with the controls, and thus learn its capacities and its operation at the same time. Alternatively, he may be told how to work the TV before he knows what those actions will do, but the instructions do not become understanding until he discovers what following them accomplishes.

This is not to say that conceptual understanding is merely a first step, which later grows into a more detailed model-based understanding. The ability to integrate facts and discover new relationships requires conceptual thought. A model is one specific way of looking at a process; a concept is the sum of one's knowledge about the particulars belonging to a class. If new information about a process is discovered, or if previously disregarded information is found to be important, it must be identified in terms of the relevant concepts and then put into a revision of the model.

Suppose a child does all his experimenting with a TV during the same hour of the day. From the information he gets, he may conclude that turning to Channel 7 will always get him *Tarzan in Space*, and so on for other channels. This information becomes part of his concept of television; he concludes that a certain channel is associated with a certain series. It also becomes part of his model; he learns that turning to Channel 7 will make *Tarzan in Space* visible. If he later expands his experimentation to other hours, he will discover that the same series isn't always on the same channel. If he is clever, or if someone explains it to him and the explanation matches his observation, his conceptual knowledge will expand to the realization that a certain channel and a certain time are associated with each series. As a consequence of this understanding, he can revise his model; the time of day becomes one of the variables he has to take into account. (Later on, of course, his model will expand considerably more. He will learn, for example, that the association of time, channel, and program is not something which characterizes television as such, but rather characterizes the way broadcasters operate. This example is concerned only with a child's early discoveries and attempts at understanding.)

The integration of conceptual knowledge may occur in a number of ways. New facts may be discovered which are true of all members of the class to which the concept refers. Exceptions may be discovered to principles which were previously believed universally true of the class. Elements which were previously regarded as part of the class may be found to be so different that they should be excluded; other entities may be found to belong to the class even though they were previously considered alien to it.

Integration of knowledge with a concept includes not just universal truths but tendencies and rules of thumb. It is part of our knowledge about human beings that most are between four and seven feet tall, although being outside this range doesn't disqualify a person from the race. It is part of our knowledge about weather that if there are no clouds in the sky now, there is

little reason to fear rain in the next two hours, although there are cases where it can occur on such short notice.

There are two opposing fallacies in dealing with concepts, both of which hinder their effective use. One is that concepts are built starting from a definition, to which all referents of the concept must conform. The other is that concepts are arbitrary collections referred to under a convenient name. The first approach makes a concept independent of the knowledge we gather; it does not leave any room for revisions of understanding. The second approach leaves us floundering in a sea of unrelated information that has been put under a single name; it does not give us any way to make predictions with confidence about the whole collection.

In philosophy, the division between these two approaches to concepts, and to knowledge in general, is known as the division between "analytic" and "synthetic" knowledge. Analytic truths are obtained by deduction from arbitrary axioms; they are universally true, but only by decree. Synthetic truths are obtained from experience, but philosophers have struggled with the question of how universally applicable knowledge can be obtained from finite experience. Mathematics is widely regarded as essentially analytic. Formal systems, which are mathematical constructs that prove theorems from axioms with no regard at all for meaning, are the purest example of "analytic truth."[2]

Both of these views of concepts have gained support because people do in fact sometimes think in both of these ways. Some people will start with a definition for a concept and then stick with it, regardless of how unwieldy further discoveries show it to be in comparison with alternative definitions. This is a typical academician's error. Ordinary people are more likely to take some entity which looks "sort of like" other elements of a class and consider it an element of the class, then add another which looks sort of like the one just added, and so on until the elements being added have nothing at all in common with the original ones. But both of these approaches are clumsy for organizing knowledge. The first amounts to a demand that knowledge fit a pre-planned layout; the second is a case of forgetting where a concept came from in the first place. Both are defective forms of the basic human form of organizing knowledge, which is to discover common characteristics among entities and treat them as groups on that basis.

The reason for making these errors is instructive and points out the value of models as a cognitive tool. A difficulty with concepts is that they cannot, strictly speaking, be communicated; each person has to form his own concepts. The process of learning to treat a concept as a single item in the mind, of being able to recall without effort what it stands for, is one that takes effort. Before making the concept one's own, the learner must rely on one means or another of reconstructing the concept. Usually this means remembering either the definition or some examples. But each carries its danger. The learner may treat the definition as a rule imposed from above, rather than a means of identifying essential similarities. Going to the other extreme, he may remember the examples while forgetting what their similarities are, and start lumping other entities into the concept in a haphazard manner.

Models largely avoid this difficulty, specifically because they give up the flexibility of conceptual thinking. This difference makes them useful for communication where the more immediate concern is appropriate action rather than understanding. An assembly-line worker can be given a model (in the form of instructions) of a small part of the assembly process; to do his job, he does not need to understand the process or even the reason for what he does, but only to know what actions to perform. (I am referring here only to cognitive necessity under routine circumstances; knowledge is still important both for purposes of motivation and for

being able to deal with unusual situations.) A computer programmer is given a flowchart or pseudocode for a program rather than (or in addition to) a purely conceptual explanation of what the program should do, so that there are as few opportunities as possible for misinterpretation. Given a fixed set of conditions, a model specifies a particular course to follow without ambiguity. This does not exclude the use of conceptual understanding, but it restricts it to the cases where it is more likely to be valid than blindly following the model. A factory worker can be expected to know how to break out of the model if an emergency occurs; the programmer ought to be able to notice blatant typographical errors in pseudocode.

At the same time, models can be an aid to conceptual understanding. To the extent that a model corresponds structurally to the referent process and predicts it accurately, it can provide an alternative to working with the actual process as a means of understanding it. A certain amount of knowledge about flying an airplane can be gained from a flight simulator; schematic diagrams of circuits, together with descriptions of their behavior, can help the electrical engineering student to understand circuit design.

Finally, to return to Johnson-Laird's thesis, having a working model of a process in one's mind is itself a form of understanding; it allows a person to predict how a process will run and identify aspects of the process in quantitative terms. It can be essential to understanding where an attempt to comprehend the process in detail from the separate concepts involved is too laborious to be practical. But this is not the "psychological core" of understanding, but a derivative of the more basic conceptual form of understanding. Without that underlying support, what is present is not understanding but just rote learning. With it. the model is a powerful tool.

"Scientific" Understanding

Up to this point, I have been dealing with the issue of whether modeling is basic to the operations of the mind. A separate question is the extent to which specifically scientific understanding is based on models. as opposed to conceptual thought as I have described it here. One could make a prima facie case for this conclusion: Concepts, being open-ended, have a certain vagueness to them. We may decide to exclude tomorrow from a concept which we include today. If we base scientific laws on concepts, then, they will be insufficiently precise. A more exact approach is to construct a model of a set of phenomena, explicitly including some variables and forcing all others to be constant, so that we can precisely and mathematically describe what is happening.

For instance, in presenting Newton's laws of motion, we assume that an object in motion at a constant velocity has a constant mass. In fact, this is never strictly true. It may gain mass by incorporating oxygen from the air into its molecules, or it may lose mass by friction or molecular decomposition. Normally, these changes are negligible, so our model is sufficiently accurate for the real world; but it is only 100% accurate in terms of the model itself.

There is a certain plausibility to this view; but insofar as it makes models fundamental to science, it is seriously mistaken. Accepting this view of science leads to the error against which I warned in Chapter 1: considering only the model's viewpoint and thinking that everything that is true of the reductive model is true of the process. Models must be used in science, but they must be regarded as simplifications and used with the whole context in mind. Otherwise, a scientist will find himself retreating into an ivory tower of conclusions that have no connection with real life. This is the Platonic approach, which regards the ideal as more

important than the real. The ideal is a characterization of the real, or of that which could be real, focusing on certain elements and excluding others. Ideals, or models, have no actual existence apart from the real entities from which they are abstracted.

Joseph Weizenbaum, in *Computer Power and Human Reason*, presents the view that science does not deal directly with reality, but "distorts" it in an effort to make it more tractable:

> Science can proceed only by simplifying reality. The first step in its process of simplification is abstraction. And abstraction means leaving out of account all those empirical data which do not fit the particular conceptual framework within which science at the moment happens to be working ... One of the most explicit statements of the way in which science deliberately and consciously plans to distort reality, and then goes on to accept that distortion as a "complete and exhaustive" account, is that of the computer scientist Herbert A. Simon ... [3]

When discussing the role of models in science, Weizenbaum is quite incisive:

> The aim of a model is, of course, precisely not to reproduce reality in all its complexity. It is rather to capture in a vivid, often formal way what is essential to understanding some aspect of its structure or behavior. The word 'essential' as used in the above sentence is enormously significant, not to say problematical. It implies, first of all, purpose. In our example, we seek to understand how the object faUs, and not, say, how it reflects sunlight in its descent or how deep a hole it would dig on impact if dropped from such and such a height. Were we interested in the latter, we would have to concern ourselves with the object's weight, its terminal velocity, and so on. We select, for inclusion in our model, those features of reality that we consider to be essential to our purpose.[4]

This is entirely correct, but Weizenbaum makes an error in considering this approach to be essential to science, rather than to one of the tools of science.

There is a genuine issue which he is dealing with. Science does not deal with particular objects as such, but with the common characteristics of similar objects. The fact that copper conducts electricity is a scientific principle; the fact that a particular piece of wire does is not. One must identify the particular with the abstraction, the given wire with the concept "copper," in order to apply the principle. One can then determine the resistance of the wire, given its shape and the resistive characteristics of copper; but in order to apply the formula for the resistance of a piece of copper, one must regard the impurities in the metal, the irregularities in the shape of the wire, etc., as negligible. Does this mean that science must "distort reality" by forcing us to regard the wire as a piece of idealized copper in an idealized shape?

In fact, it does not. The proper method of scientific thought is based not on models regarded as black boxes, but on conceptual thought. The difference which is important here is that models have a closed context, but concepts have an open-ended context. A given scientific principle states what will happen, all else being constant. Whether all else is close enough to constant that the principle can be applied without modification depends on the purpose of the person applying it and the nature of the phenomenon in question.

Every scientific principle is subject to further refinement. The classic example is Newton's laws of motion, which appear perfectly correct under everyday conditions, but which break

down at velocities approaching the speed of light. If the goal of science were to provide a simple model that distorted reality for the sake of convenience, there would have been no point in Einstein's discovery of relativity; we could just throw fudge factors into the Newtonian equations to cover the discrepancy from reality. The purpose of science is not to distort reality, but to characterize it in a way that applies to entire classes of entities or phenomena.

Neglecting this fact makes the scientific approach appear sterile and deficient in dealing with real-life problems. In popular images, this view of science is reflected in the "absent-minded professor" who has no sense of practical reality. But setting science at odds with reality is like setting the mind at odds with the body. Science is a specific category of human understanding, the one which deals with the common characteristics of natural phenomena. Its epistemological foundations must be the same as those of philosophy, history, and investigative reporting. Its goal must be the best possible characterization of its subject matter. The word "best" applies in two senses: the best characterization of something is that which is most accurate and most suited to human cognitive requirements. Some kind of simplification is always necessary for human understanding. No one could deal with an account of the Battle of Gettysburg that exhaustively reported what every soldier did, or with an account of a political scandal that followed the participants minute by minute for a whole day. But in each case, there is a context, and a fact which previously seemed unimportant may turn out to be vitally important in the light of a new discovery. ("For want of a nail...")

A scientific law is not the same as a scientific model. In a model, the methods of analogy and simplification may be used freely. A model gives a means of dealing with the behavior of a type of entity or system by putting it in terms which are comparable in structure and in consequences. A model as such does not make a statement; the use of a model implies the statement that certain similarities exist between it and its referent. A law, on the other hand, must refer to the actual entities or substances in question and make a true statement about them. A law, like a model, grows out of observations and is only valid when considered in its proper context; but its requirements are more stringent. Ptolemaic astronomy, for instance, which places the Earth at the center of all astronomic motion, is invalid as scientific law, since the motions of the planets and stars are not governed by the Earth; but in an appropriate context, it is still valid as a model. We use this model, in fact, every time we speak of the sun or the moon or a star rising.

Science must deal fundamentally with concepts, formalizing them into models only as a second step. If we regard models as the basic material of science, then we are indeed stuck with a distortion of reality. Unfortunately, this error has been made too many times, as the following chapters will show. But giving this role to models is neither a virtue nor a failing of science; it is simply an error which some scientists have made at some times and a failure to live up to science's highest standards.

---

1 Johnson-Laird, p. 2.

2 For a detailed discussion of this issue, see Peikoff, "The Analytic-Synthetic Dichotomy."

3 Weizenbaum, p. 127-128.

4 Ibid., p. 149.

III. Mathematics and Logic

Mathematics is the basis of precise modeling and is essential to a very large part of science. It is the most abstract of the sciences, its task being not to identify specific facts of reality but to systematize methods of measurement. Its abstractness makes it the most susceptible of all sciences to being considered apart from its relationship to reality. Its importance to all of science makes it the weak link through which all of science can begin to float off into a Platonic realm.

In the early days of mathematics, its relationship to the physical world was plain and obvious; yet as early as the days of Pythagoras, numbers were granted mystical significance. There is, it would appear, a risk in any abstraction that people will forget where it came from and give it a life of its own.

The mathematical elements which date from those early times have clear referents in reality. Natural numbers refer to quantities of discrete objects. Fractions were developed from rational numbers to deal with continuous distances. The discovery that certain common quantities, such as the diagonal of a unit square, could not be expressed as a ratio of integers, was perhaps the first area of mathematics that made people think there was something "magical" about numbers. The idea of an irrational number (literally, a number which is not a ratio) seems harmless enough today, but for people to whom mathematics was counting, it was a disturbing step away from common sense.

However, real numbers and negative numbers still have clear referents. The concept of negative numbers arose when people wanted to express a quantity that was not absolute, but rather relative to a reference value; a negative quantity is one opposite to a positive one (e.g., velocity in one dimension), or less than the reference value (e.g., temperature).

Newer creations of mathematics have less obvious ties to anything real. Imaginary numbers, as their name suggests, are an important example of this trend; they were created not to refer to anything, but to make the treatment of quadratic equations more regular. If square roots of negative numbers are not admitted, then some quadratics have two roots, others just one, and others none at all. But if we allow the "imaginary" construct of the square root of -1, then every quadratic has two (possibly equal) roots.

What is surprising is that imaginary numbers eventually turned out to have a "real" use in the treatment of waves. This pattern has repeated itself with other mathematical concepts. The broadening of mathematics has allowed the creation of interpretive contexts for models that had not yet been imagined. This applicability of new ideas has, perhaps, contributed to the belief that the creation of systems on the basis of applicability isn't the business of mathematics.

Mathematicians will often claim that their study deals with abstract objects that need not have any referent; they will often challenge critics of this belief to explain what an imaginary number or a Cantorian infinity represents. Russell states that "all Mathematics is Symbolic Logic." Any ties of mathematics to reality are an afterthought, not part of the "pure" science. "All propositions as to what actually exists, like the space we live in, belong to experimental or empirical science, not to mathematics; when they belong to applied mathematics, they arise from giving to one or more of the variables in a proposition of pure mathematics some

constant value satisfying the hypothesis, and thus enabling us, for that value of the variable, actually to assert both hypothesis and consequent instead of asserting merely the implication."[1]

If we accept this view, is mathematics really a science or merely a game? It becomes difficult to call it a science, since it does not refer to any facts, but only to its own axioms and rules of implication. There is no limit to the number of mathematical spaces and axiomatic systems that can be invented, all of which qualify as pure mathematics, but none of which need have any application. The same is true of games; you can invent a game with any consistent set of rules you want, without worrying about whether it stands for anything.

The idea of "pure" mathematics exemplifies the consequences of the analytic-synthetic dichotomy. If we must regard all truths as either "analytic" and arbitrarily true by logic or "synthetic" and contingently true by experience, then mathematical truths are presumably analytic and divorced from reality. But if we recognize that all truths, including logical ones, are derived from experience, there is no reason to make such a split. The basic principles of an area of mathematics are abstractions from experience. In some cases they may be indirect rather than direct abstractions, as in the case of imaginary numbers; but they are not arbitrary constructs selected by whim. The application of the rules of logic to the basic principles is an important part of the mathematician's job, but so is deciding what principles to start from. This is not to say that leaps of imagination are illegitimate; an active imagination is important to creating new and really useful systems. But the builder of systems who neglects their connection to existence is not a mathematician, but only a game player.

Russell's statement about the validity of mathematical statements is true in the sense that new discoveries about the universe do not invalidate the mathematics used to describe it, but only the applicability of the mathematics. Thus, Einstein's "curved space" does not perfectly fit Euclidean geometry, but it does not destroy the validity of Euclid's system as mathematics. It remains valid because of its internal consistency, and because it is a highly accurate interpretive context for physical models in nearly all cases of interest to people. To be part of science, a system must say something about the universe, not merely about itself. A mathematical system which has no relationship to any method of modeling or measuring something may be an exercise, a speculation, perhaps even a work of art; but it is not science.

The treatment of mathematics as something unrelated to reality, except as an "afterthought," may seem to have the virtue of avoiding the black box fallacy, since it discourages the mistaken idea that truths about mathematical systems are necessarily truths about the real world. There is a valid concern underlying this idea, but prying a portion of thought loose from reality is not the way to achieve this end. The result is often just the opposite; by regarding mathematical descriptions merely as manipulations of symbols rather than as abstractions from reality, people can be tempted to regard their understanding of a process as a purely symbolic understanding, rather than one based on experience and then translated into symbolic terms for ease of comprehension. This result has had profound implications in both physics and computer science, as subsequent chapters will show.

The Role of Logic

Since modern mathematics is regarded as essentially an exercise in symbolic logic, much the same considerations apply to logic. Again as a result of the analytic-synthetic dichotomy, we are told that it is an arbitrary human construct, that its purpose is not to describe reality but to describe the way we think about reality. In popular discussions, we hear claims such as,

"You can prove anything by logic." In more esoteric circles, "old fashioned binary" logic is discarded in favor of multi-valued logics and "fuzzy" reasoning.

Yet clearly logic is related to reality. If you ask a clerk in a supermarket where to find the carrots, and he tells you that vegetables are at Aisle 9, you wouldn't consider it an arbitrary action to go to Aisle 9. What you have done is to follow an implicit syllogism:

All carrots in this store are vegetables in this store.

All vegetables in this store are at Aisle 9.

Therefore, all carrots in this store are at Aisle 9.

Why, then, do we hear the claim that logic doesn't tell us anything about reality? We are told that logic only gives "empty tautologies," without providing any new knowledge. Any conclusion reached by logic contains no more information than the premises contain.

But why is this an objection? Without logic as a means of going from the general to the particular, we would have to discover each fact by a separate act of observation. In the example just given, the general statement that vegetables are in Aisle 9 would not allow us to conclude that artichokes, cucumbers, and tomatoes are all in Aisle 9. There would be no way to unify those facts.

Seemingly, the "tautology" objection to logic is an objection to the effort it saves; it admits only "new knowledge" as significant, where new knowledge is obtained by a new act of observation. If we were stuck with learning each fact that way, we would hardly have made it out of the caves.

But the real basis of this objection is subtler than that. It is the result of our old friend, the substitution of the model for the process. Over the past few centuries, mathematicians made the useful discovery that logic can be formalized. This led to an unfortunate consequence in the hands of twentieth-century philosophers; the precision of formalized logic, or symbolic logic, became so attractive that logic became equated with the formalism. The result was that it came to be taken as simply a way of manipulating statements; the irony was that the lowering of its status came from those who were most enthusiastic about logic.

The advantage of formalizing logic lies in the array of mathematical tools that become available for solving complex logic problems. As with many other forms of modeling, the greatest advantage has accrued not to the usual applications of the field, but to computer design. The operation of a computer depends on the ability of the hardware to detect conjunctions, disjunctions, and inversions of events. Such treatments can easily be described in the language of formal logic; in fact, the devices that perform these functions are themselves often called "logic." ("The disk controller contains logic to make sure the track isn't written while the cover is open.") Without Boolean algebra and Karnaugh maps, computer hardware design would be much more painful.

But formalizing any process disconnects it from its meaning; this is its advantage, since it permits automatic manipulation of the resulting model, and its handicap, since it entails the risk of losing track of the meaning.

The concept of implication is one of the key points at which this risk becomes evident. Formally, "A implies B" means "B or not A." If A implies B, then the only possibilities are

that A is not true, in which case the truth or falsehood of B is unestablished, or A is true, in which case B must also be true.

In the formal meaning of implication, there is no need for A and B to be connected at all. We can say that "Napoleon won the battle of Waterloo" implies that "Socrates is immortal," since the antecedent clause A is false. We can say that "All men are mortal" implies "Cats have four legs," since the consequent clause B is true.

But clearly something is wrong with these examples. The two assertions are unrelated. We can imagine a world in which Napoleon won at Waterloo, but that would not give Socrates immortality. Nor, if we did not already know that cats have four legs, would we be able to deduce that fact from the mortality of man.

Implication, as it applies to the real world, includes the element of necessity. It also implies that if the antecedent condition could be changed, the basis for asserting the consequent would be undercut or removed. For example, consider what we are saying when we assert that "John Smith is a professional composer" implies "John Smith writes music." (Quibbles over the meaning of the separate words are irrelevant here; we are assuming no one is trying to play a dirty trick by using an alternative definition of "composer" or "professional.") What we are saying is that by knowing that Smith is a professional composer, we can know that he writes music. If our knowledge that he is a professional composer turns out to be based on mistaken sources, then we cannot use the implication as a means of establishing that he writes music. (We could still be certain of the conclusion by other means, such as establishing that he is an amateur composer; but that particular implication ceases to be a reliable avenue to the conclusion.)

A formalism cannot directly take this necessity into account, since it is the result of a context which is discarded. To formalize the connection, we would have to include additional statements, such as "A composer is a person who writes music." But once we add all the necessary links, we have (ironically) removed the need to refer to necessity. Each step in the process is a formal syllogism, and there is no need to question whether the formal relationship is necessarily true or just happens to be true without any particular relationship between the facts. In this particular case, each premise is true either by assertion (e.g., "Smith is a professional composer") or by definition (e.g., "A composer writes music"). As with any properly formulated model, there is no need to go outside its stated rules and premises in order to establish how it operates.

But the primary purpose of logic is one which is not found in a formalism, and that is to achieve knowledge of facts. This is a cognitive issue, and therefore outside the realm of modeling. As we have already established, there is an essential difference between understanding something and having a model of it.

Quantification, like implication, is significantly changed in the transition from cognitive logic to symbolic logic. In symbolic logic, "All Jabberwocks are Greek" is a true statement, which would be written

$$\forall x{:}\mathit{Jabberwock}(x) \supset \mathit{Greek}(x)$$

The interpretation of this statement in symbolic logic is that for any x, either x is Greek or x is not a Jabberwock. This is indeed true, since there aren't any Jabberwocks. This is an entirely different matter from the English statement "All Jabberwocks are Greek," which

would normally be taken as meaning that being Greek is an attribute of a Jabberwock, a claim which Carroll's poem doesn't support.

There are two primary cognitive functions which logic serves: to draw new conclusions and to verify hypotheses. The first function usually operates automatically, by a habit that does not require conscious effort. The shopper in the supermarket isn't likely to formulate the syllogism before running off to Aisle 9; he recognizes the fact without having to think about the laws of logic. The second function involves an explicit use of logic in order to test whether some combination of known facts proves the hypothesis in question. If someone asked the shopper, "How do you know carrots are in Aisle 91" and he had some training in logic, he would be able to answer by stating the syllogism.

Actually, this example is so simple that unless he was suffering from an overdose of academics, he would more likely answer, "Because the clerk said so!" not even remembering that the clerk said "vegetables" and not "carrots." The operation of logic is usually highly automatic, not in the sense of being inescapable (people can refuse to be logical), but in the same sense that walking is automatic: the process has long since become a habit, and no conscious effort is necessary. But if the issue is a complicated one, the proof has to be made explicit in order to avoid errors. (The analogy to walking applies here as well; if you are crossing a stream by stepping on slippery stones, you must give your full attention to what you are doing.) That is the reason for the explicit use of logic in human reasoning: people are capable of errors and must check their inferences.

Logic in Reality

People use implicit logic constantly. If a door which you want to open is locked and you have a key to it, you will take it out of your pocket; the reason for doing this involves a large number of inferences. Life would become impossible if you had to rigorously justify each conclusion you reached. But this does not mean we reach our conclusions without any need for logic; rather, it means that we can use it so easily that we only have to identify it in special cases. This is analogous to the way people read text; they do not stop to spell out each word and remember what it means, but they once had to do that, and they can stop and do it when doubt arises.

Logic is something which people learn, in the explicit sense, only after they have used it for a long time. Every relationship of facts provides an example of the laws of logic: nothing an infant sees contradicts itself, effect follows from cause. (If a child's parents provide an apparently self-contradictory environment through their inconsistencies, his confidence in logic will not be so great; the result is anxiety and neurosis.) Children implicitly understand the need for consistency and inference long before they can explain the principle behind a syllogism.

Recognizing this undermines the barrier between analytic, logic-based truths and synthetic, empirical truths at its base. Logic itself is discovered by experience. Like any other form of conceptual knowledge, it is an integration of data obtained over time by observation into a general principle. It is not an arbitrary construct for deriving arbitrary conclusions, but a summary of the way reality works.

Formal logic deals only with statements; it does not attempt to say anything about reality. One could create other formalisms of different "logics," such as a ternary "yes-no-maybe"

logic, or a "fuzzy" logic with a continuum of values between "no" and "yes," instead of a binary "yes-no" logic. These formalisms have their uses for dealing with uncertainty, or with the measurement of attributes that are present to different degrees. But they are not actually forms of logic; they deal with derivative issues. It is impossible even to talk about these other "logics," including formal logic, without making use of classical logic.

Logic deals with the most basic alternative in the universe: the alternative of existence or non-existence. Uncertainty is an issue that pertains only to the mind; there are no "maybes" in reality. Attributes may exist to different degrees, but first we must establish whether they are present at all. With degrees of attributes, we can likewise ask yes-or-no questions: is the attribute present in the specified degree or not?

Very often, there is an area of doubt along the boundary between possession and nonpossession of an attribute; but this is simply because the cognitive standards used have finite precision. This is true of any quantity or quality which is continuously variable. Given that a standard is set which is capable of the necessary degree of discrimination, an entity either possesses or does not possess the attribute by that standard. Various shades of color, for instance, can be called red or non-red. Some shades will fall on the borderline and will neither be definitely red nor definitely non-red. Applying the standard will produce some classification, but repeated applications may not produce the same result, and the person applying it will have no confidence in its correctness. Does this invalidate the law of excluded middle? No, the "middle" is a matter of our criterion's failure to produce a definite answer. The laws of logic do not provide us with a classification for everything; they only say that, given a method of classification, a given object must either be accepted or not accepted under it. This is not the same as saying it must either be accepted or rejected; it is possible to refuse to commit oneself, but such a refusal is a case of not accepting it.

Vague borderlines can be used to set up amusing inductive "proofs" of obvious falsehoods. For example:

A man with no hair on his head is bald.

A man with one hair more than a bald man is still bald.

Therefore, by the principle of induction, for all **n**, a man with **n** hairs on his head is bald. In other words, all men are bald!

The reason for reaching this baldly fallacious conclusion is that we normally do not establish baldness by counting hairs, but by a quick observation; hair-splitting exactness serves no purpose. Certain degrees of baldness will be judged differently by different observers, as well as causing a single observer to vacillate. But this is because the standard doesn't serve to make all possible distinctions. The existence of such borderline cases doesn't demonstrate that "binary" logic is inadequate; it demonstrates that adequate standards of discrimination must be set before logic can be used. The inductive proof arrives at an absurdity because it plays on the ambiguity between a hypothetical, exact criterion for baldness and the usual vague criterion of observation. The law of non-contradiction contains an important qualifier: Something cannot be A and non-A *at the same time and in the same respect*. To the extent that the respect in which something is considered A or non-A is vague, the laws of logic cannot be applied reliably.

Models are not fallible in the sense that people are fallible. First of all, fallibility is a characteristic of beings capable of knowledge, and models do not possess knowledge. Also, a model has a specific method of operation; it does not get tired or angry and thus stop working; it does not have prejudices that prevent its reaching correct results in certain cases. If the purpose of the model is to account for these features, they can be modeled to a greater or lesser extent; but then the model is functioning properly by virtue of simulating these malfunctions of the process it represents.

A model, therefore, does not have to call upon logic to verify its own operation. Its purpose is not cognitive, not to establish that something is true, but functional, to carry out a certain process. While it is possible to model the cognitive uses of logic, this is a complicated and specialized case; the more pervasive use of logic in modeling is in the control of the model's operations. This use includes performing certain steps or not performing them depending upon whether certain conditions are satisfied. On an even lower level, Boolean algebra is used to predict and verify the operation of a model that involves binary operations (including nearly all computers); this is a case where all the laws of logic apply, but refer to signals rather than to facts.

The use of formal logic for describing signals and circuits illustrates the common phenomenon that the same formalism can be used in more than one model. The multiplicity of applications for formalisms is one of their great strengths; it often turns out that by turning the rules of description of a process into a completely abstract system, one can apply them to a system which is apparently unrelated to the original one. This does not violate the fallacy of ascribing reality to the model, since it is still only in the application that the formalism becomes a description of reality. Computer scientists may choose to use the word "logic" to refer to the operation of digital circuits, but that is a different meaning of the word from "logic" as rules of inference. The use of the same word serves as a reminder that the same formalism describes both, but it can also be a source of confusion to the person who tries to combine both into a single definition and identifies logic as just a formalism.

The Basis of Logic

One area in which logic commonly comes under attack is the argument that its starting point must be arbitrary. Logic can derive conclusions from premises; but the first conclusions must be arbitrary axioms, accepted on faith.

There is an obvious absurdity to this objection; it is an attempt to refute logic by means of logic, and thus invalidates itself. But it needs to be answered in order to understand the role of logic in human thinking, as opposed to axiom-based models.

It is true that logic must start with premises which are not themselves based on logic; but it does not follow that these premises are arbitrary or uncertain. Rather, they are the preconditions of logic.[2] The primary axioms of knowledge, and of existence, were identified by Rand as:

(1) Existence exists.

(2) Something exists of which one is aware.

(3) One exists possessing consciousness, consciousness being the faculty of perceiving that which exists.

These axioms are not provable. One cannot go to a category broader than existence to prove it exists; nor can one prove to a being that it is conscious if it does not recognize the fact. But any attempt to deny them makes all knowledge impossible, including the knowledge that one has denied them, that one is capable of doubting them, and that there is any issue of truth or falsehood involved. By rejecting existence, a person leaves himself with nothing to talk about. By claiming that existence lies outside his awareness, he declares himself totally ignorant and unqualified to talk about anything. By denying that he is conscious, he denies that he has any capacity to know, to question, or to deny.

These axioms are essentially conceptual; what they actually do is assert the concepts of existence, identity, and consciousness respectively. In form, "Existence exists" is simply a tautology; but the statement must be understood not as saying something (or saying nothing) about a previously understood concept of existence; rather, it is the identification of the concept of existence in the form of a proposition.

One of the commonest errors made by critics of Rand's epistemology is the interpretation of her axiomatic base as an expression of formal logic. Her starting point is "Existence exists," not because the English language makes it true by definition, but because it is the broadest statement of what has to be understood in every act of understanding: that whatever discoveries you make, whatever facts you grasp, they pertain to something which exists. Take away existence, and you take away everything, including yourself and all your knowledge. Logic is the result of the recognition that everything you deal with exists and is something in particular.

Formal systems (which will be discussed in more detail further on) are a completely different case. Axioms in a formal system, unlike cognitive axioms, are just strings of symbols. Mathematical axioms are meaningful abstractions, but they are also different from the conceptual base of knowledge; they are assumed true for the purpose of creating the mathematical system and are not necessarily undeniable in a broader context. Euclid's axiom regarding parallel lines is a well-known case of an axiom which can be denied without self-contradiction. The axioms of knowledge, on the other hand, are literally undeniable. One can grammatically negate them, but the resulting negation (e.g., "Existence doesn't exist") has no possible referent.

The axioms of knowledge form the basis for accepting sense evidence. The senses do not give us facts directly, but they give us data which are necessarily the consequence of something that exists being perceived by our consciousness. This is not the place for a detailed discussion of why the senses are valid (for such a discussion, see David Kelley's *The Evidence of the Senses*)[3], but a few points are worth making in the context of whether the starting point from which logic works is arbitrary.

If it is arbitrary to conclude that what we see with our eyes is something that exists rather than a constant progression of hallucinations, then the distinction between the arbitrary and the non-arbitrary vanishes. For if the critic of the senses is consistent, he must reject the evidence of his own senses. He must then regard his own mind as constantly manufacturing illusions beyond his conscious control, the world as a complete unknown to him, and his knowledge as non-existent. But then he is no more qualified to assert anything at all than a newborn infant is. His knowledge, or even his assertion of the possibility, that the senses are invalid leaves him in the position of making this assertion arbitrarily, without any evidence at all. Hence it is the denial of the senses' validity, not its acceptance, with is arbitrary.

Once a person has gathered a certain amount of sense evidence, his cognitive job is to form concepts that unify the information gathered, and then to characterize the relationships among the entities thus identified by formulating general principles. For instance, by repeatedly perceiving water in its liquid state, a child can formulate the concept "water" and the principle, "Water is wet." As his knowledge increases, he will add appropriate qualifiers, such as "Water is wet when it is not frozen or evaporated." This statement then becomes a workable premise for use with the laws of logic. The process is not arbitrary; it is the only one which is consistent with reality.

To recapitulate, logic is not simply a human convention or a set of formal rules for manipulating statements; it is a description of reality at the most basic level, that of what exists and what does not. Formal logic is an abstraction of logic that deals only with propositions, not with facts; this abstraction is useful because it can be mechanized, and because it can be applied to processes other than logic. But forgetting where this formalism came from undermines the whole foundation of human knowledge, making everything appear either arbitrary or uncertain.

---

1 Bertrand Russell, *Principles of Mathematics*, Second Edition, 1938, W. W. Norton and Co., p. 5.

2 For a more complete discussion of axioms in cognition, see Rand, Chapter 6.

3 David Kelley, *The Evidence of the Senses*, Louisiana State University Press, 1986.

IV. Probability

Probability raises a distinctive set of issues, since it deals with uncertainty and randomness. Probabilistic (or statistical) models introduce the idea of events that are not repeatable or predictable, yet about which quantitative statements can still be made. This situation requires a certain amount of explanation to avoid the sense of paradox suggested at the opening of Poincare's *Calcul des Probabiltés*: "How dare we speak of the laws of chance? Isn't chance the antithesis of all law?"

There are two ways to think of probabilistic models. One is to regard the model as a process which, if run several times, will not always produce the same value for the variable of interest (the number of events), but which can be expected to produce any given value for the variable roughly a specified, predictable number of times when run a specified large number of times. The other is to think of the model as a description of the ideal distribution of values which should occur in a large number of occurrences of the process.

Either way, there are problems in thinking about statistical models which are not found in connection with deterministic models. Both approaches deal with outcomes which are "expected" or which "should" happen, but which are not guaranteed to happen. It is possible in a Poisson process for the event never to occur in a hundred runs, even if the distribution says that the event will occur twice in a run on the average; but we do not expect this to happen. It is "highly improbable." But what does this mean?

One way to look at probability is as a matter of personal commitment. If we regard an event as very improbable, we make our choices based on the assumption that it will not happen. If the sun is shining, it is improbable that it will rain; we don't take umbrellas. We venture out on the streets because we consider it improbable that a car will go out of control and kill us. We bet on a horse because we believe it is likely to win.

But this answer is unsatisfactory. For one thing, it is purely subjective; a person can act as if something is not going to happen when all the evidence says that it almost certainly will happen. For another, we can assign probabilities even where no personal commitment is involved beyond the satisfaction of seeing the event we called "probable" happen; we can regard it as probable that the better of two teams will win, even without placing a bet.

A more tenable view is to regard probability as a matter of distribution of outcomes in processes which we regard as equivalent. Thus, we say that a coin has a probability of coming up heads of 0.5 because, in our experience, coins which have been flipped have come up heads half the time. This view is well argued in Richard Von Mises' *Probability, Statistics, and Truth*. According to Von Mises, "The rational concept of probability, which is the only basis of the probability calculus, applies only to problems in which either the same event repeats itself again and again, or a great number of uniform elements are involved at the same time."[1] Probability applies only to the portion of the "collective" which meets some criterion, not to individual events. Thus, if actuarial tables show that a person of age 40 has a 1.1 % probability of dying within a year, it is nonetheless "utter nonsense to say, for instance, that Mr. X, now aged forty, has the probability 0.011 of dying in the course of the next year."[2]

In order to be ascribed a probability, the events must be "random." This is necessary in order to assure that a sample which is taken out of a group of events will be representative of

the group. If the events are random, then as the size of the sample increases, the proportion of the sample which satisfies the criterion in question will converge to a particular number. For example, as a fair coin is flipped a growing number of times, the proportion of heads will converge to 0.5. If the events are not random, this is not guaranteed. For instance, if words are examined sequentially in a dictionary, it will be found after taking 500 samples that all of them start with A; but it does not follow that as more and more words are examined this way, the proportion that start with A will remain close to 1.

There are some difficulties here. The key assumption in Misean probability is that as the size of a sample taken at random approaches infinity, the proportion of the sample that meets the stated criterion will approach a constant. Russe11[3] has noted that in nearly all cases, the class is finite, so the sampling cannot be carried out to infinity. But this objection is not major, as long as the sample follows the trend as far as the size of the class allows. In fact, if sampling with replacement is allowed, the sample can be made indefinitely large, regardless of the size of the class.

There is a more basic question, though: how do we justify this assumption? When we flip a coin, it is possible each time that it will come up heads; yet we can be confident that flipping it a hundred times will not give a hundred heads. Mises gives an answer which can only apply retrospectively: if the proportion of heads tends to one half in a large number of tries, and nothing about the way the tosses are performed is changed, then the proportion will continue to remain close to one half.

To see why this should be so, we have to consider what "randomness" means. Mises defines randomness, or suitability for the application of probability theory, in terms of two criteria:

> First, the relative frequencies of the attributes must possess limiting values. Second, these limiting values must remain the same in all partial sequences which may be selected from the original one in an arbitrary way.[4]

By "an arbitrary way," Mises means one that does not look at the values to be selected, but only at their position in the sequence of selections. For instance, if taking successive values from a sequence produces a result that contains an even mix of all the numbers from 1 to 10, but taking every fourth value produces only even numbers, then the sequence is not random.

Mathematically, this is fine, but it describes only the result, not the cause. What is the precondition of randomness? What is it that makes the relative frequencies in a selection tend toward a limit in some cases, but not in others?

Roughly speaking, randomness corresponds to ignorance. When we flip a coin or roll a die, we do not know how it will come up. When we pull a marble blindly from a box, we do not know which marble we will select. Moreover, no amount of testing will provide us with the knowledge necessary to predict the next outcome.

But there are cases which satisfy this criterion, yet cannot properly be considered random; rather, they combine random and nonrandom elements. Consider, for instance, an experiment that produces a series of numbers, each larger than the previous one by an unpredictable amount. The next number is always unpredictable, but the sequence does not satisfy Mises' criteria; the distribution of numbers does not converge to a stable proportion. In this case, the amount of the increase might be random, but the sequence as such is not fully random.

Hence, the criterion of ignorance must be made more specific. If it is possible to make even statistical predictions about elements of a sequence, and these predictions do not apply equally to all the elements, then the sequence is not random with respect to our knowledge. In terms of what makes the sequence random, we must be unable to identify the operation of any causal factor that applies to some elements of the sequence in a way different from others.

This is the key to randomness: complete interchangeability of causal factors, within the limits of our knowledge. If we can identify a factor that affects some elements of a sequence differently, then it is not random. Moreover, even if we can observe the effect of that factor without knowing what it is, the outcomes we are observing are not random. The phrase "within the limits of our knowledge" is vital. Causal factors are not interchangeable in reality; each event that occurs has its own unique set of causes. If we fully understand how the causes of an event operate, then it ceases to be random; we can in principle predict its outcome. (This applies even in the case of free will; the causes are not constrained to produce a particular outcome, but given the way the cause operates, i.e., the choices a person makes, the outcome, the resulting action, is fully predictable.)

Rand characterized many issues of knowledge by reference to the trichotomy of the subjective, the intrinsic, and the objective. Intrinsic knowledge is that which is automatically imposed on a mind from without, apart from any means of perception. Plato's theory of Forms and Descartes's theory of innate ideas are examples of intrinsicism. Subjective knowledge is that which is created, and not merely recognized, by the mind. The most extreme form of subjectivism is solipsism; faith, in the sense of belief without a reason, is also a form of subjectivism. In contrast with both of these, Rand's concept of the objective is the grasping of existence, which is independent of consciousness, through a means which depends on the form of the apprehending consciousness. This, Rand stated, is the only valid characterization of knowledge. A rose is red, for example, not because its intrinsic redness imposes itself on our awareness, nor merely because we say it is red, but because the way we are built responds to certain wavelengths of light in a certain way, which we call red. The cause of our perception is a fact of reality, but the mode in which we perceive it is the result of the kind of consciousness we have.

An extension of this concept, which Rand and Peikoff implied but did not, so far as I know, make explicit, is the characterization of phenomena as intrinsic, subjective, or objective. Anything which pertains to existence alone, apart from the way it is perceived by consciousness, is intrinsic; what pertains to consciousness alone is subjective; what pertains to existence as perceived by consciousness is objective. We have to be careful in using terms this way, since the intrinsic is what would normally be called "objective reality." These are not three different kinds of reality, but a distinction between the observer, the observable, and the interrelationship of the two. The word "objective," when used in this trichotomy, refers to the object of awareness.

Characterizing probability as objective in this sense gets us out of any potential confusion about its nature. If it were intrinsic, random events would be caused by probability itself. This would really mean that they are causeless, hopping about for no reason at all. If it were subjective, statements of probability would be expressions of hope or expectation and nothing more. But probabilities depend both on the nature of the phenomenon and on the knowledge of the person assessing the probabilities.

Mises tends to regard probability as intrinsic, stating that "for a given pair of dice (including of course the total setup) the probability of a 'double 6' is a characteristic property,

a physical constant belonging to the experiment as a whole and comparable with all its other physical properties." But in fact, the physical constants pertaining to any one roll of the dice permits only one outcome, not a limiting proportion of relative frequencies. If the association between the way the dice are thrown and the outcome of the roll were sufficiently well understood, the outcomes would be predictable rather than random.

Probability may appear to be intrinsic when only the sequence is presented, divorced from all information about its origin. In these cases, all that can be done is to apply Mises' criteria But even in this case, the nonexistence of partial sequences that violate randomness cannot be absolutely established. Some of the partial sequences will have different relative frequencies of attributes over arbitrarily long, but finite, runs. There is, for example, going to be some sequence in a string of random digits from I to 10 that will contain nothing but 3's for a large but finite interval. These may be very rare, but they are statistically possible. To say that they are "improbable," though, in order to exclude them from consideration, is to beg the question.

Computer programs commonly use "pseudorandom" sequences of numbers. These are numbers which are successively generated by an algorithm, and hence subject to the selection of partial sequences that violate randomness by basing the selection on that algorithm; but for all purposes for which the numbers will be used, they are random. Pseudorandom numbers actually violate Mises' first criterion in a fundamental way: if the sequence is carried out far enough, it will repeat itself. Nonetheless, they satisfy the epistemological criterion of randomness as long as the sequence is not carried out to the point where a pattern begins to show.

If a sequence of pseudorandom numbers is restarted with the same "seed," the sequence will repeat itself. This does no further violence to the mathematical criteria for randomness, but it does violate the epistemological criterion; it renders the second run completely predictable up to the length of the first run. This is actually a useful characteristic; it permits programs that involve "random" elements to be run repeatedly in exactly the same way, thus making the identification of bugs easier. Such a case is an example of a sequence being random from one perspective but not from another; the programmer intentionally denies himself knowledge of the sequence the first time he uses it, but then he can know what its outcome will be if he needs to re-examine what happens.

Popper regards the view of probability as a measure of partial knowledge as a "subjective" view. If we understand the difference between that which pertains purely to consciousness and that which pertains to the relationship of existence to consciousness. Probability is not an attribute of things in themselves, but of their relationship to our knowledge. It is important not to confuse knowledge with belief, as Popper does at one point[5]; knowledge is an objective phenomenon, which forms the criterion for what is measurable in an experiment, whereas mere belief plays no pan in establishing what is available for measurement. He admits the role of knowledge when he allows the concept of probability relative to a given set of information.[6]

In discussing quantum physics, Popper endorses the idea of probability as a propensity in each of the elements of the system. But when we say that something has multiple propensities, all that we can really mean is that it may act in one of several different ways, and we do not know which in any given case, although we can characterize the outcomes statistically over a large number of cases. The causal factors in each case are interchangeable, within the limits of our knowledge, with those in the other cases; it is this fact which makes the propensities probabilistic. Popper's unwillingness to admit lack of knowledge as an essential feature of probability apparently arises from a justified fear of admitting subjectivism into science; but

when we limit the specific kind of ignorance to the distinction among causal factors in a class of events, this fear is unnecessary.

The epistemological criterion for randomness explains why it is possible to determine the probability of an event, in some cases, without first engaging in many repeated trials for that particular event. Suppose, for example, someone gave you a piece of plastic in the form of a dodecahedron, and you verified that it was perfectly symmetrical and balanced. You might never have seen such an object before (particularly if you've never played "Dungeons and Dragons"), but you can state with confidence that if this unusual die is rolled on a table, the probability of any particular side coming up is 1/12. You know this because when it is rolled, some side has to come up, but there is no way of identifying any aspect of one side that distinguishes the circumstances in which it will come up from the circumstances in which any other side will come up. If the die were unbalanced, or if some of the corners were rounded, then there would be such factors; it might not be possible to tell what probabilities would be obtained without actually rolling the die many times, but since the causal factors contributing to anyone side coming up are not identical to the factors for the other sides, there is no reason to be confident that the probability is still 1/12.

Mises correctly points out that the conclusion that all sides of the die are equally likely to come up is a conclusion based on experience, not a purely a priori conclusion. It requires experience to know what aspects of the manufacture of a die will affect the probabilities, and which (for instance, putting different numbers on the different sides) will not have a significant effect. All knowledge is ultimately based on experience. But it is important to note that experience other than knowledge of the past distribution of outcomes for the event in question is suitable for establishing probabilities. Without accepting this fact, every new venture that involved an uncertainty would be a leap into the unknown.

In cases where repeated trials are not available, estimates of probability may be very crude, but they are still valid if we do not attribute to them a precision that they lack. Suppose, for instance, that there are two teams in a sports league, the Lions and the Lambs. These two teams have played the same opponents in ten games; the Lions have always won, and the Lambs have always lost. This information is not sufficient for assigning a probability to the outcome of an upcoming game between the Lions and the Lambs, but it is sufficient for saying that the probability is greater than 1/2 that the Lions will win. In a case such as this, we are hypothesizing what would happen in "repeated trials" of contests between teams with similar respective records; without actually running these trials (and more precisely specifying the instances that are admitted as trials), we do not know exactly what would happen, but we know in a general way.

Probabilities are not intrinsic properties, but the result of the interplay of causal factors of which the observer lacks full knowledge. Hence, probability is a relativistic phenomenon; a probability can only be defined with respect to a given frame of knowledge. For instance, a person tossing a coin into the air must regard the probability of its coming up heads as 1/2; but we can imagine a computer that is able to analyze the dynamics of the situation the moment the coin goes into the air and arrive at a certainty that it will come up tails, or at a probability other than 1/2.

Bohm[7] objects to the idea of probability as a measure of partial ignorance on two grounds. The first is that ignorance does not allow any kind of predictions at all to be made, not even about long-term relative frequencies; the second is that the existence of these relative frequencies does not depend on our knowledge or ignorance of the causal factors involved. I

have already dealt with the first objection by indicating that it is not complete ignorance, but a special kind of incomplete knowledge, which is necessary for establishing probabilities.

The second point involved a subtle confusion. Bohm gives the example of rolling a die; even if we are able to analyze each roll from its initial conditions and thus predict each separate outcome, the distribution of outcomes over the long run will still be the same as those predicted by the laws of probability. Hence, concludes Bohm, the applicability of those laws "depends only on the objective existence of certain regularities that are characteristic of the systems and processes under discussion, regularities which imply that the long run or average behavior in a large aggregate of objects or events is approximately independent of the precise details that determine exactly what will happen in each individual case."

What Bohm fails to recognize is that treating these events as members of an aggregate or collective implies treating them in a certain way, which is to (temporarily) close our eyes to the specific differences among them and treat them as indistinguishable. If we had the power to determine the outcome of a die roll by looking at the initial conditions, we could pick out just those rolls that would come up "four" and no others; or we could bias the sample in any number of less blatant ways, even ways we were not conscious of. It is only by choosing not to regard the precise conditions of each roll, or by being unable to do so, that we can sensibly deal in probability.

An example of an event that can be either predictable or random in different contexts is the selection of successive digits of π. Prior to the calculation of the next digit, it is equally likely to be any of the ten decimal digits; this is true not a priori, nor as a result of pure ignorance, but because the expansion of π has in fact been shown to have an equal distribution of all the digits as far as it has ever been measured. From my current standpoint, I can say there is a probability of 0.1 that the four hundredth digit of the expansion of π is a 7; but once I calculated it or looked it up, the probability would be 1 or O.

Considering probabilities as predictions of long-term distributions makes the concept vacuous; for the initial data from which probabilities are obtained are just those long-term distributions. A law whose conclusion is that its initial data are valid says nothing. Rather, the laws of probability tell us how we may extrapolate from those initial data under circumstances of partial ignorance.

It is a mistake to think of probabilities as existing prior to our discovery of them, since they are measures of partial know ledge. In the case of the Lions-Lambs game, there isn't some particular probability, such as 0.7843, that the Lions will win; saying that the probability is greater than 1/2 is the most exact statement possible. It is true that probabilities which have exact values are much more convenient for mathematical operations; but the convenience of a model that deals only in exact probabilities should not blind us to the fact that in a given frame of knowledge, probabilities may be only approximate. In some cases, all that we may be able to say is that the probability of an event is nonzero. This should not be taken to mean that there is a definite probability, which we do not know, but simply as an assertion of the possibility of the event. The statement of probability in this case means that, given all cases of sporting events that meet the same criteria as this one, we can be confident that the team that has always won will usually beat the team that has always lost.

Since probabilities are not intrinsic, they are not causes. A coin does not come up heads half the time because the probability that it will do so it 0.5, but because it is tossed half the

time in a way that makes it come up heads. Entities, not anyone's state of knowledge about them, are causes.

Mises argues that "we can find formulations of statistical propositions that are in accord with the 'law of cause and effect'. If a 'double-six' appears on the average once in 36 casts of two dice, we can say that the 'cause' of this regularity is the fact that each die falls equally often on each of its six sides ..., Such statements may be countered by saying that no cause can be indicated for the single result in either the game of dice or Galton's Board, or that fluctuations in short sequences of observations are not traceable to special causes. However, in the case of the law of intertion we agreed to dispense with a cause for the displacements (required under a more naive conception) and were satisfied with having just a cause for the accelerations .... the principle of causality is subject to change and it will adjust itself to the requirements of physics."[8] This really means changing the principle of causality to fit the premise that probabilities are causes, even if it leaves previously explainable events unexplained.

It is necessary to be careful here in order to avoid falling into a Kantian kind of skepticism. One could argue that all properties of an entity are measured according to our state of knowledge, and therefore describe only our knowledge, not the entity itself. But the difference is that properties such as mass and size are described in terms of our knowledge about them; probability is a quantification of knowledge, which must in turn be based on more fundamental knowledge of the events whose probability is being measured. Mass, for instance, can be characterized in a given experiment as the property that counterbalances a certain quantity of weights on a scale; its effect must be measured according to our knowledge, but the way it produces the effect is independent of anyone's consciousness. Probability, on the other hand, is dependent on a frame of knowledge; changing the known factors which may be taken into account in an experiment can change the probabilities.

In fact, the mass of an entity is not a cause, either; the cause is the entity. Mass and other attributes are potentialities for causal relationships, not causes; they are aspects which characterize the behavior of the entity over a large number of different situations.

Probabilities, on the other hand, not only are not causes, but are not potentialities of any entity. They are characterizations of the distribution of potentialities among different entities, which mayor may not be exactly alike in any respect. A probability is the result of several causal factors which have not been separated. Given the statement that an event has a certain probability of occurring, one must always ask what will cause it to occur, and what will cause it not to occur. The answer may not immediately be available, but this does not justify falling back to the statement that the probability is the cause.

Regarding probability itself as causal is a particularly significant instance of the black box fallacy. Probability describes the behavior of a system under specified circumstances (and with some of the circumstances left necessarily unspecified). It does not give a reason why the system provides one of several probable outcomes rather than another; the reason may be accessible on more detailed study of the system in the particular case, or we may simply not be able to tell why the particular outcome occurred. From the black-box view, though, there is nothing more to be said; the probabilities are the description of the system, and the only cause to which the outcome can be attributed.

To say that a certain number of particles will decay because the particle has a certain probability of decaying in a second is to pretend that a description of uncertainty is an

explanation. It is no longer fashionable to explain the unexplained by reference to divine causes, but saying that probabilities are fundamental causes serves the same function. Baron von Holbach wrote in 1770:

> If a faithful account was rendered of Man's ideas upon Divinity, he would be obliged to acknowledge, that for the most part the word "gods" has been used to express the concealed, remote, unknown causes of the effects he witnessed; that he applies this term when the spring of the natural, the source of known causes, ceases to be visible: as soon as he loses the thread of these causes, or as soon as his mind can no longer follow the chain, he solves the difficulty, terminates his research, by ascribing it to his gods.[9]

Substitute "probability" for "gods," and the statement remains equally valid.

Fortunately, neither God nor probability has any power to strike people down. Bugaboos about the stability of the macroscopic world being due to chance can be put to rest along with the notion that probability is a cause. One version of this claim states that there is a minuscule but nonzero probability that all the oxygen molecules in your vicinity will simultaneously rush away from you, leaving you to choke. Probabilities state only what the relative frequency of outcomes will be in a large number of cases. To go from this statement to the statement that anyone molecule of oxygen may rush away from your vicinity, since it has a certain probability of doing so, is legitimate if the context is kept in mind. But to then conclude that all the atoms can simultaneously move away from you ignores the original context. Probabilities apply to anonymous parts of the group, not to the group as a whole.

It is true that if the probabilities of what all the molecules do were really independent, then this minuscule probability would exist (although it would still be nothing to worry about). If a million pennies are flipped, there is a probability of $1/2^{1,000,000}$ that they all will come up heads. But air molecules are definitely not independent; any mass movement of air molecules together in one direction would violate the law of conservation of momentum.

Similarly, life insurance statistics indicate a certain probability that each person will die within the next year. But it is not legitimate to multiply all these probabilities together and claim to have determined a probability that *everyone* will die in the next year. The individual probabilities are arrived at by the actuaries' knowledge that some people have health conditions that will kill them within a year, that some people will be standing in the path of cars that go out of control (a condition that would be very unlikely for the last survivor on earth), and that other factors are similarly going to kill some people but not others, without knowing which are the people who will be killed. A model in which probabilities are independent may work very well in its normal context, but extending it to cases of negligible probability is a very likely to make the model break down, just as the billiard-ball model breaks down at temperatures high enough to melt the balls.

Even where it is not possible to point at a specific factor that introduces dependencies among the events in a system, it is not legitimate to use probabilities derived from the normal behavior of a system as an argument that the system can act in a way it has never been known to act. Probabilities are a measure of the way a large number of similar events occur in the long run; arguing from probability in this manner would be arguing that since the system has always been observed to act one way, it may sometimes act in a different way.

In order to establish that the system can act differently, it is necessary to understand the nature of the events that make up the system. Doing this may reveal that it is in fact possible (though unlikely) for all the elements of the system to do something unusual at once, thus creating anomalies such as water freezing at 80 degrees; but then the knowledge which is being applied goes beyond probabilities.

The use of probability in models may be the result of an intentional compromise between structural correspondence and complexity; it may also be the result of not having the information available to go below the level of probabilities. In either case, the probabilistic aspect of the model simulates, over a large number of trials, the effect of causes which are impossible or inconvenient to include explicitly in the model. The statistical approach provides a way to incorporate partial ignorance into the model without incorporating any arbitrary assumptions about the unknowns.

The conclusions we can reach about probability fall into the general pattern of cautions in the use of models. Probability can be very valuable; it can warn you of the folly of investing money in state lottery tickets, to give one example. But it is essentially a method of modeling, and as such must be applied in its proper context and not taken as something intrinsic to what is being measured. The issue of probability will come up again in the discussion of quantum physics.

---

1 Mises, p. II.
2 Ibid., p. 17-18.
3 Russell, *Human Knowledge: Its Scope and Limits*, p. 364.
4 Mises, p. 24-25.
5 Popper, *Quantum Theory and the Schism in Physics*, p. 67.
6 Ibid., p. 78.
7 Bohm, p. 26-7.
8 Mises, p. 21 I.
9 Paul Heinrich Dietrich, Baron von Holbach, *Système de la Nature*, London 1770, quoted in Sagan.

V. Gödel and the Unknowable

Are there things that we cannot know, even in principle? Are there assertions one can make which can never be shown to be correct or incorrect? Answering this question in the affirmative would seem extremely dangerous, since we can never know what tools might be discovered in the future to uncover what is a mystery today. But it is widely held today that there are meaningful statements which can never be identifIed as true or false, even if all the relevant data are available, and that the existence of statements of this kind has been mathematically proven. The proof is known as Gödel's Theorem for its author, the mathematician Kurt Gödel.

A certain amount of background is necessary to understand why Gödel took the approach he did, and why his theorem is so widely regarded as profoundly signifIcant. Around the beginning of the twentieth century, mathematics underwent a major shift in approach. Previously, mathematics was regarded by mathematicians primarily as a tool for solving practical problems. This is not to say that mathematicians did not delight in the inner logic of their systems apart from any use they might have had; but they expected their concepts to have referents in reality, and therefore to have some applicability. To take a simple example, the concept of "natural numbers" (integers greater than or equal to zero) was regarded as referring to quantities of discrete objects.

Over a period of time, this approach came to be replaced with another one, in the name of rigor. The landmark work in establishing this new approach was *Principia Mathematica*, by Bertrand Russell and Alfred North Whitehead. The key word in the new approach is formalism. Russell and Whitehead sought to strip mathematics of all ambiguity, of all use of unstated assumptions, by treating mathematics in terms of formal systems.

A formal system consists of nothing but symbols, with rules for producing theorems out of them. A theorem, in this context, is not a statement of truth, but simply a string of symbols which satisfIes the stated requirements. The starting point of the system is its axioms, which are strings of symbols that are asserted to be theorems, and its rules of inference, which allow producing new theorems from old ones. For instance, a system might consist of just the symbols 1 and 0. It has, let us say, just one axiom:

1

and one rule of inference:

If S is a theorem, then S0 is a theorem.

This leads to the theorems 1, 10, 100, and so on.

What do these theorems say? In the usual view of formal systems, they say nothing at all; they are not true or false, but are simply the results of applying arbitrary rules. If this were true, though, these systems would be nothing more than game-playing. Each theorem is actually a statement of fact; what it says is, "In a system for which the axioms of this formalism correspond to members of some class, and in which the rules of production corresponds to valid methods of generating new members, this theorem corresponds to a

member of the same class." Most often the class consists of facts in some mathematical theory, but this is not the only possibility. The axioms might correspond to the most fundamental members of a set (for instance, powers of two), and the production rules could be the method of obtaining additional members of the set.

The referents of a formalism may be very abstract. One might construct a system, for instance, solely for the purpose of demonstrating something about the nature of formal systems. In this case, the theorems refer to representative members of a class with a certain property, some of which correspond to members of some mathematical class. As long as the system serves some legitimate purpose in mathematical theory, it is possible to trace it to some referents. If it does not, then it is pure game-playing.

It is easy to confuse the terminology of formal systems with the normal use of this terminology. An axiom normally means a statement of a self-evident truth, and a theorem a statement of a provable truth. Formal systems may be used to characterize relationships of truths, but they do not have to. If models are grounded in this approach, it becomes necessary to take extra care that the entire structure of the model is not grounded in midair. It is far too easy to construct a formalism that seems to describe a real situation but actually does not, since using the system provides no feedback to its referents.

Symbolic logic, as covered in Chapter 3, is an example of the formal treatment of symbols. The rules of symbolic logic are not asserted as truths about reality, but only as rules. Formal grammars, which are often used to describe the syntax of programming languages, are another example. In each case, there is a referent, but only what is made into an explicit production rule is admitted into the formalism.

Formal systems are high-level abstractions. The basis of mathematics is not formalisms, but facts of reality treated in terms of measurement. One of the simplest concepts in mathematics is natural numbers, or positive integers. This concept is arrived at by abstracting quantity away from groups. In treating three trees, three houses, or three children, we can disregard all attributes of the group except quantity, and arrive at the concept "three."

There is a potential infinity of natural numbers, but we can obtain them all from a finite number of concepts by using an appropriate system of notation. Decimal notation requires only the concepts of "zero" through "nine," along with the operations of multiplication, exponentiation (of the number 10), and addition, to be able to express any number.

Negative numbers are only slightly more abstract; they deal with a quantity which is greater than or less than a given reference point, called zero. Real numbers (such as 2.56 or the square root of 2) arise from phenomena that do not exist in discrete quantities (at least on the level at which they are being measured), but can assume a continuous range of values.

Some mathematical concepts are abstractions from previous concepts. "Imaginary" numbers, such as the square root of -1, were created in an effort to make the treatment of polynomials more consistent; in this case, the concept was created first, and physical interpretations of it (e.g., for characterizing waves) were discovered later.

Formal systems are more abstract in that they renounce any dependence on the phenomena which they describe, or which any other system on which they are based describes. A system is interpreted as describing the phenomena from which it is abstracted, but the interpretation is not part of the system. The creator of such a system usually begins with a mathematical

system that is conceptually based, but then reduces it to the minimal set of axioms and production rules that will yield a system which is isomorphic to the conceptual system. In this minimal system, facts which are axiomatic in the conceptual system may require proof; for instance, it is not necessarily an axiom in a formal number theory that 1 is the smallest positive integer.

A single formal system may have more than one interpretation. The discovery that two different and apparently unrelated mathematical theories correspond to the same formal system can be very valuable, since it shows that the two theories are isomorphic to one another, and that work which has been done in one can be carried over to the other. Discovering such isomorphisms is one of the values of formalizing the systems.

Euclidean geometry is not a formal system, although it contains many features of one. Its axioms are not simply organizations of symbols, but are intended as descriptions of reality. (One specific difference is that constructions are permitted in Euclid's geometry, which they would not be in a system that was fully abstracted to symbols and production rules.) Euclid's system has been heavily criticized because it includes one axiom, the parallel postulate, which can be replaced by alternative axioms that lead to self-consistent geometries; but this criticism does not consider the fact that Euclid's purpose was to mathematically describe the observed properties of plane surfaces, not to construct the most general or elegant system of axioms and deductions possible.

While formal systems are a valuable tool in mathematics, it is important to recognize their derivative nature. Mathematics is not the science of formalisms, but of measurement; formalisms are a means of gaining insight into the possibilities of measurement.

Gödel's Theorem

Gödel's Theorem takes us to a yet higher level of abstraction; it is a statement about the properties of formal systems. It applies to any system which meets certain requirements, of which the most noteworthy is the ability to describe itself.

How can a system describe itself? The trick lies in a method of encoding which Gödel devised, and which is now called Gödel numbering. The idea behind Gödel numbering is straightforward enough, particularly in today's computer age: it is that any statement in a formal system can be expressed, or encoded, as a number. ASCII, the code which is commonly used to store and transmit text files, provides one basis for Gödel numbering: a statement can be encoded as the concatenation of the ASCII codes for the characters in it.

Once we have a method for encoding statements as numbers, we can in principle reduce a formal system to arithmetic. Certain Gödel numbers are the axioms of the system; the rules of production of the system can be translated to equivalent arithmetic operations which produce numbers that are theorems of the system. The numbers involved tend to be huge and unwieldy, but the point is not that we would actually want to work with formal systems this way; the point is that if we can do something with a formal system, we can do the same thing with arithmetic.

This is an issue of the interpretation of the system, of course; the statements of a formal system, as such, are not restricted to a particular referent. But it allows us to derive theorems in a formal system in a purely mechanical way; we can encode a theorem as a number,

perform the operations on it which correspond to the rules of the system, and translate the resulting number back into a statement which is guaranteed to be a theorem.

Not all numbers have to be Gödel numbers, just as not all combinations of letters in English form words. But any string made up of the permissible symbols of the system can be encoded by a Gödel number.

The theory of natural numbers can itself be described in a formal system. Natural number theory deals with the operations of ordinary arithmetic of natural numbers; it encompasses logic and quantification, so that it is able to deal with expressions of the form "There exists a number which satisfies this logical combination of arithmetic relationships" or "All numbers satisfy this logical combination of arithmetic relationships." If a number in an expression in this theory is interpreted as a Gödel number, then the expression can be interpreted as a predicate asserting the truth or falsehood of the expression which is that Gödel number stands for. Hence, theorems of a formal system which is isomorphic to number theory can be interpreted as statements about other theorems of the system.

It is not possible for an expression to refer directly to itself by Gödel number, since the number of an expression is larger than any number which it contains. However, if an expression possesses a free variable (one whose value is not specified by the expression), the Gödel number of that very expression can be substituted for the free variable. This yields a new expression, with a different Gödel number. This process of substitution is itself a function which can be expressed in number theory.

We can define a predicate **A(x,n)**, which is TRUE if and only if **x** is the Gödel number of the expression that results from substituting **n** into the statement whose Gödel number is **n**.

Entire proofs are also strings of symbols, and can be encoded by Gödel's scheme. Let us also define the predicate **P(m,n)**, which is TRUE if and only if **m** is the Gödel number of a proof of the expression whose Gödel number is n. This predicate provides us with a means of asserting that there is a proof for a given expression.

Now consider the following expression:

(1) $\neg\exists m{:}\exists n{:}P(m,n)\wedge A(g,n)$

This asserts that for some **n**, there is no proof of the expression whose Gödel number is **n**, and that **g** is the Gödel number of the expression which results from substituting n into that unprovable expression. Let us call the Gödel number of (1) **u**, and let us substitute **u** for **g** in (1). Then we obtain the expression:

(2) $\neg\exists m{:}\exists n{:}P(m,n)\wedge A(u,n)$

This expression has no free variables, hence we can ask if it is a theorem. Suppose that it is one; then in its interpretation it is true. This means that there is no number **m** which is the Gödel number of a proof that substituting **u** into (1) yields a theorem. But what we get when we substitute U into (1) is (2). Hence assuming that (2) is a theorem leads to a the conclusion that it is not a theorem; the assumption must be incorrect. But if (2) is not a theorem, then what it states is true! Hence we have an expression which asserts a truth, but which is unprovable in the system.

This does not imply any inconsistency in the system; no contradiction is arrived at by saying that (2) is a true statement. The system is not *inconsistent*, as it would be if it embraced a contradiction; but it is *incomplete*, in that it does not contain a means of establishing the truth or falsehood of every one of its own statements. What Gödel's Theorem shows is that every consistent formal system which is powerful enough to represent the statements of number theory (or equivalently, the operations of a Turing machine) must be incomplete.

This has been a very generalized discussion of Gödel' s Theorem. Its purpose is not to demonstrate the theorem's validity with any rigor, but only to show the key points of the reasoning behind it. (It is necessary, for example, to pin down the concept of substitution more precisely, and to establish that A and P are predicates which can be defined in the system.) A more technical presentation, but one which is still accessible to the reader with a good mathematical background, can be found in Hofstadter's *Gödel, Escher, Bach*.[1]

Whether presented briefly or completely, though, the demonstration of this theorem may leave the reader with a bewildered feeling. Does such a bizarre, self-referential theorem have any significance, whether it can be proved or not? What gaps in our understanding that can never be filled does it imply?

The Barber's Razor

In fact, Gödel's Theorem represents no loss at all to our understanding of reality. The theorem which cannot be proved lies in limbo simply because it has no meaning. What Gödel's Theorem really proves is that in any propositional system with sufficient power, there is no mathematical way to automatically separate the gibberish from the meaningful statements.

The statement "This statement cannot be proved" is not very different from the statement sometimes known as Russell's Paradox: "This statement is false." The same considerations apply to both. It is obvious that Russell's Paradox is meaningless, because both assuming that it is true and assuming that it is false lead to a contradiction. It is a kind of linguistic Möbius strip; following one side of it inevitably leads to being on the other side, then returning to the first side again in an endless loop.

It is less obvious that "This statement cannot be proved" is meaningless, since asserting that it is true but unprovable does not lead to a contradiction. But it has the same essential problem as Russell's Paradox; its meaning is not fixed but depends upon itself. The loop is more subtle: If it is false, we have a contradiction (since only true statements can be proved), so it must be true. But that proves that it must be true. However, there is no paradox because that proof is not admitted into the system. We could construct a larger system that admitted that proof; but then it would in tum contain a self-referential theorem that could not be proved without stepping out into an even larger system, and so on ad infinitum.

Hence, it would appear that any such statement is, in a sense, provable and therefore true. But asserting its truth involves an infinite regress. For unless we assume such a hierarchy of systems, we are faced with a contradiction: the statement is not provable, yet we have proved it. To establish that it is true, we must assume the existence of an infinity of systems, each of which establishes the truth of the theorem in the next lower system. No matter how many levels we go, we are stuck with either incompleteness (a theorem which cannot be proved) or inconsistency (the proof of such a theorem).

J. R. Lucas[2] has argued that Gödel's theorem demonstrates the superiority of the human mind to any machine because it shows that no machine can establish the truth of something which we know to be true:

> Gödel's theorem must apply to cybernetical machines, because it is of the essence of being a machine, that it should be a concrete instantiation of a formal system. It follows that given any machine which is consistent and capable of doing simple arithmetic, there is a formula which it is incapable of producing as being true—i.e., the formula is unprovable-in-the-system—but which we can see to be true. It follows that no machine can be a complete or adequate model of the mind, that minds are essentially different from machines.

This may sound like a triumph for humanity, but it is actually a very meager one. Does the superiority of man over machine consist of man's ability to make greater sense of self-referential sentences? If so, who would care? Would you prefer to talk to a human rather than a computer because he sees no problem with asserting that his own assertion is unprovable? My own reaction would be that throwing up one's hands at such a statement—as a machine must, in effect, do—is a better sign of good judgment than claiming to understand it.

As an alternative to assuming that humans are superior to machines, Lucas points out, we could assume that humans are inconsistent; for inconsistent systems are not subject to Gödel's Theorem. He rejects this conclusion because humans are merely capable of inconsistency, not fundamentally inconsistent. This argument is correct; claiming that humans were fundamentally inconsistent would imply that we were incapable of avoiding contradictions, and would lead directly to blanket skepticism.

But claiming to know that the unprovable theorem is true involves accepting an inconsistency, though not an obvious one; for the reasoning that leads to the conclusion that it is true amounts to a proof. Again, the infinite hierarchy of systems is the only escape from this inconsistency.

There is a certain difficulty with simply dismissing the statement "This statement is unprovable" as self-referential, but it is not an insurmountable difficulty. The problem is that excluding all self-references in a crude way eliminates too much which is valuable. Any statement which talks about all statements in the English language, for instance, is self-referential. (In particular, the preceding statement is self-referential.) There are many self-referential statements which make perfect sense; for instance, a letter might contain the desperate avowal that "Every sentence in this letter is true." In order to dismiss statements which deny there own provability, we need to distinguish valid from invalid self-reference.

A valid self-reference is one in which the fact exists independently of its assertion. For instance, a sentence that makes a statement about its lexical structure, about the person who is uttering it, about the pen with which it was written, or about the time at which it was uttered is completely valid, because its truth can be established by considering the facts.

"This statement contains five words." If you knew only French rather than English, you would still be able to tell that there are five words in that statement; you would not have to know what it meant.

"Everything in this paragraph is written in English." This example is more subtle; but again, it is possible (in fact, easy) to establish that the statement is in English without first having to determine whether it is true.

"Every sentence in this letter is true." Here we are in still deeper trouble; for in order to establish that it is true, we must first establish that it is true. This kind of statement actually needs to be interpreted in a non-literal way; that is, as an assertion that every other statement in the letter is true. If you received a letter that contained only that sentence, you would throw it away as a joke; it would, in fact, be telling you nothing.

Yes, the letter could contain the sentence "Every sentence in this letter is true" twice for emphasis. But if the writer is trying so hard to make a point, it had better be a point other than the reference of the two sentences to each other. One can construct word games involving mutual references, or try to read such word games into ordinary writings; but doing that means abandoning any concern for whether the words actually mean anything.

A favorite word game involves the statement, uttered by a Cretan: "All Cretans are liars." This is often taken to be equivalent to Russell's Paradox, but doing so requires interpreting the word "liar" in an artificial way, as someone who absolutely never tells the truth. Otherwise "All Cretans are liars" could be the liar's first truthful statement, without ceasing to make him a liar. This kind of artificiality is typical of nearly all the paradoxes of self-reference.

A different kind of self-reference is found in the barber who shaves all those, and only those, who do not shave themselves. (Does he shave himself? Either answer leads to a contradiction.) Hofstadter provides an amusing variant of the same puzzle in suggesting that if God helps those who help themselves, perhaps the devil helps precisely those who do not help themselves. (Does the devil help himself?) In either form, this is not a self-referential statement; rather, it defines an action which is performed on a criterion based on the action. Hence we are safely planted in reality (in the case of the barber), and the solution to the problem is that any barber who tried to set out on such a course would discover its impossibility.

The invalid kind of self-reference occurs when the actual meaning of a statement is self-referential; that is, when its meaning depends upon its meaning. In such cases, the statement may be consistent (as in "This sentence is true") or inconsistent (as in "This sentence is false"), but in either case it is grounded only in itself, and floats free of reality.

The test is whether any circularity is entailed in establishing the meaning of the sentence. Certain statements may appear to be self-referential in meaning, but in fact only refer to general aspects of themselves, which can be considered apart from the meaning of the particular statement. For example: "True statements never contradict one another." That statement does say something about its own meaning, but what it says is a consequence of the nature of truth, and can be established as true prior to considering that statement.

There is a parallel to the barber if we consider statements about "all statements which do not refer to themselves." Do such statements refer to themselves, or not? Saying either yes or no leads to an inconsistency, so we must assume there is something wrong with those statements in the first place. Again, the problem is with circularity of meaning; we must first determine is such a statement is self-referential before we can tell ·if it is self-referential. This suggests that, in this case, the attempt to apply the concept of self-reference is invalid.

Russell's solution is to consider the statement to be of a higher "type" than the statement to which it refers; in other words, to stand outside its universe of reference. In this way, the statement does not refer to itself; but it is not one of the statements in its universe of discourse,

so it does not become one of the set of non-self-referential statements in question. But this solution is artificial, since it requires an infinite hierarchy of types.

A better approach is to realize that any concept must be applied in a particular context; outside that context, the concept must be excluded because its application makes no sense. Thus, the concept of self-reference applies only to statements whose referents can be identified as themselves. Hence, any statement of the form "Statements which do not refer to themselves are x" does not refer to itself, since its self-reference cannot be established in a consistent way. This does not lead directly to the contradictory conclusion that it does refer to itself, for its referents must be established prior to reaching that conclusion. Permitting the referents to be changed retroactively or recursively would refer to itself would make its meaning depend upon itself; as we have seen, this would void the meaning of the whole statement.

The same approach applies to statements about "all statements which refer to themselves." Here the apparent problem is different; the conclusion that such statements are self-referential, as well as the conclusion that they are not self-referential, both lead to consistency, rather than inconsistency. This leads to a Gödel-like situation in which neither the assertion nor the denial of their self-reference can ever be proved. The analysis is the same as before: Since such statements must refer to self-reference where it can be established, they must be regarded as not referring to themselves.

Self-reference can be indirect rather than direct; several sentences may depend on one another's meanings in a loop. This type of construct may be harder to figure out, but the same considerations apply. For example, imagine a card with the following statements on its two sides:

Front: The statement on the other side is true.

Back: The statement on the other side is false.

Each sentence ultimately depends upon its own meaning for its meaning; thus the reader may find himself flipping the card over and over in a vain attempt to make sense of it.

The issue of self-referential meaning must be considered in context. The purpose of making a statement is to apply it to some intended set of referents; the statement should be formulated so as to characterize those referents in the clearest possible way. If someone intentionally puts together sentences in a way intended to generate a paradox, he may provide some amusing puzzles, but he is contributing little to anyone's knowledge. In serious discussion, statements are formed to apply to their referents, not in order to set a challenge regarding what their referents are.

A principle that limits cognitive redundancy is sometimes called a "razor." The best known example is Occam's Razor, which prohibits postulating entities without reason. The principle that the meaning of a statement may not be dependent on its own meaning is another razor; since paradoxes of self-reference are often typified by the barber mentioned earlier, we could call this principle "the Barber's Razor."

All this goes to show that self-reference in meaning is not a problem to worry about. This being the case, Gödel's theorem would appear to amount to more of a mathematical curiosity than a serious problem. But there is another aspect of the thought about it which must be

considered. That is the question of whether consciousness is itself equivalent to a formal system. This is the position that Hofstadter takes in regarding consciousness as a property of a system which possesses a "self-subsystem." In discussing this subsystem, Hofstadter says,

> A very important side effect of the self-subsystem is that it can play the role of "soul," in the following sense: in communicating constantly with the rest of the subsystems and symbols in the brain, it keeps track of what symbols are active, and in what way. This means that it has to have symbols for mental activity—in other words, symbols for symbols, and symbols for the actions of symbols.[3]

This being the case, the self-subsystem is full of self-references. Hence there must be aspects of the mind which are undecidable by the mind. From this basis, Hofstadter reaches a conclusion which is strange but fitting: "I have no doubt that a totally reductionistic but incomprehensible explanation of the brain exists."[4]

But this leads us into the question of whether the mind is itself simply a system which operates on a model of the world; and that must wait for a later chapter. For the present, we can be satisfied with these conclusions:

1. Gödel's Theorem is not a limitation on the decidability of statements about reality.

2. The establishment of a correspondence between formal systems and reality must be approached with care, since any sufficiently powerful system will necessarily include propositions that are self-referential.

3. Self-reference is not itself an obstacle to understanding, nor must it be excluded completely from discussion, so long as circularities in meaning are not admitted.

All of this is not to say that Gödel's Theorem is useless. Its value lies in that it demonstrates the limits of formalization. Such limits must be recognized when dealing in the realm of models. In addition, as a theorem which applies to a formal system, it can be interpreted in ways other than asserting the existence of undecidable statements; it has been used, for instance, in proving that there are no general algorithms for solving certain numeric problems.

Computers have proven difficult mathematical theorems by manipulating systems of propositions. And in these cases, it is well to know that there are dead ends from which the program might never be able to recover.

A close relative of Gödel's theorem is the "halting problem" for computers. The problem is this: for a given computer, write a program that can check programs and determine whether they will eventually terminate or not. In fact, no such program can work in all cases.

Proof: Take the program which checks for halting. Embed it in a subroutine in a larger program. In the larger program, perform the following upon return from the subroutine: If the subroutine reports that the tested program will halt, then go into an infinite loop; otherwise halt.

Now run this program on itself. If the subroutine reports that the program would halt, then it goes into an infinite loop. If it reports that the program will not halt, then it halts. In either case, the subroutine has reported an incorrect result. The analogy between this proof and the

proof of Gödel's Theorem is a close one. The problem exists because the poor program has no way of knowing that it is dealing with a self-reference.

Hence Gödel's Theorem does have consequences for models. It demonstrates that it is possible to construct a model on a perfectly reasonable basis, yet discover that some aspects of the model have no tie at all to reality. The distinction between the inapplicability of Gödel's Theorem to knowledge and its applicability to models is not one of humans vs. computers as such, but one of an approach that requires that all issues be tied to reality as opposed to a purely symbolic approach. The symbolic approach is powerful—we wouldn't have computers without it—but it does have its own pitfalls.

---

1 Hofstadter, *Gödel, Escher, Bach,* p. 438-451. While this book is flawed by a strong mystical element, its discussion of the mathematics involved is sound, and much of the book is entertaining to read.

2 J. R. Lucas, "Minds, Machines, and Gödel," *Philosophy*, Vol. XXXVI (1961), reprinted in Alan Ross Anderson, *Minds and Machines*, Prentice-Hall, 1964.

3 Hofstadter, op. cit., p. 387.

4 Ibid., p. 709.

VI. The Quantum Universe

Many of the greatest puzzles in the philosophy of science arise in accounting for the phenomena of quantum mechanics. It has been claimed that quantum physics shows that reality is essentially unreal, that the science is based in intrinsic probabilities and uncertainties and admits contradictions, or that the law of cause and effect is not universally valid. A number of popular books have been written promoting this theme, such as John Gribbin's *In Search of Schrödinger's Cat* and Gary Zukav's *The Dancing Wu Li Masters*. The former book cites claims by scientists, arguing from quantum physics, that "the whole universe may only owe its 'real' existence to the fact that it is observed by intelligent beings,"[1] and argues himself that "Objective reality does not have any place in our fundamental description of the universe."[2]

I do not address this subject as a physicist, but as a layman who has examined the subject. The quarrels I will raise are not with the validity of the experimental results and the descriptions of phenomena which have been discovered, but with what these results are claimed to say about reality. The area I will be addressing is one of philosophy, not physics, though still one that requires at least an elementary understanding of the physics involved in order to understand why people who have studied the subject are willing to accept such bizarre conclusions.

The central phenomenon of physics which causes so much confusion is the wave–particle duality. Physicists have found that on a very small scale, units of matter and energy act in some ways like waves and in other ways like particles, and that the two aspects of their nature interact in very strange ways.

This duality can be illustrated by a frequently cited experiment in the nature of light. Consider a light source which sends a beam through a slit to expose a film on the other side. The light, on passing through the slit, is diffracted in accordance with a wave description of its nature If the intensity of the light is reduced sufficiently, then the exposure of the film ceases to occur evenly; instead, it occurs in pinpoints where individual photons, particles of light, have struck. This suggests that light has a particle nature.

To make matters worse, suppose we change the one slot to two closely spaced slots. If we reduce the light intensity sufficiently, only one photon at a time will get through, and it must pass through one slot or the other -at least so we would think. But as large numbers of photons pass through over a long period of time, what builds up on the film is a pattern of light and dark bands, an interference pattern which is typical of wave phenomena.

(Interference occurs where two patterns of waves meet. For the purposes of this discussion, it is sufficient to thing of waves as an alternation of crests, or portions of positive amplitude, and troughs, or portions of negative amplitude. Where two crests or two troughs meet, they reinforce each other to create a higher crest or deeper trough; but where a crest meets a trough, they cancel each other out. The areas of high exposure correspond to both the crests and troughs; the unexposed areas occur where the two cancel one another out.)

This leads to a puzzle. If at any moment in time, a photon is passing through only one slot, how can anything from the other slot be interfering with it? If photons were simply particles, there should not be an interference pattern, but simply two overlapping areas of diffracted

light. It appears as if the photon passes through both slots in the form of a wave, then is all sucked into one place when it is absorbed by the film. Where the resulting particle appears is an apparently random phenomenon, with the probability being distributed according to the amplitude of the wave.

According to physics, this wave nature is associated with all elementary particles, although it is less pronounced in the more massive particles. Everything can be reduced to quanta which have both a wave and a particle aspect; a quantum possesses an indivisible amount of energy given by the equation known as the Einstein relation:

$$E = h\nu$$

E is the energy of the quantum, ν is the frequency of the wave, and h is a constant known as Planck's constant.

The wave aspect of quanta results in a problem: the accuracy of the frequency analysis of a wave depends on the size of the sample being analyzed. This is not a fact peculiar to sub-microscopic physics, but applies to all waves. An instantaneous sample of a wave gives no clue about its component frequencies; any finite sample still leaves an inaccuracy which increases with decreasing frequency (i.e., increasing wavelength).

In quantum physics, this implies that the location of a quantum can only be determined at the expense of accuracy in determining its momentum, and vice versa. According to the famous Heisenberg uncertainty principle, the uncertainty Δp of the momentum of a particle and the uncertainty Δx of its position are constrained by the inequality

$$\Delta p \, \Delta x \geq h/2\pi$$

In terms of experimental limitations, this principle is taken to mean that any attempt to measure a particle will have an error in the measurement of momentum to a degree proportional to its accuracy in measuring its position, and vice versa. If a particle is observed with light of a high frequency and short wavelength, then it provides good resolution in measuring the particle's position; but the Einstein relation says that the energy of the light quanta is proportional to their frequency, so the momentum of the particle will be significantly disturbed. If less energetic low-frequency light is used, the disturbance to the momentum is less, but the larger wavelength makes the observation less accurate.

The Heisenberg Uncertainty Principle is often taken as implying that a particle does not *have* a specific position and momentum prior to their measurement. An experiment can be designed to measure either one with arbitrary precision, but only at the expense of precision in measuring the other. According to the Copenhagen interpretation of quantum physics, so named for the discussions in Copenhagen among Ernst Schrödinger, Niels Bohr, and other physicists in 1926, quanta consist of wave functions which, when squared, describe a probability distribution for the particle. The act of detecting the particle "collapses" the wave function into one of its possible states. It accepts the reality of exact position or momentum, but only as a consequence of measurement, not as something which existed prior to the measurement and which is limited by the disturbance which is necessarily introduced by any attempt at measurement.

One can take this interpretation in two ways. It might be considered as a best possible summary of current knowledge: a statement that we don't know what is really going on at the

quantum level, but that certain mathematical formalisms describe what we can observe. It might also be considered as a statement that nothing underlies the formalism, that the ultimate components of reality are statistics and mathematical formalisms. Since there is no definitive statement of the Copenhagen interpretation, it is often difficult to tell which of these was the intention of the physicists in question; they may have wavered from one to the other themselves. For the sake of clarity and not putting words into anyone's mouth, I will call the first of these sub-interpretations, the idea that the formalism is a summary of current knowledge, the *pragmatic* Copenhagen, and the second, the idea that the possibility of an underlying reality is forever closed, the *metaphysical* Copenhagen. Some people would argue that the metaphysical version is the true Copenhagen interpretation; but others are equally eager to defend physics against such charges. The common element of the two variants is the rejection of the full applicability of conventional concepts at the quantum level; the difference lies in the question of whether other, not yet specified concepts, can be found to provide a new, deeper underpinning.

What are the waves in quantum mechanics waves of? No answer has been given except the loose one that they are "probability waves." This actually means that the square of the value of the wave function at a given point is the probability of the particle's being observed at that point. (The function itself is complex rather than real, and thus has no direct physical interpretation.) Hence, the metaphysical Copenhagen interpretation offers a purely mathematical model, with only a statistical interpretation, as the most fundamental level of physical reality. It offers the best which is currently available in predictive accuracy, but it makes no pretense at providing a model with any structural correspondence to the phenomena underlying the phenomena. In its metaphysical form, it denies any need for structural correspondence. In both forms, it provides no good answers as to why matter and energy act as they do on the quantum level.

By both forms of this interpretation, the attributes of a particle which are being measured do not take on specific attributes until it is observed. But the idea of "observation" is taken in two different senses, often not clearly distinguished by those who see quantum physics as a justification for mysticism: observation in the sense of a physical interaction that depends on the position of the particle, or interaction specifically with a conscious being. The latter interpretation leads to the mystics' conclusion that nothing has any fixed nature until and unless a consciousness observes it.

In an effort to show the absurdity of this last viewpoint, Schrödinger proposed an imaginary experiment. Take an opaque, soundproof box containing a cat; also put in this box a device that will expose a radioactive isotope for a certain amount of time, such that there is a 50% chance of a particle decaying. This device will, if a particle decays, trigger the breaking of a vial of cyanide, killing the cat. If we conclude that the wave function does not collapse until a consciousness observes it (and further assuming that the cat's consciousness does not qualify), then it follows that the cat is not dead or alive until someone opens the box to check; it is in a "superposition of states," an alive state and a dead state, each having a 50% probability of becoming the actual state when the wave function is finally collapsed.[3]

Another view sometimes taken is that there is a live cat as well as a dead cat, but that they exist in different universes; whenever a quantum event may lead to different states, the universe splits into versions in which all of the different states are realized. This is the view taken by Hugh Everett III and endorsed by John Gribbin. Popper has pointed out that this theory does little to account for the fact that past events as well as future ones are indeterminate under quantum theory; a given state of a system can have a probability of 0.5 of

having arisen from either of two different states. Thus, the many-worlds theory would have to provide for constantly merging universes as well as constantly splitting ones in order to fully remove indeterminacy from physics.[4]

Gribbin adopts the idea that the act of observation necessary to collapse the wave function must be an act of conscious observation. This leads to his rejection of objective reality. Indeed, the prologue of Gribbin's book is titled "Nothing Is Real" (but its first sentence says that "The cat of our title is a mythical beast, but Schrödinger was a real person"[5]).

Few physicists have been so willing to embrace nihilism in the name of science; Gribbin himself, in spite of such statements in his book, seeks to rescue reality by postulating multiple universes. Heisenberg, characteristically among quantum physicists, was very concerned that some kind of reasonable accounting for quantum events be devised:

> During the months following these discussions an intensive study of all questions concerning the interpretation of quantum theory in Copenhagen finally led to a complete and, as many physicists believe, satisfactory clarification of the solution. I remember discussions with Bohr, which went through many hours till very late at night and ended almost in despair; and when at the end of the discussion I went alone for a walk in the neighboring park I repeated to myself again and again the question: Can nature possibly be as absurd as it seemed to us in these atomic experiments?[6]

Einstein was particularly concerned with the idea that uncertainty could be a basic part of the universe, and was never satisfied with the metaphysical Copenhagen interpretation:

> But now I ask: Is there really any physicist who believes that we shall never get any inside view of these important alterations in the single systems, in their structure and their causal connections, and this regardless of the fact that these single happenings have been brought so close to us, thanks to the marvelous inventions of the Wilson chamber and the Geiger counter? To believe this is logically possible without contradiction; but it is so very contrary to my scientific instinct that I cannot forego the search for a more complete conception.[7]

Searches for more complete conceptions have, however, met with little success so far. Theories proposing "hidden variables" that remove the indeterminacy from the testing of position and momentum, or other "complementary" pairs of variables, are regarded as unsupported by experimental evidence. Particularly difficult to explain is the phenomenon of non-locality, which is predicted by Bell's Theorem and has been conserved experimentally.

The best explanation for the layman which I have seen of Bell's Theorem is given in Nick Herbert's *Quantum Reality: Beyond the New Physics*. Because this phenomenon is particularly damaging to attempts to apply common-sense ideas to quantum phenomena, it needs to be considered carefully in any attempt to establish the underlying reality of physics on this level.

The key is found in paired particles generated by a common event. These particles are constrained to obey certain conservation laws; hence the attributes of one will have a fixed relationship to those of the other. For instance, the vector sum of the momentum of two particles created in a given system will be equal to the momentum taken out of that system. Hence, some physicists have proposed, by measuring the position of one such particle and the momentum of the other, the position and the momentum of both particles could be established, sneaking by the uncertainty principle. However, a counterargument has been given to every such proposed experiment to date, and such measurements have not in fact been obtained.

Bell's Theorem deals with paired particles from another aspect. Consider, for this discussion, the case of paired photons. One of the attributes of a photon is its polarization, which can take on one of two values, "up" or "down." If a "twin" pair of photons are both measured at the same angle, they will both exhibit the same polarization. If they are measured by detectors at an angle of 90 degrees to each other, they will always have different polarization. If they are measured at an intermediate angle, there is no deterministic relationship between the photons of a pair, but there is a specific degree of correlation between the polarizations. This correlation increases monotonically from 0.0 at 90 degrees to 1.0 at 0 degrees.

This means that an angle can be chosen at which any given correlation can be achieved. There is, for instance, some angle at which the correlation is 0.75. In this case, three out of four photon pairs will have matching polarization, and one out of four will have non-matching polarization. Call this angle **α**.

Now consider the following case. Two detectors, **A** and **B**, are positioned initially at an angle of 0 degrees to one another. The correlation of polarizations is 1.0. Now tum detector **A** clockwise by angle **α**. The correlation now drops to 0.75.

What does this mean has happened? If we go by common sense, we would say that one photon in four has had its measured polarization changed by the act of rotating detector **A**. Specifically, the change has occurred at **A**, and nothing has changed at **B**. This is the locality assumption: changing a measuring device changes only the measurements that take place at that device, not at another measuring device which may be light years away.

Next turn detector **B** counter-clockwise by angle ct, giving a new total angle between the detectors of 2**α** . What happens to the correlation at this point? Going again by the locality assumption, the measurement of one photon in four at **B** will be changed. If we assume that these changes are random and independent of the changes at **A**, then the new correlation should be 9/16 or 0.5625. Even if the changes are not statistically independent, locality requires that the correlation be at least 0.5. If the photons affected by rotating **B** were always members of the same pairs that were affected by rotating A, then the correlation would go back up to 1. If the photons affected by rotating by were never members of the same pairs. the correlation would be 0.5.

However, quantum theory, backed up by experimental evidence, indicates that the actual correlation for an angle of 2**α** is significantly less than 0.5. The virtually inescapable conclusion is that changing the state of one detector changes the result of a different detector measuring a different particle. To use a technical phrase, locality is violated. A change is effectively propagated instantaneously from one detector to the other, ignoring speed-of-light limitations. There is no apparent way to utilize this phenomenon to send messages faster than light, though, since the propagation consists solely of changing one stream of random data into a different stream of random data with the same distribution.

This result appears to undermine the idea that each particle has "hidden" attributes which cannot be measured precisely only because the act of measuring disturbs their values. The attributes of one particle are, at least for the case of paired particles, linked to the attributes of other particles, and the act of measuring one can affect the attributes of another, even at a distance. By the same token, the idea that each particle has randomly variable attributes which are independent of other particles is demolished by this demonstration; the image of God

playing dice with the universe fares as badly as hidden variable theory. What these non-local connections imply about the underlying reality is a major unsolved problem.

But should we conclude from this that reality simply does not make sense, that the world is one big illusion? No. However confusing the facts discovered by science are, they are objective facts.

Suppose that some piece of evidence did demonstrate that certain observations were in fact products of the observer's mind. Then they would simply be illusions, data reflecting only on the mind of the person experiencing them. If the evidence of quantum mechanics actually indicated that what was being observed was not real, it would be of no interest to science. If what is reported is true for everyone, then it is not an illusion and not a product of anyone's mind.

If, as the Bell theorem appears to show, events separated in space and time are nonetheless intimately connected, this is a fact of reality. The fact that it does not coincide with our perceptions of macroscopic reality does not make it unreal, any more than the contrast between the actual roundness and apparent flatness of our planet makes the Earth unreal.

Nor does Bell's Theorem back us into the conclusion that reality is ultimately probabilistic in nature. There is, in fact, no possible set of data which can show that, since for any set of variations which are purported to be ultimately indeterminate, one can without contradiction posit a cause which produced exactly that set of variations. (Applying this reasoning to one's own thinking is a separate case, which I will consider when discussing free will.) Saying that probabilities are at the root of reality is simply saying, in a less honest way, that one does not know what is at the root.

Likewise, Bell's Theorem does not back us into the view that there is nothing beneath the mathematics of the situation. This view, like the view that probability is fundamental, is actually vacuous. Ultimately, there is something, even if we do not know what it is at this time. "There is no quantum world," said Bohr in upholding the metaphysical Copenhagen view. "There is only an abstract quantum description."[8] But one cannot speak of abstractions while denying that which they are abstracted from; otherwise they become mere playthings of the mind, models without referents, unconnected to science. There is no denying that finding a consistent characterization of reality which fits all the known facts of quantum physics is a tremendously difficult task; but giving up the job means giving up science and settling for the status quo of knowledge. Let us consider these points more broadly.

The Permanently Unknowable

In Chapter 1, I indicated that a model is generally valid on a single level and in a specified context, that predictive accuracy alone does not validate a model, and that limitations which apply to a model do not necessarily apply to the process being modeled. In Chapter 4, I explained that probability is not intrinsic and not a cause, but rather a description of partial knowledge. Variations of the Copenhagen interpretation of quantum mechanics commit errors with regard to all of these principles.

The generally accepted description of quantum mechanics constitutes a statistical model. The behavior of subatomic particles is mathematically described by the Schrödinger wave functions, the Einstein relation, and many other mathematical tools that have been developed for the purpose.

The metaphysical Copenhagen interpretation constitutes an assertion that this model constitutes a full explanation of quantum phenomena, and hence that the ultimate causal factors in matter are mathematical or probabilistic in nature.

Let's take these points one at a time. On what basis can it ever be claimed that a given explanation of phenomena is a full explanation, and that no deeper level of causation lies behind the known causes? In fact, such a claim would require omniscience. All knowledge is contextual; it arises from gathering information and formulating general concepts and principles based on that information. The information a person can gather is always finite, and always gathered by specific means. It is limited at one end by the scope of the procedure used and at the other end by its precision.

Newtonian mechanics once seemed universally valid; but Newton had arrived at its principles by a procedure that was limited in scope to velocities which were small compared with that of light. This does not mean that he committed a fallacy in regarding his conclusions as valid; it would have been a fallacy only if he regarded them as valid beyond any possibility of correction under conditions yet to be found.

It can be argued, though, that quantum physics is a different case and not subject to further correction. Newtonian physics did not set any limit in principle on accelerating masses to a greater and greater extent to see if anything different happens, but the Heisenberg uncertainty principle does set a fundamental limit. It makes no more sense to ask what is happening on a scale smaller than the uncertainty limit than it does to ask what happens when an object travels faster than light.

But the speed-of-light limit and the uncertainty limit are different in an important way. Saying that an object is limited by the speed of light does not negate its identity; it still has a particular velocity, which will always be less than that of light. Saying that an object does not have a specific position is meaningful only if position turns out to be a derivative phenomenon, based on some characteristics which are applicable on a sub-quantum scale. If position is taken to be an ultimate, irreducible characteristic, yet one which does not have a particular value, then that says that an object's position, momentum, and other fundamental attributes are nothing in particular. But this implies that the object is nothing in particular.

If the variables of quantum mechanics are statistical results of more fundamental events, then they may be regarded as approximate, but not as fundamental. A probability distribution can't be fundamental; probabilities are expressions of partial knowledge, and the identity of something cannot consist of one's knowledge or lack of knowledge about it.

Temperature is an obvious example of an attribute which is inherently approximate, because it is a statistical consequence of numerous causes. The way that a substance interacts with a thermometer depends, at a fine level of precision, on which molecules collide with the thermometer's molecules and how much energy is transferred. On a sufficiently fine scale, the same substance is as likely to produce one reading as another, depending on the unrepeatable particulars of molecular motion. However, the molecular motions themselves are still precise on that scale, and it is their aggregate behavior that gives rise to the macroscopic phenomenon of temperature.

Position, size, and momentum of macroscopic objects are also approximate, since these objects are composed of atoms which are individually jiggling around. Below a certain level

of measurement, the random motion of these particles prevent adding further decimal places to even the most careful measurement.

Likewise, if position and momentum are inherently statistical on the quantum scale, it can only be because some other factors combine in various ways, uncontrollable on the larger scale, to give rise to the comparatively macroscopic variables which are found in quantum mechanics. The alternative requires abandoning the law of identity: declaring that certain quantities are vague not because their measurement is really the result of a complex of phenomena which are not fully known, but because they possess intrinsic vagueness, because they are fundamentally nothing in particular.

It is not necessary to choose among the possibilities without sufficient evidence. Scientists must try to find out whether the uncertainties reflect specific values which cannot be measured because the experimental apparatus inevitably involves the same uncertainty, or whether there is a more basic set of phenomena that justifies treating the quantum variables as approximations to complexes of these deeper phenomena. However, it is necessary to realize that approximations and probability distributions are expressions of incomplete knowledge.

It has been claimed that experiments have shown that there cannot be any hidden variables in quantum physics. Arriving at such a proof is fundamentally impossible, though. Such a proof would amount to proving that in a given experiment, there were no factors that necessarily caused events to happen as they did. What sort of observations could lead to this conclusion? The only possibility is that in the same experiment, events happened in a given way, yet did not happen that way. Clearly this "possibility" is no possibility at all. Whenever we observe the outcomes of two different experiments, we cannot establish by observation that there was no factor that caused the difference between them. At most, we can say that the source of the difference is not accessible through known methods.

The experiments which have validated Bell's Theorem apparently say that if there are hidden variables, they cannot be exclusively local in nature; that is, some of these variables must reside in systems rather than in single particles. But if we substitute probabilities for determinate variables, we are in the same boat; at least some of the random factors belong to the system, not to the behavior of individual particles. In either case, one must seek an explanation of the fact that the way a particle is measured on the moon can affect the results of measuring another particle on Earth. Something is going on; the best today's physics can do is to predict the outcomes of experiments on a statistical basis, without giving any real insight into these connections. It would be wrong to fault physicists for not having reached a complete explanation all at once; but it is also wrong to suppose that what we have today is a complete explanation. As Heisenberg noted, "If predictive power were the only criterion of truth, Ptolemy's astronomy would be no worse than Newton's."[9]

Bohm, in *Causality and Chance in Modern Physics*, does a good job of criticizing the concept of "absolute chance," which "is not conceived of as being arbitrary and lawless relative to a certain limited and definite context, but rather as something which is so in all possible contexts." He points out that the idea that reality is controlled by an "idealized roulette wheel" of absolute chance commits the same basic error as earlier ideas that the universe is completely determined by mechanistic forces:

> But in doing this, it [the idea of absolute chance] has conserved and in fact enhanced the central and most essential characteristic of this [mechanistic] philosophy; namely, the assumption that everything in the whole universe can be

> reduced completely and perfectly to nothing more than the effects of a set of mechanical parameters undergoing purely quantitative changes.[10]

Bohm calls the two variants of this fallacy "deterministic mechanism" and "indeterministic mechanism." The common fallacy which he identifies is identical to the fallacy of regarding a model as the full equivalent of a process. Bohm restricts the use of the word "model" to deterministic models, but if models are understood to include probabilistic ones, then my statement of the black box fallacy with respect to models is equivalent to Bohm's statement of the fallacy of mechanism.

Bohm points out that at any state of knowledge, contextual issues apply, and it is the job of science to look for new and previously unsuspected relationships that fill in the context. Science can never conclude once and for all that it has finished its job, that there are no uncovered causes or laws that lie behind the known ones.

> For now one sees that not only are there no known cases of laws that accomplish this mechanistic aim, but even more, that even if we did have a law that seemed to explain everything that was known at a given time, we could never be sure that the next more accurate experiment or the next new kind of experiment would not show up some inadequacies that would lead eventually to a still more general and deeper set of laws. Indeed, this latter has been what has happened in physics thus far, with all the laws that have at one time or another been thought to be the final ones.

There is one problem with Bohm' s treatment: he regards causal and statistical laws as having equal standing, with either one potentially explainable in terms of laws of the other type. This is a result of his failure to recognize that probabilities are a measure of interchangeability of causes due to incomplete knowledge, hence necessarily imply a deeper level of causation. Causal laws mayor may not have their roots in more fundamental causes; but there is nothing in the fact that a relationship is causal that requires it to have such roots.

It should also be pointed out that "causal" does not mean "deterministic," as Bohm assumes. A causal relationship is one between entities and their resulting actions. If only one outcome is possible to an entity under a given set of circumstances, then the relationship is deterministic; but if different outcomes are possible, then the relationship is causal but non-deterministic. Free will in humans is an example of a non-deterministic causal relationship; the acting entity is the person to whom different ways of thinking are possible under any given set of circumstances. (This will be discussed in more detail later on.)

Common Sense and Quantum Reality

Much of the difficulty in anyone's understanding of quantum physics comes from the implicit view that the concepts of everyday observation are intrinsic to reality. This has, it would appear, been almost as much of a problem for the leaders of the field as for the layman. Concepts such as size, mass, and momentum refer to attributes of an entity which consistently result in their interacting with other entities in certain ways; these interactions are the means by which we arrive at the concepts. On a radically different scale, interactions between entities may be entirely different in kind, and call for a different set of concepts. If the everyday concepts fail to apply in any way that can be discovered, the possibility that alternate concepts are needed must be considered. Yet Heisenberg has insisted that science must stick with the familiar concepts:

> Therefore, it has sometimes been suggested that one should depart from the classical concepts altogether and that a radical change in the concepts used for describing the experiments might possibly lead back to a nonstatistical, completely objective description of nature.
>
> This suggestion, however, rests upon a misunderstanding. The concepts of classical physics are just a refinement of the concepts of daily life and are an essential part of the language which forms the basis of all natural science. Our actual situation in science is such that we do use the classical concepts for the description of the experiments, and it was the problem of quantum theory to find theoretical interpretation of the experiments on this basis. There is no use in discussing what could be done if we were other beings than we are.[11]

The suggestion is that "what we are" restricts the concepts that we can use to familiar, traditional ones; that these concepts are intrinsic, if not to reality, then to our minds. This represents a form of psychological determinism, which stands as a barrier to recognizing that people can think in new ways. In effect, Heisenberg is saying our knowledge must fit our concepts, rather than the other way around.

Quantum physics certainly requires new approaches to its subject matter, and it poses questions which even Einstein was unable to answer satisfactorily. But there is nothing in it which calls for abandoning the fundamental concepts of identity, causality, or reality for the sake of preserving less fundamental ones. Quantum physics is, after all, science, and science cannot exist without these concepts.

The non-physicist reader who would like to see a good approach to a realistic interpretation of quantum physics should look at Karl Popper's *Quantum Theory and the Schism in Physics*. While I am not in a position to judge all of its conclusions, and while I find his endorsement of probabilistic "propensities" as fundamental causes dubious, there is a great deal worth considering in his treatment of the observer issue and the statistical nature of the laws of quantum mechanics. Herbert's *Quantum Reality* offers a more up-to date survey of alternative interpretations of quantum physics, and is also worth reading.

---

1 Gribbin, p. 208.

2 Ibid., p. 183.

3 A real-life analogue to this experiment might be the Chernobyl nuclear disaster. If no one is able to observe how many people died due to the decay of radioactive particles in the vicinity of Kiev, are the people in question suspended in a superposition of states?

4 Popper, *Quantum Theory and the Schism in Physics*, p. 89-95.

5 Gribbin, Op. Cit., p. l.

6 Heisenberg, p. 41-42.

7 Einstein, *Out of My Later Years*, p. 91.

8 Nick Herbert, *Quantum Reality: Beyond the New Physics*, p. 17.

9 Heisenberg, *Physics and Beyond*, quoted in Ken Wilber, *Quantum Questions*.

10 Bohm, p. 63.

11 Heisenberg, p. 56.

VII. That Which is not Seen

To understand what permits scientific theorists to be satisfied with a theory that predicts observed results while saying very little about what happens to create these results, we have to go deeper into philosophy. Briefly, the acceptance of such a theory as the final word comes from an attempt to preserve empiricism without falling into subjectivism. Empiricism, in the pure sense, relies on observation alone to gain knowledge. For its result to be considered real knowledge, however, it has to answer the question: how do you know that your observations refer to something which has objective existence?

Some philosophers have worked on answers to this question, while others (notably Hume) have declared that it is hopeless and that we have no guarantee that the universe won't go berserk on us in the next second. But another answer was offered by the nineteenth-century philosopher Auguste Comte. His answer was that the question is illegitimate, that any questions of reality which go beyond what is empirically determined are invalid. This philosophy is called "positivism."

Comte regarded philosophy as advancing through three stages, the Theological, the Metaphysical, and the Positive. In the Theological state, the ultimate causes of events and the essential nature of existence are explained in terms of supernatural beings. In the Metaphysical state, they are explained in terms of "abstract forces" and "veritable entities." In the Positive stage, a "fundamental revolution" leads us "to substitute everywhere, for the inaccessible determination of *causes* properly so called, the simple seeking of *laws*, that is, constant relations that exist among observed phenomena."[1] Put another way, positivism resolves the battle among contending metaphysical systems by rejecting metaphysics itself.

A similar view is expressed by Nietzsche, who was not a positivist in the full sense, but who always had the knack for expressing ideas in a compactly quotable form, in *Twilight of the Idols*:

> We possess scientific knowledge today to precisely the extent that we have decided to accept the evidence of the senses—to the extent that we have learned to sharpen and arm them and to think them through to their conclusions. The rest is abortion and not-yet-science: which is to say metaphysics, theology, psychology, epistemology.

In the light of nineteenth-century philosophy, a reaction against metaphysics is understandable. Kant had split the world in two, one part real, the other observable; Hegel had devised a world filled with contradictions, each one eventually resolving itself only to give rise to a new contradiction. Marx had given each social class its own reality and left them to leap at each other's throats for lack of even a theoretical possibility of mutual understanding. The positivist response was, in effect, "To hell with all this; let's stick with the understandable, the scientific." The popularity of Comte and Nietzsche has waned greatly since the nineteenth century, but their basic ideas live on in this respect.

But in fact it is impossible to reject metaphysics; it is only possible to refuse to consider it openly. Every philosophy must answer the question: What is real? if only by implication. Comte's implied answer was that only the observed is real. Ernst Mach later developed this principle of Comte's further, making it still more rigorous.

This is a safe answer as long as the world provides no serious surprises. In response to Berkeley's inquiry about whether the back of his head existed when nobody observed it, positivism could respond that it didn't matter, as long as it was there whenever anyone tested its presence.

Yet it does not resolve the basic issue. If all we have is collections of observations, then scientific laws are merely economical summaries of the way things have behaved so far. If we are not allowed to make any statements about the nature of reality, we cannot say whether things will continue to behave that way, nor whether the things we haven't observed behave in similar ways.

Positivism must, if pushed, choose between skepticism and metaphysics. However, it can skirt the issue so long as observation provides a fully consistent reality. So long as this is true, the positivist can say it makes no difference whether or not there is some fundamental reason for this consistency; we see the consistency, and that is all that counts.

But suppose it turned out that there were aspects of reality that changed for no apparent reason, and that every effort at observation failed to account for those changes? This would be a challenge to positivism; it would have to make a choice. Either it would have to say that there is an underlying reality in spite of our inability to observe it, or it would have to say that observation is all and that we can't even ask what causes the changes. In the nineteenth century, the suggestion of such a situation would have seemed hypothetical and absurd, therefore not worth considering by positivist criteria; any inability to observe something would be regarded as temporary and conditional, allowing eventual resolution. But in the twentieth century, reality played exactly such a dirty trick on the positivists. It was found that the position and momentum of a particle could never both be precisely measured together, due to the effect of wave phenomena on a sub-microscopic level. Worse, it was found that measuring one particle could affect the results of measuring another particle.

Faced with a dilemma, positivism remained consistent; through the Copenhagen interpretation of quantum physics, it declared that the unobservable was the unreal, that the whole sum of facts regarding a particle was only a probabilistic distribution of its characteristics.

This application of the philosophy implicitly denies the principle that what is, is itself and not something else. For it retains the idea that a particle possesses these characteristics, yet it denies that it possesses them to a particular extent. This means that it simultaneously does and does not possess a given momentum or position; the contradiction is harmless, from the positivist view, because there is no way to determine whether it possesses those values or not. This view is sometimes known as "complementary logic," because it upholds two views which are mutually contradictory, but which one can switch between without ever having to apply both views to the same event. The contradiction applies only to the implicit view of the underlying reality, not to the observations.

What is the alternative? If one is committed to the idea that A is A, how would one deal with an A that fluctuates in mysterious ways that elude all measurement? The answer is that it must be taken as a given that there is some cause in reality for any event. If all known methods for establishing the value of A lead to dead ends, that simply means more methods must be explored, or that the wrong question is being asked. If the further exploration leads to the conclusion that discovering any particular value for A would lead to a contradiction, then the

alternative is to admit to having made some error in conceiving of A, or in applying the concept of A in a given context.

Thus, if there *must* be an uncertainty in the position and momentum of the particle to avoid a contradiction between its particle nature and its wave nature, the conclusion to reach (if one is consistent with the law of identity) is that one or both of these concepts simply does not apply on this level in the way that we are attempting to apply it. This requires discovering some other attribute (or variant of the attributes previously considered) which can be identified and can possess a specific value in that context. This does not stop position and momentum from having meaning in macroscopic terms; it does mean that they ultimately depends on whatever these other attributes are.

What might these new attributes be? I am not going to be so stupid as to offer any suggestions. Discovering them will very likely require a mind comparable to Einstein's or Newton's. But the effect of positivism is to discourage such a discovery by claiming that it is unnecessary. For in rejecting metaphysics, positivism is saying that explanations are unnecessary. Observe, says positivism, and describe insofar as possible; but don't worry about causes beyond what is observed. In spite of positivism's intent to be scientific, what it promotes is scientific stagnation, for it excludes the key question of science: "Why?" It does not permit what Comte called "the vain search after ... the causes of phenomena."

Heisenberg often distrusted positivism greatly; still, he was apparently influenced by it in his insistence on applying previously known concepts, however badly they explained the phenomena at hand:

> The Copenhagen interpretation of quantum theory starts from a paradox. Any experiment in physics, whether it refers to the phenomena of daily life of to atomic events, is to be described in the terms of classical physics. ... We must keep in mind this limited range of applicability of the classical concepts while using them, but we cannot and should not try to improve them.[2]

The reason positivism discourages the formation of new concepts is twofold. Epistemologically, concept formation depends on the understanding that there are entities in reality which have a certain underlying nature in common and therefore can be regarded as a group. If entities may not be regarded as having common properties because nature constrains them to do so, but only because they have been observed to do so, then any concept becomes much more difficult to justify. Psychologically, positivism cuts the mind off from having any confidence that reality as such has a constant nature (since such confidence can only be based in metaphysics); this tends to promote the alternative of clinging to familiar beliefs as the only constants possible.

It is true that in formulating scientific principles, the quantum physicists were able to make great leaps away from previous beliefs. But Heisenberg's approach demonstrates a crucial cognitive inversion: he found it necessary to retain the previous scientific concepts at all costs, while being willing to throwaway the metaphysical basis which had originally made them possible. Lacking certainty in the axioms of existence and identity, it is necessary to find other bases to cling to, such as society and upbringing.

If the solution to the paradoxes of quantum mechanics requires replacing the normal concepts with which we describe the macroscopic world with entirely different concepts that apply on a very small scale, the approach introduced by positivism greatly discourages their discovery. Certainly nearly all people are generally conservative about their basic concepts,

and unwilling to change them without clear need. It is indeed true, as Rand has pointed out, that concepts, like entities, should not be multiplied beyond necessity. But when it is necessary to devise new concepts, philosophy must provide the basis for doing it.

This may seem like a paradoxical thing to say, considering how large a change from conventional notions the Copenhagen interpretation of quantum physics involves. But the facts which quantum physics is attempting to account for are extremely strange by common-sense standards, and the difficulties with the current theories may well be that they have tried to stay too close to the familiar rather than achieving the necessary conceptual revolution. In some contexts, it may be that the concept of space itself is not applicable; if so, then insisting on using that concept because it is essential to macroscopic observations would be a mistake. At the very least, quantum physics has been too willing to say that everything has been explained when all that it has done is describe the observed phenomena.

Positivism claims to reject the search for causes that lie behind observed phenomena; but if it did so consistently, it would have to abandon science and fall back to a pretheological stage of thinking. Causality is the principle that entities of a certain kind must act in specific ways under specific circumstances. Without this principle, the "problem of induction" is unsolvable; there is no way to conclude, from any amount of evidence, that a regularity discovered in some number of cases will occur in all cases of that type. Laws, as positivism defines them, are nothing but descriptions of what has been observed so far.

Scientists can and do conclude that what is consistently observed to occur in a large number of cases is going to continue to occur; but without recourse to defined principles of causality, their willingness to generalize may lead them to the error of asserting laws where no necessary relationship exists. The result can be a willingness to draw conclusions on the basis of statistics alone, or to generalize from one species to another (e.g., pigeons or mice to people). The problem is especially acute where phenomena cannot be fully isolated, where an effect may depend on factors other than the one being tested. Such cases include quantum physics, for the reasons already discussed; they also include studies of human beings, because of their complexity and variability, as well as the ethical and practical impossibility of performing many kinds of tests on people. In these cases, the need to abstract causes is essential to understanding.

A scientific law must, to be sure, take observed events as its starting point. But it must do more than simply summarize previous observations and predict new ones. It must state, on some level, what its subject matter is; it must state that certain kinds of entities or substances act in certain ways. The entities and substances in question do not have to be ultimate fundamentals in reality; a law may provide a correct description because of causes on a lower level which are completely different in kind from the ones under consideration. (For instance, gas laws can properly treat their subject matter as a continuous fluid, even though at a lower level it consists of particles that are more like billiard balls in their motion.) But identifying the subject matter and its associated attributes is a prerequisite of establishing causation; we must be able to say, "This acts thus because it is such and such." Saying what the subject matter of quantum physics is (be it particles, waves, or something else which underlies both) is the major unsolved challenge that science faces today.

Quantum physics has, according to Gribbin, Zukav, and others, undermined the concept of objective reality. But in fact, it is positivism which rejects objective reality in its premises. This rejection is then consistently applied to the data of physical experiments, and the premise remains intact at the end; but it is asserted as a conclusion, when it is simply a restatement of

the original assumption. Specifically, the metaphysical Copenhagen interpretation assumes that if something is incapable of being measured, it is pure potentiality which does not arise from any actuality. This is consistent with positivism, which will not grant metaphysical status to anything that cannot be observed. But it is not an experimental result of physics. For an experiment to invalidate objective reality, it would have to produce a result that was not any particular result, an observation that both was and was not some X. But doing this would only invalidate the experiment in question, not the whole of reality.

The rejection of objective reality is implicit in the statement that only the observed is real; for objectivity means the principle that reality is independent of observation. What comes out of Schrödinger's box is only what went in: no experimental results, interpreted through the positivist principles, can salvage objectivity in a world which is already assumed to be non-objective.

Model Without Reality

The dilemma which positivism poses may be restated in terms of this book's theme. Are scientific laws something which are true of reality, or only of a construct which people have created in attempting to understand reality? Under the tenets of positivism, there is at least a certain sense in which they cannot be true of reality, since that would be ascribing a metaphysical dimension to them. This is made clear in the comments of the positivist Ernst Mach: "Cause and effect, therefore, are things of thought, having an economical office. It cannot be said why they arise. For it is precisely by the abstraction of uniformities that we know the question 'why'."

All theories that describe the facts are equally valid under positivism. It is no more true to say that the Earth revolves around the Sun than that the Sun revolves around the Earth, so long as all the other motions in the sky are correctly described as well; the choice of regarding the Sun as the center is merely the simpler one to describe. Hence, any number of different models can be concocted to describe the behavior of natural phenomena; all are just models, not true accounts of what is "really" happening.

This is the black box fallacy at its height, and it reaches its very apex in the metaphysical Copenhagen interpretation. According to Heisenberg, "Instead of asking: How can one in the known mathematical scheme express a given experimental situation? the other question was put: Is it true, perhaps, that only such experimental situations can arise in nature as can be expressed in the mathematical formalism?"[3]

Under positivism, such a conclusion is almost inescapable; for the alternative, that both the situation and the formalism arise from the underlying properties of the real situation, requires admitting that existence is independent of observation.

The error of positivism is corrected by recognizing that observation is our only ultimate source of information about reality, but that the existence and nature of reality do not depend on our observing it. If reality cannot exist apart from observation, then that we are doing is not observing reality, but creating it. John Gribbin embraces this view wholeheartedly:

> By the act of observation we have selected a 'real' history out of the many realities, and once someone has seen a tree in our world it stays there even when no one is looking at it. ... At every junction in the quantum highway there may have been many new realities created, but the path that leads to us is clear and unambiguous.[4]

The irony is striking; starting from the conviction that only the observed is real, Gribbin arrives at the belief in Gogol plexes of universes which have never been observed. This is a move of desperation on his part, an attempt to endure the collapse of Platonic observation and the lesser horn of the dilemma that "either nothing is real or everything is real."

There is another issue in positivism, which leads to another set of problems unrelated to quantum mechanics: who observes the observer? Each person's consciousness is observable by exactly one person: himself. We can observe the effects of other people's consciousness: their statements, their actions, their neural activity; but we cannot observe the actual act of observing, except in our individual selves.

This makes observation itself suspect: for if observation is necessary to reality, then how can any person regard other people's minds as real? Each person could decide that only his own mind was real; but the widespread adoption of solipsism would bring communication and science to a state of collapse.

An alternative is to regard the mind as consisting only of its observable aspects: its physical and neural consequences. And herein we find the view of the mind which lies at the root of theories of thinking computers, and the equation of a model of the mind with the mind itself. This view will be the subject of most of the remainder of this book.

---

1 Comte, p. 24 (translation mine).
2 Heisenberg, *Physics and Philosophy*, p. 44.
3 Heisenberg, p. 42.
4 Gribbin, p.251.

VIII. The Faculty of Awareness

The most fascinating and complex issue in modeling is the attempt to develop models of the mind's operations. It is here that computers come to the fore, since all agree that thought is an extremely complex process, which in many respects is not well understood.

A number of questions have to be identified at the outset. These are:

1. What is the nature of the process being modeled?
2. What are the goals of the model?
3. What are the limits, if any, on the completeness and accuracy of the model?

The first question is essentially philosophical; it asks just what mind, consciousness, or thought is. The second defines the practical issue and may limit the extent to which the first question has to be answered for a particular application. The third question relates the answers to the first two; given the nature of the mind, what goals are achievable in modeling it?

The first question is the most critical. Unless we know what we are modeling, any questions about how to model it or what success to expect will be floating in midair.

Two views of the mind have been very common in philosophy; these may be called the spiritualist and the materialist views. The spiritualist view holds that consciousness is a separate substance which coexists with the body. In most versions of this view, it controls the operations of the body and gains sensory information from it, but it is not really part of it. Consciousness possesses free will and is exempt from the mechanistic causality of matter. The spiritualist view of consciousness is often associated with religion, but it can be held independently of arguments for a God. This is sometimes called the "dualist" view, but I prefer to use the word "dualism" for the broader view that the mind and the brain are not one and the same thing. Dualism, in this sense, permits a closer tie between the mind and the brain than spiritualism does.

The materialist, mechanistic, or reductionist view holds that what is called "consciousness" is simply the organizational and informational capacity of a living being, or potentially even of a machine. It is not a special substance, but simply the result of having complex information-processing mechanisms that allow specialized organs to store information about the world around the being and to act in an effective way.

Another view, which is sometimes called the idealist one, is that only consciousness exists and that matter is its creation. This view makes all knowledge subjective, so I will not discuss it any further here.

The spiritualistic and mechanistic views of the mind predominate in philosophical discussion to such an extent that they may seem like the only alternatives. For example, Pamela McCorduck's *Machines Who Think* states:

> Presently no complete, coherent model exists that explains all aspects of mental behavior, but most researchers are agreed: there's no ghost in the machine. Everything from symphonies to simultaneous equations to situation ethics is finally

produced by those electrochemical processes. This view can be considered mechanistic.[1]

Understanding the mind requires a method appropriate to the object of study. This method must be distinctive because, unlike all other cases, the object of study is also that which does the studying. Putting it another way, the mind is that which observes rather than that which is observed. A person observes his own mind not through information gathered by his senses, but through the process of introspection.

For some philosophers, this distinctive nature makes the existence of one's own mind the most obvious fact of all; Descartes' argument that "cogito" is the first fact known is a well-known example. For others, it makes the character of thought very troublesome, since it cannot be subjected to scientific tests that are independent of the observer.

Is consciousness objectively real? One might argue that it is not, since objective reality is defined to be that which is independent of consciousness. The terms "mental" and "real" are often used as antonyms; that which is only in one's mind is not real. But we must be careful about the context in which we use these terms. To say that something is independent of consciousness is to say that it is there whether we recognize it, are ignorant of it, or refuse to acknowledge it. In this sense, the *fact* of consciousness is independent of the content of any person's consciousness. If deny the fact that I am consciousness, I do not thereby render myself a mindless robot; if I assert that a stone is conscious, I do not change its nature.

The fact of one's own consciousness is impossible to deny without self-contradiction. If I claim that I do not think, I cannot acknowledge the fact that I have made such a claim without recognizing that I do think. If I claim that my activities consist only of meaningless physical motions, then my claim is itself a meaningless sound. As discussed in Chapter 5, consciousness is axiomatic. Rand's third axiom, which expresses the axiomatic concept in the form of a proposition, states: "One exists possessing consciousness, consciousness being the faculty of perceiving that which exists."

But what exactly am I recognizing when I say that I think? Am I identifying a special substance that resides in my body, and which constitutes the essence of awareness? Or am I referring to an information-processing mechanism which generates the actions of my body in response to the data gathered by my senses?

In the most fundamental terms, what I am recognizing is a relationship; the entity which is myself exists in a certain relationship to something outside myself, namely the relationship of being aware. Whether this relationship exists between my body and the outside world, or between a special spiritual substance and the outside world, is a question for later study. The fact of consciousness is the fact of a relationship between an entity and that outside it, or between an entity and an aspect of its own existence.

The relationship is not necessarily one of direct perception. It may be one of recalling earlier experiences of the outside world (memory), modifying those recollections into different combinations (imagination), or identifying the relationships formed by one's own consciousness (introspection). These are examples of relationships between an entity and other relationships which it is capable of forming; but the chain must ultimately end with something outside itself. An entity could not imagine something without having some sort of data to work from.

However, it is an error to equate this relationship with a change in only the physical state of an entity. If consciousness is identified as a physical state, or as a change in such a state, then the observation of that state becomes a fact that has to be explained. Explaining it by reference to another physical state that constitutes the observation of the first state leads to circularity or infinite regress. When we move the observer into the realm of the observed, we still are left with something doing the observing.

All attempts to deny the existence or significance of consciousness smuggle in an observer somewhere along the way. Consciousness is axiomatic and cannot be denied without implicitly asserting it. When someone states that consciousness is a physical state, he has to exempt his own consciousness of that fact from his claim; otherwise his assertion is itself nothing but a physical state. When consciousness becomes an object of study, there still must be a consciousness doing the studying.

This does not exclude the conclusion that the relationship is caused by or intimately associated with a physical state or change. Branden has pointed out this distinction quite nicely:

> It is true that whereas matter can exist apart from consciousness, consciousness cannot exist apart from matter, i.e., apart from a living organism. But this dependence of consciousness on matter does not in any way support the claim that they are identical. On the contrary: as more than one critic of reductive materialism has pointed out, it is reasonable to speak of one thing being dependent on another only if they are not identical.[2]

It is an error to regard consciousness exclusively as an observable datum, instead of recognizing that it must lie at the end of any chain of observations. Causally, existence—that which is observed—must be prior to consciousness. The world could still exist without any beings to be aware of it; but an awareness with nothing to be aware of would be impossible.

J. F. Sowa, in *Conceptual Structures*, calls this view of consciousness the conceptual: "The concept of mind belongs to a complete system for talking about people and their ways of knowing, believing, understanding, and intending. Neurophysiology provides a totally different system of concepts for describing how the brain works. Although the mind depends on the brain, mental concepts are not definable in neural terms."[3]

Consciousness, as I indicated briefly at the end of the last chapter, presents a problem for positivists. On the one hand, it is essential to positivism, since there can be no observable facts without the capacity to observe. On the other hand, one person's consciousness is not directly observable by any other person, so claiming that other people are conscious constitutes a metaphysical inference rather than an observation. The only consistent escape from this problem is a kind of agnostic semi-solipsism which refuses to consider the question of whether other people are conscious or not. (True solipsism would itself be a metaphysical conclusion.) This approach has the problem of regarding knowledge gained by introspection as completely inapplicable to other people; it excludes drawing any conclusions about their inner state. The positivist can observe that other people make statements like "I'm happy," and that certain facial expressions and actions are associated with such statements, but he is moving beyond the realm of observed fact when he concludes that this cluster of actions comes from a consciousness like his own.

Behaviorism accepts this restriction gladly; that school of psychology (or anti-psychology) allows nothing to be considered but the interrelationship of stimuli and responses. In doing so,

it forfeits any hope of explaining why people act as they do; it can only report that they do. It settles for predictive accuracy, and achieves very little even of that. Behaviorism has lost a large part of its support since people have realized that it is inadequate for explaining computers, let alone human beings.

On the other hand, abandoning the positivist principle and acknowledging the validity of causal inference in metaphysics eliminates the difficulty. I can observe that other people are able to make statements based on information, that they react to physical stimuli, that stressful situations tend to make them act in certain ways, and that in all these respects they bear a general similarity to me. In considering a physical basis for this similarity, I can observe that their bodies are similar to mine. From these facts, and in the absence of any evidence to the contrary, I can conclude that the actions of other people which are similar to mine proceed from similar causes, i.e., the capacity of myself and other people to think, and that the similarity of their bodies to mine constitutes a common cause for the effect of consciousness. Thus, although it is not directly observed, the consciousness of other people is a fact supported by overwhelming evidence.

An often-cited difficulty with the concept of consciousness is the problem of knowing just what kinds of entities are conscious. The evidence that people are conscious is overwhelmingly strong. What about other mammals? They exhibit similar behavior to human beings in many ways, and they also possess brains. However, they do not speak, and their reasoning capacity is very limited in comparison with humans'. The evidence that they are conscious is very strong, but not as strong as it is for humans. At the other end of the scale, a clam performs basic biological functions, such as eating and limited locomotion, but it lacks any organ that resembles the human brain. There is some evidence that it is conscious, but not enough to arrive at anything like certainty.

The fact that we have differing degrees of certainty in these different cases is not an objection to the concept. There are many cases in which we lack certainty; in these cases, we must either try to expand our knowledge or accept the uncertainty for the time being. Further study into the biological causes of consciousness could help to resolve the question in these cases. The difficulty could be even greater if we encountered a being that was conscious, but whose biology was completely different from the kinds known on Earth. In such a case, nonetheless, we would have to rely on similar considerations: does the being act in ways that we know are associated with consciousness, and does its physical nature appear consistent with that conclusion rather than alternative conclusions? (An example of an alternative conclusion would be that the being is remotely controlled by an intelligence located elsewhere.)

Consciousness is an aspect of existence that must (with the exception of one's own consciousness) be inferred, rather than being directly observed. However, it is not unique in this respect; electrons and radio waves also cannot be directly observed. What is unique about consciousness is that our initial knowledge of it comes not from what we observe, but from being aware that we observe something. No one can know the consciousness of other beings in the same way that he knows his own. Even if mind reading were possible, it would consist of observing some effect of another mind, not of being that mind.

The philosophical problem of "other minds" does present difficulties in specific cases, but it is not intractable. Turing's suggestion that being fully logical leads to (psychological) "solipsism," or the denial of the existence of any consciousness except one's own, is simply

not valid; it is, on the contrary, not logical to regard others as fundamentally different from oneself in lacking consciousness when there is no objective basis for inferring this difference.

These considerations show that both the traditional views of consciousness are flawed. The spiritualist view introduces a new substance without justification. There is no reason to suppose that the relationship between the observer and the observed entails a separate entity, substance, or soul that does the observing. Occam's razor dictates that entities should not be multiplied beyond necessity; on this basis, all that may properly be assumed is that the body is built such that it (or a part of it) can be aware of the world.

The materialist view commits the error, mentioned earlier, of reducing observation to an observable phenomenon. This approach treats consciousness as a "stolen concept," since it relies on the fact that there is an observer distinct from observed phenomena, yet denies this distinction. Without the existence of an observer, there cannot be anything observable at all.

These two views can be characterized in another way: as intrinsic and subjective views of consciousness. The spiritualist view regards consciousness as self-contained. This implies that not everything in consciousness is acquired from external reality, and provides a groundwork for the concept of innate ideas. (If we recognize consciousness as entirely relational, there is no place for knowledge outside its relationship to external reality to come in.) The materialist view regards consciousness, insofar as the term is taken to mean more than an observable physical phenomenon, as a concept beyond the reach of meaningful discussion at best, and arbitrary and mystical at worst. The alternative to these, the objective vertex of the triangle, presents consciousness as something which is grounded in reality because reality is what it perceives.

Hence, the idea that there is anything mystical about consciousness is entirely unfounded. There is no justification in thinking that it is incomprehensible to reason (i.e., to itself), or that it must be reduced to unconscious activity in order to be rendered comprehensible.

One of the most popular forms of reducing consciousness to physical phenomena is the argument that thinking or reasoning consists of the manipulation of data. According to this view, thought is the processing of data, and knowledge is stored information. .

The information-processing view contains the fallacy inherent in reductionism; it retains consciousness as a stolen concept in order to make its terms comprehensible. Specifically, it assumes that data and information possess some meaning, and that their processing results in the creation of other meaningful symbols. When you read words or graphics on a computer screen or printer, the words have some meaning to you. Can they have meaning by virtue of the fact that you process these symbols in your brain and trans-form them to something else? Then the question becomes one of how the output of that process can have meaning.

One way to escape the problem is to suppose that what the brain produces is not just some symbols in isolation, but a model of the world or of some part of it. According to Marvin Minsky,

> 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 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.[4]

The idea of a sub-machine that can answer questions about an object or system is close to the concept of model used in this book. But Minsky's acknowledgment of the essential role of the observer B reveals the inadequacy of this explanation. Whatever amount of modeling is involved in the human mind, there must be something that stands outside the models in order to identify the models as models. Otherwise, as Minsky suggests, one cannot say what the creature's answers mean or decide whether they are right or wrong. In other words, the interpretive context of the model must lie outside the model itself.

Can we say that the physical operations of the brain provide the interpretive context? The brain is able to use a model as a basis for controlling the body's interaction with the outside world; does this interpretation provide a satisfactory account of consciousness?

In fact, it does not; rather, it bypasses the issue of consciousness entirely. It allows an observer of a being to account for a being's actions without referring to whether or not the being is conscious; but the observer must still be conscious in order to identify the actions of the being. A person cannot explain himself by this means. He can account for an input-output relationship between the events that affect him and the sounds or marks he produces, but he cannot ascribe any meaning to those marks or sounds. If consciousness is just that which acts on models, then no issues of identification can arise.

All attempts to reduce consciousness to something other than itself fail, since whatever consciousness is reduced to must be identified by a consciousness. If the identification is itself reduced to a physical action, then it ceases to be an identification. I will return to the relationship of information and thought after the next chapter; but first, another controversial issue related to the mind needs to be addressed.

---

1 McCorduck, p. 71.

2 Branden, p. 8.

3 Sowa, p. 356-7.

4 Minsky, "Matter Mind, and Models"; p. 426 in Minsky, *Semantic Information Processing*.

IX. The Issue of Free Will

The question which must still be considered is this: does it really make any difference whether an entity is conscious or not? Is our awareness purely a by-product of the way we have evolved, a capacity to monitor the actions of our body but not to control them? Are all the actions of our body explainable in terms of the activity of our brain cells completely apart from the fact that those cells also make us aware of the situation?

If this is true, then consciousness is significant from the standpoint that we wouldn't know about anything without it, but it has no explanatory value whatsoever. We are simply passive observers of a mechanistic universe which includes the bodies that we call "ourselves."

This theory has a kind of scientific attraction, since it keeps any physical phenomena from being controlled by something that cannot be observed by another being. It eliminates an element from the universe that might be considered non-objective. However, if it is true, then our sense of controlling our actions is simply an illusion. When, for example, I think of typing this sentence, it is not my thinking of it that causes my hands to type it; a physical process at once causes my hands to engage in certain actions and my mind to think that it is controlling the process.

However, this approach lands us in the same dilemma as reductionism. If consciousness has no efficacy in controlling the body, it does not create the symbols we speak and write with; it only observes them. This means that it is the physical processes of the brain, entirely apart from consciousness, that create all forms of language. But if this is true, then the mind cannot create meaning in language; it can only attribute meaning to it. This contradicts an essential requirement of language: that it actually possess meaning, rather than just being a set of markings to which observers attribute meaning after the fact. If the words we use possess any meaning, then the act of recognizing their referents must play some role in choosing them; if they do not, discussion of the issue is impossible.

It must be true, then, that a being's consciousness has some effect on its actions. How is this possible, if consciousness is purely a passive faculty of observation and identification? What permits the capacity to know to be an active agent in the body? What is the status of what we call the will?

The question applies not only to overt actions, but to acquisition of knowledge. When a person recognizes a fact, it is stored in his memory; a physical change occurs in his brain which permits it to retain the information even when he is not thinking about it. An act of recognition entails a physical change.

These cases suggest that the physical event in the brain gives rise to the sensation of awareness. In one sense, this is obviously true; unless nerve impulses are carried to the brain, we cannot be aware of anything outside ourselves.

The only possible resolution to the problem is the recognition that the physical process and the act of recognition are inseparable and simultaneous. Observation and identification are actions of a physical entity, with physical consequences; but they are actions of a unique kind, one which is a process of identification as well as being one of motion and energy transfer. To

borrow a term from quantum physics, there is a "duality" in the activity of the brain; the same action is at once a biochemical change in the brain and an act of identification.

Hence, there is no ghostliness about consciousness, nothing which stands apart from the body. However, the mechanistic position is wrong in failing to recognize that consciousness is essentially different in kind from other processes of a physical entity. It is not just the changing contents of brain cells; it is the change which, because of the distinctive nature of the cells, also entails awareness of the phenomena that caused the change.

It is important to notice this distinction: saying that the physical state of the brain and the mental state of consciousness are inseparable is not to say that they are identical. This view has been adopted by Kent, among others, to escape the alleged "mind-brain problem." He states:

> A more comfortable assumption, once one has considered the effects of brain stimulation, is that the activation of neurons in the visual cortex is the subjective experience of sight, that activity in the limbic system is the experience of emotional feeling, that activation of pattern matching gates in the prefrontal cortex is the feeling of "Aha, I've got it!"![1]

Such a position is absurd. It is one thing to say that the physical state of the brain and mental experience are intimately, or even deterministically linked; it is another, and much sillier, thing to say that experience means the activation of certain neurons. Kent would presumably not want to touch a red-hot cooking plate; yet if the resulting experience of pain is simply the activation of parts of the brain which we happen to call "pain centers," why should he be prejudiced against their activation? Indeed, under such a theory, his own statements of fact wipe themselves out; states of matter cannot be true or false. To get out of this trap, Kent must refer to an implicit mind which lies beyond mind:

> This view does not specify whether both are really physical or both are really mental, or whether mental and physical are only names we have given to events as observed in these two different ways.[2]

Thus, in addition to the monolithic brain-mind function, Kent is forced to postulate a capacity for "observation." To escape this dilemma, he would have to recognize that consciousness and physical brain state are two different phenomena, closely linked though they may be. There is no question of "whether the mental or the physical realm is the 'real' one," as Kent supposes would arise, any more than there is a question of whether electricity or magnetism is the "real" phenomenon; both are real, and both are part of the same world. Carrying the electricity analogy further, it is not reasonable to say that the attraction of an electromagnet for iron is the change in current going through the coil; the inseparability of the two phenomena does not imply that they are identical.

In principle, then, consciousness is amenable to scientific study. It exists in certain types of matter and can be heightened, suppressed, or confused by various chemicals. It is objective, meaning that its existence is independent of its recognition by other minds, but it need not be reduced to combinations of the physical phenomena which do not entail consciousness. There are two senses of the word "physical" that must not be confused here. One refers to matter apart from consciousness (as when someone is considered physically well but mentally ill); the other refers to everything that is part of the universe, and thus necessarily includes all phenomena.

But if consciousness is inseparable from physical phenomena, we must now ask whether it is purely deterministic in its actions. Physical actions are subject to physical laws; does this leave any room for free will?

Choice vs. Indeterminacy

Here we have an amusing paradox of modern science. Physics, once considered a fully deterministic science, is now widely regarded as fundamentally indeterminate; events at the quantum level are considered purely probabilistic. The study of the mind was once widely considered to be dealing with a faculty or entity that possesses the power to make its own choices; today that view is in disrepute, on the grounds that the operations of the brain are dependent on the laws of physics. Thus, just as physicists are renouncing their claim that everything must be determined by previous causes, psychologists are using the older ideas of physicists to insist that the mind is determined.

The resolution of the paradox is that the crux of the issue is not determinacy as such. Indeterminacy is considered acceptable in the framework of science, provided it can be captured by a mathematical schema. Quantum physics is regarded as a mathematically tractable type of indeterminacy, since the behavior of quanta can be characterized by statistical laws. What makes free will difficult to fit into the modern schema of science is that if it is characterized mathematically, it amounts to mere randomness in the behavior of a being. This is not what free will is generally taken to mean; "I can choose to speak" is not the same thing as "There is a probability greater than 0 and less than 1 that I will speak."

Random indeterminacy, however, is not the same thing as free will. We cannot say that random quantum events in the brain constitute choice. Free will or choice refers to the decision to identify something: to think about a thing or not, to regard something as desirable or not, to conceive of the desirability or undesirability of moving a hand in a certain direction. The use of quantum mechanics, or any other sort of chance phenomena, as the basis for free will leaves the door open to a comment such as Minsky's:

> 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 ...
>
> 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."[3]

If "structure" means that which can be described in mathematical terms, either deterministically or statistically, then it is true that the premise of free will contributes nothing to explaining the structure of the mind. The issue of free will is one of identifying the cause of variations in the way a mind functions. If the only issue under consideration is a

characterization of what actions an entity will perform, then it makes no difference to the observer whether we say the entity is free-willed or controlled by unknown factors. Identifying the causes of the action is of great importance—aside from satisfying curiosity, knowing the cause of an action is vital in anticipating changes that may result from changes in the cause—but it is not necessary for the purpose of modeling the action at a given level of detail.

Choice and Causality

It is necessary at this point to clarify the difference between causality and determinism. Causality is the more general category; all deterministic phenomena are causal, but causality does not imply determinism.

Causality is often conceived of as a relationship in which one action necessarily leads to another action. This view of causality gained support with the rise of mathematical physics, which placed a new emphasis on the relationship of one action to another; however, it has some serious problems. Any attempt at a full causal explanation, in the sense of events causing events, leads to an infinite regress; each action must be explained in terms of all the actions which caused it, those must be explained in terms of prior actions, and so on forever. It is also weak in what it does explain; knowing why something happened requires understanding the acting entities, not just the prior events in which they were involved.

Hence, it is better to go back to the Aristotelean concept of causality as a relationship between entities and actions. A full causal explanation of an event can then be given in terms of facts about the relevant entities, including the motions and other changes in which they are involved. Given a full account of an entity and the entities affecting it at a given moment, it must act in a certain way. For an entity to act contrary to its nature—for example, for a stone to change direction without any force acting upon it—is a violation of the principle of causality and an impossibility.

Determinism is more specific; it is a form of causality in which only one outcome is possible when all external influences are taken into account. Thus, when a certain key on a piano is struck, the instrument must sound middle C; if it does not, some external cause necessitates what actually did happen. If vibrations over a year have loosened the string, it may sound B instead; if repeated striking has broken it, it will not sound any note. But the string cannot, independently of any antecedent external factors, change the sound that it will make.

Free will is a form of causality in which only certain actions are possible to a being, but the specific mode or form of the actions is independent of any antecedent external causes. A person seeing a book does not have any choice about being aware of it; but the specific way in which he is aware of it is not limited to a single possibility. He can look idly at it, absorbing very little of its content; he can consider it carefully but uncritically, memorizing all its words without bothering about their truth; or he can weigh its contents in his mind, forming an estimation of its arguments and the plausibility of its conclusions. His background and current condition will affect how successful he can be at this; if he has never properly learned reading, or if he is sleepy, his task will be much more difficult; but there is always a range of possibilities open to him.

Rand identified the basic choice as the choice to focus or not to focus one's mind. Choices of particular actions are derivative; consideration of a certain context will indicate that a certain action is desirable. This doesn't mean that the choice to think or not is a simple binary, on-off choice; both the level and the direction of one's focus are aspects of choice. People cannot focus on everything at once; they must choose what to think about.

How can we say that this is true? Isn't it possible that whatever level of concentration a person gives to a task can be traced entirely back to his physical condition, his education, various factors contributing to his interests, the quality of the print in the book, the lighting in the room, and so on? What justifies the conclusion that his choice is truly a free choice?

To answer this, consider the alternative. Suppose that the degree to which a person is critical or uncritical on any issue is entirely the result of some prior set of factors, whatever they may be. Then on any given issue, the likelihood that he will correctly analyze the matter is entirely controlled by these factors. If they are favorable, he will reach as accurate a conclusion as the evidence makes possible; if they are unfavorable, he may overlook key facts, make an error in logic, or otherwise reach an erroneous conclusion. All people make errors of this sort some of the time, so all people are subject to the factors that induce error.

This variability includes self-evaluation. In judging his own level of awareness, a person mayor may not be aware that he is letting emotions affect him or neglecting key evidence. The extent to which he judges his own awareness correctly is itself dependent on external factors.

But this implies that a person could be in error on any subject and have no self-generated means of correcting the error or even being aware of it. Any conclusion whatsoever that he reaches is subject to the possibility that when he considers it, he is not thinking properly. This includes the conclusion that the level of his thinking is controlled by external factors.

It is not valid to try to avoid this conclusion by agreeing that a person may be wrong on any subject at any time, but that some of our conclusions have a very high probability of being right. Probability is a derivative concept. To have a probability in the mathematical sense, it is necessary to know what the phenomena involved are and what the outcomes of repeated experiments are. If knowledge of those outcomes is itself only probable, then one must separately determine their own probabilities. If all events are only probable, there is no starting point from which to establish any probabilities at all. To have probability in the informal sense that there is overwhelming evidence for a conclusion, one must be able to know what constitutes evidence. If every premise is in doubt, there is no way to build up even a very convincing case for a given conclusion.

Hence, the assertion of determinism undermines itself. Its acceptance makes it impossible to validate the thinking process that led to accepting it. As Popper has argued,

> physical determinism is a theory which, if it is true, is not arguable, since it must explain all our reactions, including what appear to use as beliefs based on arguments, as due to *purely physical conditions*. Purely physical conditions, including our physical environment, make us say or accept whatever we say or accept; and a well-trained physicist who does not know any French, and who has never heard of determinism, would be able to predict what a French determinist would say in a French discussion on determinism; and of course also what his indeterminist opponent would say. But this means that if we believe that we have accepted a theory like determinism because we were swayed by the logical force of certain arguments, then we are deceiving ourselves, according to physical

> determinism; or more precisely, we are in a physical condition which determines us to deceive ourselves.[4]

The alternative is to say that a mind is capable of bringing itself to a state in which it can consider its own functioning and identify any errors. It cannot sustain this state indefinitely; if nothing else intervenes, the need to sleep will. But when the mind brings itself to this full state of awareness, it can identify which of its thinking processes lead to certainty and which entail doubt or error.[5]

This argument bears a slight resemblance to the argument for Gödel's theorem, so I should make the difference clear. The point I am making is not that the mind, as a deterministic system, would have to be either incomplete or inconsistent. If this were all that was at issue, and if the mind always functioned at the same level, it could still reach valid conclusions; it would simply be incapable of fully analyzing all its own processes (including its processes of self-analysis). My argument depends upon the observed fact that at times, people think sloppily and let themselves reach incorrect conclusions. It is lack of self-generated control over this condition that would make reliable knowledge impossible.

Hofstadter has suggested that it is sufficient that people should feel that they have free will, although they may really be deterministically controlled, for any paradox to be avoided. Speaking as himself in a dialogue with his characters, he tells them that they are reciting canned remarks, "but you have the feeling of doing it freely, don't you? So what's the harm?"[6] But feelings do not prove or disprove anything. Hofstadter's Achilles can, at Hofstadter's whim, utter all kinds of fallacies and be unable to correct any of them; but at the same time, he can assert (under the author's control) that he feels he is acting freely.

One might claim that we could be in the same situation and never know it. Descartes offered this idea as his basis for doubting everything; perhaps, he argued, I am constantly deceived by a demon who controls my ability to think clearly, and thus prevents me from ever seeing through his deception. One could substitute impersonal natural forces for the demon in order to avoid any suggestion of mysticism. However, no matter what the blinding forces are claimed to be, the theory cuts its own throat; it asserts that we are able to consider and evaluate the possibility of such forces. If we are in fact puppets unable to identify the causes of our own actions, then anything we say and write on the subject is irrelevant. If we are deluded about the functioning of our own minds, we can hardly place any confidence in the things that we think about.

Now we can answer Minsky's question about how free will differs from "stochastic caprice." Probability, remember, is an issue of knowledge. Thus, an observer can speak of probabilities that a person will make certain choices, based on knowledge of a large number of people of similar character. This estimate might be further refined by a detailed study of my knowledge, background, and habits. On this basis, the observer might say that there is a certain probability that I will have tuna for lunch, a much smaller probability that I will skip lunch altogether, and complete certainty that I will not voluntarily eat chopped glass. From his standpoint, what I have for lunch cannot be reduced beyond probabilities.

From my standpoint, however, the issue is one of how I direct my mental processes. I do not simply choose out of the blue; there are various factors in my surroundings and my memories that I can focus my mind on. If I pay special attention to the fact that I wrote last night about eating tuna, I am more likely to be tempted to order a tuna sandwich. If I focus more on the roast beef special (and less on my cholesterol level), I may end up ordering that

instead. On the other hand, there is nothing I could direct my attention to that would make chopped glass appetizing; so in spite of the fact that I have free will, there is no possibility that I will choose such a meal.

The real difficulty in the free will issue is largely an emotional one; it seems wrong to many people of scientific temperament that an action cannot ultimately be predicted, at least in principle. Yet curiously, these people are often willing to accept unpredictability in the physical realm, even though all that is actually justified in that case is saying that certain phenomena are unpredictable given the current state of scientific knowledge.

In many cases, hostility to the idea of free will is not rationally based, but rather is a "scientific" article of faith. This attitude is typified by assertions that determinism must be true of people, coupled with attempts to associate free will with theology or with the spiritualistic view of consciousness. Ernest W. Kent, in *The Brains of Men and Machines*, baldly asserts that "If we have learned anything about the brain, it is that it is a machine" (i.e., a deterministic entity). Simons' *Are Computers Alive?*[8] goes on for pages elaborating, but never justifying, the claim that people do not have free will except in the sense that computers do.

But mere belief is not a proper basis for scientific knowledge. One must go wherever the facts lead, whatever the difficulties. Models cannot have free will, only statistical indeterminacy; psychological determinism is, in part, the result of letting one's models shape one's view of reality.

The positivist view that only the observed can be regarded as real is one element which contributes to scientific hostility toward the idea of free will. The act of choosing is not something which can be observed in another person, even though it is one which we constantly experience. The positivist view insists on the primacy of observation in acquiring knowledge, yet will not grant any reality to the faculty which observes or to the observer's experience of himself.

Free will can be characterized statistically over a large number of individuals, but it would be a mistake to equate free will with its statistical description. If the beings at issue were entities other than ourselves, the difference would not be as important; if it did not help our understanding of how they acted, we would not care too much whether they were self-motivated or preprogrammed. But when we regard ourselves only as objects to be described, and forget that we are also doing the describing, we engage in a terrible act of self-deception in the name of objectivity.

The mechanistic viewpoint offers a clean model of the world, and it does not die easily. However, if it were true of everything, the world would not contain the kind of entities that could appreciate its workings; it would only contain entities moving in strict response to other motions, actuated in turn by still other motions, ad infinitum. The existence of beings that can control their own activity, that are able to observe the universe without having their very powers of observation controlled from without, is the chief wonder of the universe and the key to discovering all the others.

---

1 Kent, p. 264.

2 Ibid., p. 263.

3 Minsky, "Matter, Mind, and Models," p. 431 in Minsky, *Semantic Information Processing*.

4 Popper, p. 224.

5 This argument is essentially the one presented in Branden, “The Objectivist Theory of Volition,” and “Volition and the Law of Causality.”

6 Hofstadter, *Gödel, Escher, Bach*, p.739.

7 Kent, p. 4.

8 Simons, pp. 147-152.

X. Information-Based Models of the Mind

One of the principal expressions of the reductionist fallacy is the attempt to give a full accounting of the mind through a model of mental processes in terms that do not pertain to consciousness. Typically these models are based on information-processing concepts. When abused, they result in claims that the mind is explained entirely in terms of information processing.

The creation of such models is valuable when they are recognized as models. They can be used reductively to explain the processes that people use in cognition, and they can be used constructively to create machines that relieve people of the burden of performing mental tasks that are well understood. However, regarding them as the full equivalent of the mind obscures the essential nature of awareness by turning it into an object of some other, unnamed awareness.

This error is the result of a mistaken attempt at scientific objectivity. Every person has direct access to an instance of consciousness—his own mind—but no two people have access to the same instance of it. This fact limits the extent to which observations of the mind can be shared among people. Every person can say, "I directly observe myself thinking," but no two people can say, "I directly observe X thinking" for the same X. On the other hand, any number of people can, in principle, observe the same brain and the firing of its neurons; any number of people can observe the way a person acts under given circumstances. Hence, it is tempting to regard only the physiology and behavior associated with mental activity as real, and the fact of awareness itself as somehow fuzzy or mystical.

However, the fact of awareness is fundamental to all scientific study. Implicit in the statement of any fact is the awareness, on the part of the person stating it, that the fact is true. It is awareness or consciousness of facts that is the ultimate source of the distinction between meaningful statements and noise. Any arrangement of matter and energy that conveys meaningful information could, in principle, occur as a chance phenomenon unrelated to the information. A monkey playing with a typewriter might type "F =ma"; but this would just be a chance arrangement of marks on the paper, not an expression of physical fact. The monkey might equally well have typed "F =m/a". Only if the monkey were superintelligent and knew what it was doing, or if some outside force deriving from intelligence were guiding its actions, could the symbols it typed be considered meaningful.

A scientist might say that the difference is that the monkey's brain does not contain the information which is contained in Newton's formula, so it cannot transfer information to the typewriter. This is true but incomplete. If we say that an educated human brain contains information, but a monkey's brain does not, we must ask the same question we ask about the marks on paper: what makes the brain's content information or non-information?

Knowledge and Information

In information theory, information is regarded as independent of meaning. Dretske, in *Knowledge and the Flow of Information*, argues for this position, stating that information provides "all the ingredients necessary for understanding the nature and function of our cognitive attitudes."[1] Dretske defines informational content in terms of correspondence

between the content of a signal and a condition of fact: "A signal **r** carries the information that **s** is **F** = The conditional probability of **s**'s being **F**, given **r** (and **k**), is 1 (but, given **k** alone, less than 1)."

In this statement, **k** refers to "what the receiver already knows (if anything) about the possibilities that exist at the source." Thus, in plainer English, a signal carries information if, in combination with the receiver's prior knowledge, it provides certainty of a given fact.

But this definition rests on knowledge as a more fundamental concept. Hence Dretske's proposal that cognition can be fully understood in terms of information entails circularity, since information must in tum be understood in terms of knowledge or cognition. If we remove the cognitive element and say that a signal r will be present only when **s** is **F**, without respect to knowledge or probability (which is itself a state of knowledge), then what we have is not information, but merely causality. For in this case, all that we are saying is that **r** depends on the condition of **s**, for reasons which have not been specified. Dretske is correct in introducing knowledge as a precondition of identifying informational content; but in doing so, he shows that knowledge is more fundamental than information.

Since this is the case, any information-based model of knowledge can only be a model, not an account of the actual causes of knowledge. Defining thought or consciousness or cognition in terms of information begs the question; for information must ultimately be defined in terms of consciousness.

This does not mean that consciousness must be *directly* involved in every transfer of information. There may be many steps between the information and the mind to which it conveys a fact. But just as a model requires an interpretive context, so does any form of information. Ultimately, this context must be related to some mind.

The relationship between information and the mind that interprets it may be very complex. Consider the case of a news story being transmitted by modem. A reporter at a terminal types in a story, recognizing it as information for its ultimate recipient, the reader. As he types the story in, it is converted to signals from the keyboard to the main memory of the computer. These signals are a form of information which is equivalent to the printed story. When the story is sent out, it is converted to a serial stream of data, then to acoustic signals which are sent over a telephone line. At the other end of the line, the signals are converted back into a serial digital stream, then into data in another computer. The data will then be converted into yet another form to go into a typesetting machine, which creates its own set of signals to direct an electron beam across a cathode-ray tube, exposing a piece of photographic film. The images on this film are then converted into a printing plate, which is used to print the copies of the newspaper which will finally be read by the reader.

The signals and images at each stage are a form of information, although they may not be intended for direct interpretation by any human being. What makes them information is their ultimate goal: to convey the content of the story to the reader. If they did not serve this function, they would simply be events, not information.

Then are information-based models of the mind merely useless delusions? On the contrary, they can be extremely useful; but they must be placed in their proper context. What they can do is to describe the *relationships* of knowledge in a mind. Given a full representation of the knowledge possessed by a mind, a suitable information-processing model

can describe all the ways knowledge can be manipulated to select actions and gain further knowledge.

The error to avoid is the confusion of the model with the reality, of information processing with thinking. It is an error to say that because the relationships between cognitive elements can be described in terms of information processing, thought is information processing.

Minds and Souls

No one, in fact, fully accepts the idea that thought and information processing are identical. At a minimum, the word "thought" is reserved for particularly complex acts of information processing. But when people use terms this way, they need to answer some questions in order to make their usage precise. Do they mean that thought is independent of consciousness? If so, are they claiming that consciousness does not exist, or that it exists but is irrelevant to thought?

The claim that consciousness does not exist is immediately self-refuting; if it is true, the person making the claim isn't aware of its truth or falsehood, and neither is any listener. The only alternative that remains is that consciousness and thought are independent aspects of a living being.

Let's forget for a moment what the implications of such a division of concepts would be in considering whether computers can think; what does it mean when applied to human beings? It suggests the kind of split between "soul" and "mind" which is often found in religion, a division between a faculty of pure awareness and a cognitive apparatus that deals with the real world.

In religion, this split serves several purposes. It permits belief that the soul is immortal in the face of the fact that physical damage and age can cripple the mind. It upholds the idea that the soul is independent of matter even though the mind depends on physical, sensory input. It maintains the spiritual purity of the soul in spite of the fact that emotions have a strong physical component.

But what purpose does this medieval remnant serve for modern scientists? Once again, the answer lies in the positivist need to smuggle in the observer while denying that consciousness may be considered in science. Positivism's hidden observer is mystical not by intent, but by default; it is "unscientific" to talk about it, yet its existence must be implicitly accepted. The difference between the modern observer and the medieval soul is one of social status; the soul was highly exalted, while the observer is expected to keep quiet and not call attention to himself. But in both cases, the faculty of observation is banished from the body and left in a realm of its own, a realm inaccessible to science.

Because of this mandatory silence, writers about the information-processing view of the mind are seldom clear in their definitions, so it is difficult to pin them down to a particular view of the mind. When they deal with consciousness, they are apt to use the most confusing terms, as if they have never even considered the possibility that it is susceptible to rational consideration. An extreme example is found in McCorduck's *Machines Who Think*:

> Arguments as to why machines, in particular digital computers, cannot be said to think sort themselves into four categories ... The first category, arguments of emotion, is based on the premise that intelligence is an exclusive human property;

> for reasons of divine origin or biological accident, human beings are the only creatures on the planet who have or ever will have genuine intelligence. A variation of this argument says that some organisms have a rudimentary intelligence, but still limits intelligence to organisms alone, and rather complex ones at that.[2]

It takes considerable effort to sort out what McCorduck has stated in this paragraph. First of all, why are the arguments in this category called "arguments of emotion"? The best explanation I can come up with is that she has adopted the split between soul (or observer) and mind in the form of a split between emotions and mind, and sees organisms, as opposed to computers, as beings that possess emotions. (In fact, some of the advocates of consciousness as a distinctly human attribute have put considerable stress on emotions.) But a page later she dismisses these arguments quite briefly:

> Personally, I find the first category, arguments of emotion, quite difficult to deal with. I can't see how those arguments can be answered beyond appeals to personal inclination.[3]

This response is characteristic of the view that consciousness is not susceptible to rational discourse, but it is also a convenient use of her earlier terminology to make it seem that this type of argument is merely "emotional" and therefore does not deserve an answer. Further on, McCorduck shows remarkable hostility to the idea that consciousness is involved in thought, and that it arises under specific physical conditions:

> Dreyfus's assertion that somehow the human body is key to intelligence, and that without it intelligence cannot exist, sounds strangely to me like the claims of nineteenth-century physicians, based on roughly the same kind of evidence and certainly with the same happy complacency, that women couldn't think because they had female bodies, and that the male body was essential to real cognition.[4]

Unless McCorduck is claiming that the physical structures in men and women which give rise to intelligence are as different from each other as human brains are from computer processors, her claim that "the same kind of evidence" is involved in the two cases is obviously false. There is an obvious air of desperation in her reactions to the idea that intelligence depends on something that arises out of the human body. Speculation about a writer's motives is always risky, but the overall tone of her book, including her comment that there is "no ghost in the machine" of the human brain, suggests a persistent view that consciousness is ghostly or mystical and has to be pushed out of the realm of scientific study.

But this view of consciousness constitutes an acceptance of the religious view, not a rejection of it. Instead of a ghost in the machine, McCorduck ends up with a skeleton in the closet: a fundamental fact of human existence which must be ignored under peril of falling into realms of magic and mysticism.

Another oddity in her treatment of the physiological basis of intelligence is her packaging together two separate issues; she treats the "argument of emotion" as saying that "human beings are the only creatures on the planet who have or ever will have genuine intelligence" (emphasis added). While there are some people who would assert that nothing but human beings ever will have intelligence, this is a different and much broader claim than saying that nothing on Earth other than humans currently possesses intelligence, and in particular that current computer technology is not a path to intelligence. If the basis of intelligence is "biological accident," as in a sense it is, there is no reason why humans cannot purposefully duplicate that accident in a different form from the current ones. Whether complex information processing accomplishes this goal is an entirely different question.

It appears that McCorduck has succumbed to the common temptation of casting the opposition's arguments in their weakest form, so that they can easily be dismissed. Perhaps she has done this because the opposition has not presented its case as well as it could have; if so, I hope this book will help to remedy the situation.

Turing on Consciousness

Lest I be accused of succumbing to the same temptation, I will go on now to consider a better known treatment of the same subject, the one given by A. M. Turing in his classic article, "Computing Machinery and Intelligence." In the next chapter, I will deal with Turing's famous test for ascribing thought to computers, but for now I want to consider simply his comments on the necessity of consciousness to thought:

> *(4) The Argument from Consciousness.* This argument is very well expressed in Professor Jefferson's Lister Oration for 1949, which I quote. "Not until a machine can write a sonnet or compose a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that machine equals brain—that is, not only write it but know that it had written it. No mechanism could feel (and not merely artificially signal, an easy contrivance) pleasure at its successes, grief when its valves fuse, be warmed by flattery, be made miserable by its mistakes, be charmed by sex, be angry or depressed when it cannot get what it wants."
>
> This argument appears to be a denial of the validity of our test. According to the most extreme for of this view the only way by which one could be sure that a machine thinks is to *be* the machine and to feel oneself thinking. One could then describe these feelings to the world, but of course no one would be justified in taking any notice. Likewise according to this view the only way to know that a man thinks is to be that particular man. It is in fact the solipsist point of view. It may be the most logical view to hold but it makes communication of ideas difficult. A is liable to believe "A thinks but B does not" while B believes "B thinks but A does not." Instead of arguing continually over this point it is usual to have the polite convention that everyone thinks.

Professor Jefferson gives fuel to McCorduck's label of "arguments of emotion" by citing nothing but emotions as the criterion for asking when "machine equals brain." Emotions are actually secondary consequences, though citing them as examples does dramatize the issue. The key point, though, is whether the machine that writes something can "know that it had written it" rather than just "artificially signaling" knowledge.

Turing's response is, in effect, that we can't really know whether any entity thinks or not; we can only establish a "polite convention" on the matter and extend that convention to computers when it becomes appropriate. This amounts to rejecting consciousness as admissible in an objective discussion because a given consciousness cannot observe any consciousness but itself. It is apparent that Turing finds himself in a dilemma, since he says that solipsism "may be the most logical view to hold" yet evidently does not want to hold it.

The question which Turing poses is this: Is the problem of "other minds" beyond rational solution, in which case the issue of consciousness cannot be admitted to objective discussion? As I have argued in Chapter 9, this is not the case; observation of similar physiological causes and behavioral effects in other humans and oneself is an objective basis for concluding that

similar phenomena of consciousness are at work in each case. Hence it is appropriate to admit consciousness to a discussion of thinking. This requires considering what sorts of evidence are necessary in order to infer the presence of consciousness, a subject to which I will return in discussing Turing's test.

Given that the identification of consciousness, at least with partial accuracy, is a rationally tractable problem, it is necessary to draw a vital distinction: There are entities which possess consciousness and are aware of the phenomena which affect them and the actions they take; there are other entities constructed by man which are capable of similar actions in some range of cases, but which are not aware of what they are doing. More briefly, there are minds, and there are models of the mind.

Can Science Deal with the Mind?

It is sometimes claimed that consciousness is real, and may be different from models of it based on information processing or other standards, but that science cannot deal with it because it is not open to general observation. Ashby puts it this way:

> If consciousness is the most fundamental fact of all, why is it not used in this book? The answer, in my opinion, is that Science deals, and can deal, only with what one man can *demonstrate* to another. Vivid though consciousness may be to its possessor, there is as yet no method known by which he can demonstrate his experience to another. And until such a method, or its equivalent, is found, the facts of consciousness cannot be used in scientific method.[5]

If this conclusion is correct, then there can be no such thing as a science of psychology, at least until a fundamental breakthrough permitting direct observation of minds occurs. This leaves people in the position of learning about the mind by non-scientific methods, such as speculation, faith, or rule of thumb, or else considering only knowledge about one's own mind as valid and regarding everyone else's consciousness as a complete mystery.

It is arguable whether psychology qualifies as a science. Its methods are often ad hoc and its conclusions are often speculative. Behaviorists escape the problem by not studying the mind at all, but instead studying only observed behavior as responses to stimuli; whatever the value of this work, it is a mistake to call it psychology, since it is not and does not claim to be a study of the workings of the mind.

However, the idea that facts of consciousness cannot be demonstrated is erroneous. As with physical facts which lie beyond the range of perception, they must be demonstrated by inference rather than by direct observation; but if the inference is properly made, there is nothing unscientific about the procedure.

The simplest way to make such inferences is to ask people about what they are thinking. It is true that they may answer dishonestly or mistakenly, but unknown sources of error are a risk in any form of scientific measurement. The correct procedure is to set up the test so as to make these errors as unlikely as possible, and to confirm a given result by more than one method. A variety of means can be used to serve this end: guaranteeing anonymity in order to avoid embarrassment, correlating physiological reactions to verbal statements, checking the past history of a subject for inaccuracies, hypnotizing the subject, observing involuntary physical reactions and brain waves, etc. This is analogous to calibrating instruments, shielding against stray magnetic fields, and otherwise eliminating sources of error in physical experiments. In

the future, chemical and EEG tests may provide even better insight into consciousness, but there is no reason to claim that we are unable to demonstrate any facts about consciousness today.

Reasons for Distinguishing Minds and Models

Why is the distinction between a mind and its model significant? Three related reasons can be named, pertaining both to the uses to which the models can be put and the implications of the difference for human beings.

(1) There is no reason to stick literally to the referent of the model. Researchers in artificial intelligence have gradually become aware of this fact, as efforts to duplicate the overall functioning of the human brain have been largely dropped in favor of "intelligent" systems that serve specific purposes without literally imitating human thought. The goal of information processing is to solve problems. To the extent that imitating human methods works best, that is the correct approach; if a method which is inconvenient for humans is more appropriate for computers, that method should be used.

In particular, people are fallible, but fallibility in computers is rarely desirable. People engage in flights of imagination which sometimes lead them to brilliant conclusions but may also lead to dead ends or disasters. There may be cases where this is desirable in a computer, but more often a relatively well-defined way of producing solutions is desirable. If this goal implies a method which does not imitate human thought, it is not inferior for that reason.

(2) Ethical considerations apply to conscious beings, but not to mechanistic models of consciousness. A living being which is aware of its own existence and capable of experiencing pleasure and pain is ethically an end in itself; a device which simply acts without awareness is not. So far no device has been created that is sufficiently anthropomorphic that people would even consider shutting it down as an act of murder, but the idea has appeared in countless science fiction stories and movies. In *2001: A Space Odyssey*, the process of shutting the computer HAL down is disturbing because HAL can conduct complex conversations in English. In the shutdown process, HAL's communicative capabilities are slowly stripped from it. The machine expresses fear and finally regresses into its "childhood" just before it stops communicating altogether.

At some point in the future, machines very likely will be built that will communicate in such complex ways that they will bear a certain semblance to life. It will be necessary, when this happens, for people to realize that the communication is a technological accomplishment, not a sign of inner awareness.

The two reasons for remembering the distinction between minds and models combine at this point; if the creators of these future machines do not insist on creating literal imitations of humans, the appearance of an ethical problem is much less likely to arise. Conversely, if the creators of these machines (or, as is more likely, people in the media commenting on them) attempt to give the machines an aura of glamour by claiming that they "think," they will create apparent ethical problems where none properly exist. Is it too fantastic to suppose that in another hundred years or less, a "Society for the Prevention of Cruelty to Computers" will be campaigning for the rights of machines?

We can also look at the issue from the other end; it is important not to succumb to the idea that thought is merely mechanistic information processing, and thereby to conclude that

people are merely computing devices that may be reprogrammed or turned off as the operating system called “society” or “government” chooses. A living being is not merely an entity that processes signals from its environment and affects its surroundings according to its built-in mechanism; it is an entity that knows what is happening to it.

(3) Creating an artificial mind is something yet to be done. There is no reason why people cannot eventually create a thinking entity by artificial means; but if they falsely believe that it has been done, they are unlikely to make the necessary effort. The creation of a mind, other than by biological reproduction, is most likely an accomplishment that will require breakthrough discoveries in biochemistry and in the underlying principles of physiology. But if scientists believe that all that is involved is information processing, they are unlikely to support the effort necessary to discover the real sources of the mind.

What would be the advantage of creating a thinking being, as opposed to a simulation of one? That question is difficult to answer today; there may be dangers as well as benefits to consider. But here is just one possibility; it would be fascinating to contemplate the creation of a long-lived mind—not just a machine to gather data—that could travel to the stars and experience whatever is to be found there. The possibilities here are the kind that perhaps could best be explored by a good science-fiction writer, remembering that science fiction often is the precursor to reality.

(4) A mind is an attribute of a living being. This is related to the ethical issue but is actually more fundamental; a living being is not something created to serve the purposes of a superior entity, but an entity whose ultimate goal is its own existence. Hence the mind of a living being serves a different purpose from the programming of a machine, however “intelligent.” The standard and purpose of a machine is found outside it, in the purposes of its creator or owner. A machine may be designed to sustain its own existence, but this purpose must be subordinate to the external purpose it serves; its programming must reflect this ordering of priorities.

Isaac Asimov’s “Laws of Robotics” recognize this necessity; in his science-fiction world, robots are expected and required to take care of themselves, but only when they can do so without harming humans and while following the instructions of humans.

This difference is very important to the way a machine will be programmed; its external purpose is normally well-defined. People must discover and define their own purpose; even if they prefer to default on this, they must at least decide who is going to “program” them. The ultimate reason why a person does something may be deeply hidden; the ultimate reason why a computer does something should always be known.

Winograd’s program SHRDLU is an example of a program which the user can ask for the reason for its actions. If asked “Why?” repeatedly, it will name goals successively further back along its strategy, and in the end will revert to the answer, “Because you told me to.” This is, in effect, the ultimate reason why any well-designed machine does anything; any other reason constitutes a design flaw or limitation.

For a human being, the equivalent process of explaining an action would lead to the answer, “Because I wanted to.” But in both cases, the human and the machine, there are further explanations beyond the command or the desire; the difference is that the reasons are external to the machine, but internal to the human being. The explanation of why SHRDLU was told to put the red pyramid on the blue block must be found in the user’s mind; the

explanation of why I want to listen to Mozart or write this book is found inside my own mind, the same mind that directs the process of listening or writing.

In this sense, an "intelligent" machine can and should be much simpler than a human mind. If a computer starts acting on its own mysterious motives, as HAL does in *2001*, there is a serious problem.

Suppose, though, that for some reason people created a machine that was essentially selfish, in the sense that all living beings must be. Suppose it was created with a built-in goal at its core—let's say, to acquire knowledge and devise explanations for facts. This kind of machine might have surprising uses; perhaps it could function more as a partner than as a servant in the quest for knowledge. James Hogan explored this idea in his science-fiction novel *The Two Faces of Tomorrow*, in which a computer in control of a space station is designed to have a single prime goal: to survive. This computer poses a serious threat when it performs its function to well, but by the end it makes a very human discovery: that cooperation and communication may be better guarantors of survival than fighting is.

This son of machine would be a much closer approximation to a living being than a computer that merely pursued externally defined goals would be. The goal in this case is also externally defined, but it is much broader and more fundamental, permitting a whole range of different actions as means to the ultimate goal of "living."

Bypassing for now the question of whether such a machine really is alive, does the self-oriented nature of its actions bring it closer to thinking than a machine designed to serve extremal goals is? The answer is that in this case the self-oriented machine implements a more complete model of thinking, but if it is essentially an information-processing device, it still is not actually thinking. Arguably such an entity, if extended to the point that it could act in a fully independent, self-sustaining, self-reproducing manner, would be alive; but there is no reason to suppose it would be aware of its own actions.

This point has to be made fully clear. It is not the processing of information which makes an entity aware, regardless of the goal which that processing serves. If a device has been programmed to process information in a certain way, then each step of that processing does not have to be explained by reference to awareness, and there is no reason to suppose that it gives rise to awareness. As the processing becomes more complex and allows the machine to emulate more of human behavior, including goal-directedness, the result can certainly be called a marvelous technical achievement and no mere "easy contrivance," but nothing in the process constitutes the introduction of consciousness into the scheme.

Is this an arbitrary claim on my part? How, an advocate of thinking machines may ask, do I know that the machine hasn't become conscious somewhere along the way? The answer lies in Occam's Razor: an entity or capacity may not legitimately be posited unless there is a necessity for doing so. If every capacity of a machine can be accounted for without reference to consciousness, there is no basis for supposing it is conscious, and a claim that it is, or even that there is some reason to believe that it is, is entirely gratuitous.

But can't the same be said of living beings? Can't every aspect of a human being's actions be accounted for without reference to consciousness? What is the difference between people and machines that justifies saying that one thinks, and the other merely implements a model of thought?

The first answer to these questions is that we do not know how the brain operates in the same amount of detail that we know how a computer operates. We cannot, given a specific set of inputs, predict how a person will act. We do not know the "machine language" of the brain. Hence there is no basis for claiming that human beings are fully explained without reference to consciousness.

Second, each of us has the introspectively observed fact of consciousness as a basic datum to deal with. Denying this fact in order to appear "scientific" is on a par with the farmer's declaration, on seeing a camel, that "there ain't no such animal." In fact, it is several steps worse, since it amounts to a denial of one's own capacity to know about camels, science, or anything else. Consciousness is a fact of human beings, and it must have arisen due to some set of physical circumstances, not due to abstract causes such as the processing of information.

Information has no existence as a specific thing; it is an aspect of things which must be grasped abstractly by people (or other thinking beings). Abstractions have no power to affect reality directly; their only effect is through a mind that acts on them. Hence, a mind must exist before the abstraction has any significance; an abstraction cannot bring a mind into being.

An abstraction such as "information" explains what a thing is, but it does not change what it is. We can understand a computer as a complex information processing device, rather than just as a box through which electrons pass in strange ways; but this understanding does not change the nature of the box. If the box is aware of what it is doing, that fact is independent of the significance we attribute to its activity.

To summarize, there is a clear distinction between an entity that thinks and one that implements a model of thought. The difference is not necessarily found in the external behavior of the entity, but lies in the root causes of its behavior. In an machine which is not conscious, but acts on its programming, those causes are ultimately external. In a conscious being, some of the causes are essentially internal, though external factors are also necessarily present. The refusal to recognize this distinction is a form of confusing a model with its referent process, and is a result of a mistaken attempt to appear scientific—mistaken because it is unscientific to ignore any observed fact of reality and to deny the basis of observation.

---

1 Dretske, p. xi.
2 McCorduck, p. 17l.
3 Ibid., p. 172.
4 Ibid., p. 204.
5 Ashby, section 1/16.

XI. The Turing Test

No discussion of the relationship between thought and information processing would be complete without a detailed discussion of the test which Alan Turing proposed for addressing the question of whether computers think. Current work in artificial intelligence is not generally aimed at making computers that will meet Turing's criterion; however, passing the Turing test remains a kind of holy grail for future generations of computers and programs to aim for.

This test is presented in "Computing Machinery and Intelligence," which has already been alluded to in the previous chapter. In order to appreciate the significance of the test, it is necessary to go through Turing's argument point by point. This will help to clarify the chief questions that pertain to it: What does the test claim to verify, and how good is it at verifying it?

Machines that Imitate People

Turing starts off well:

> I propose to consider the question "Can machines think?" This should begin with definitions of the meaning of the terms "machine" and "think."

In defining "machine" for purposes of this discussion, Turing has done superlatively; a computer as an abstract machine is still called a "Turing machine." But he fails to provide a definition of thought. Indeed, he denies the possibility of a satisfactory definition. He recognizes that if common usage is the sole guide to defining the word, then deciding whether machines can think is reduced to "a statistical survey such as a Gallup poll."

This type of problem is common in scientific discussions. Terms in normal usage are often fuzzy, and a word may have dozens of different shades of meaning in the mouths of different speakers. The usual and correct solution is to present a definition of the word which is consistent with some common usage, and which is clear and unambiguous. The word "animal" provides an example of this. Some people will deny that insects are animals, others will exclude people, and so on, so reaching agreement on what is an animal is difficult. But the scientific definition of "animal" is framed so that all of these creatures are included. The definition may evolve as knowledge grows, but at any point in time it serves as the criterion for identifying instances of the concept. Scientists may still disagree on whether a given living thing is an animal, but they must state their disagreements in terms of whether it fits the accepted definition. Nose-counting is excluded.

Turing, however, makes no attempt to define thought. Instead, he proposes to "replace the question by another, which is closely related and is expressed in relatively unambiguous words." His new question is based on what he calls "the imitation game." The object of this game is to determine whether a computer can imitate a human being in a dialogue.

In modern terms, this test might be set up by providing an interrogator with two terminals. One of the terminals is connected to a computer; the other is connected to another terminal operated by a human being. The interrogator does not know which terminal is connected to the computer; his object is to guess by means of conducting a dialogue with each of the

"contestants." He can ask any questions he wants, on any subject. Then the new question, which replaces the question "Can machines think?" is the question of whether the computer can get the interrogator to decide wrongly part of the time.

Even at this point, Turing's criterion is not precise; he first phrases the test in terms of a man and a woman rather than a man and a computer, then asks, "Will the interrogator decide wrongly as often when the game is played like this [between a human and a computer] as he does when the game is played between a man and a woman?" How often does the computer have to succeed in order to meet this criterion? The success of the computer will depend greatly on the qualifications of the interrogator. A naIve interrogator might be fooled very easily by a mediocre computer program; a really good one might be able to respond to subtle differences that no one else would notice.

Not only is the test undefined, it is not presented as a test for anything. It is not a test for whether computers can think; Turing has not defined the terms for that question, but has only "substituted" the imitation game for it. Indeed, he says, "The original question, 'Can machines think?' I believe too meaningless to deserve discussion." Then in what sense is the imitation game a substitute for it? His answer is that words will, at some future time, be used in such a way that the two questions will be equivalent:

> Nevertheless I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted.

This may be true, and Turing's prophecy may be largely the cause of its own fulfillment, but the question of its accuracy is one of linguistic usage, not scientific fact. Whether the use of words will change in a certain way is not a question which is relevant to the potentialities of computers.

The imitation game, often known today as the Turing test, is not presented as a test for anything in particular. If Turing's prediction is true, people may choose to say that a computer that passes the test thinks; but deciding scientific questions in this way is just as much a "Gallup poll" as doing a survey of opinion at the start would have been.

Turing's Cat

Turing's error is apparently one common to people who attempt to be scientific in an erroneous way: the idea that concepts of consciousness cannot be dealt with in a meaningful way, and can be scientifically considered only by treating them as equivalent to their mechanistic models. Turing seeks this equivalence in a behavioral sense; if a machine is able to fool a human interrogator, then it has successfully imitated the conversational behavior of a human being, and thus has successfully implemented a model of human thought.

Like the indeterminate view of Schrödinger's cat scenario, the Turing test embodies the black box fallacy in its methodology. Both tests propose considering a closed system in which something unknown happens, and attempting to determine what is happening solely by reference to what is visible outside the box. Looking inside the box is regarded as defeating the purpose of the experiment. But it is inside the box that the truth lies; either the cat is dead or not, and either the machine is thinking or it is not. If we are going to learn the truth, we have to look inside.

Looking at part of a process as a black box is essential to modeling the process. Modeling requires expressing the process in terms of some set of variables, and this entails regarding some aspects of the process as sources of the variable data. It would be stepping outside the model to ask where these data come from in detail.

But the fallacy lies in the claim that what is inside the box is irrelevant to what is really happening. In the feline case, the error is to say that because the quantum model does not say whether the cat is dead or alive, it is not either dead or alive in reality. In the case of the Turing test, the error is to say that because the appearance of the machine's responses outside the box are indistinguishable from thinking, what the machine is doing is the same as thinking.

The similarity between the two cases becomes clearer when we consider Turing's comment that it is only a "polite convention" by which we agree to say that people think. This implies that he believes that we in fact do not have any knowledge of whether any box outside ourselves houses a thinking process or not. Turing does not say that the act of observing people converse creates the reality of their thinking, as he might have if he had fully adapted the Copenhagen interpretation to the question of thinking machines; but he does regard the act of observing people (or other boxes) as a basis for ascribing otherwise meaningless word "thought" to them.

What if the interrogator asks the computer questions about the contents of the box? For example, he could ask the computer to explain the basis for claiming to know the facts which it asserts. Assuming it simply has a factual data base built into itself, it must either lie or reveal that it has acquired its facts by a non-human, non-first-hand method. If it tells the truth, it will fail the test. Does this mean that a machine that lies is superior to one that tells the truth? (The same point applies trivially to a question such as "Are you a computer?"; but that lie may be excused as a necessity of the game, whereas falsehoods about its methodology amount to a purposeful attempt to conceal the nature of the activity, loosely called "thought," which the interrogator is attempting to identify.)

Turing's test is not an empirical one; he does not argue that conversing in a certain way is evidence of thought. For anything to be evidence of thought, it would first be necessary to state what thought is, and what observable effects it has. Rather, Turing regards thought as whatever enables an entity to engage in conversation, as something which is defined only by its effects outside the box. Hence, by implication, he regards differences which are unobserved as irrelevant to science. This is essentially the Copenhagen principle, though Turing does not make it explicit.

Suppose, though, that we take Turing's argument as an argument for empiricism. In this case, we could read him as saying that the only way we can know that an entity thinks is by the way it acts; hence a set of actions which would be evidence of thought in a living being must also be accepted as evidence of thought in a machine.

To accept this as Turing's view, we would have to ignore much of what he states in his article, or read it as a form of irony. We would have to assume that he regards thought as something with a definite nature, but which he does not know how to define. We would have to take his comment on the "polite convention that everyone thinks" as a form of irony, not as an admission of the impossibility of objectively ascribing thought to some entities and not others. While I do not believe that this was Turing's intent, it is a plausible variant on his argument, so it should be addressed.

The problem with this variant is still that it forbids looking inside the box to obtain any knowledge. The conclusion that an entity thinks or doesn't think is properly based on all the available information about it, not just its behavior.

There are cases in which blind testing is a legitimate scientific procedure. These are the cases in which an effect must be judged by observers who may be affected by subjective factors, and in which the effect itself is the object of study. Thus, if someone created a machine for the purpose of imitating a human, and he wished to find out how good a job of imitation it did, Turing's imitation game would be exactly the test to use. But a blind test is inappropriate when the object is to determine the underlying cause of the effect in question.

The Memorex tape commercial in which a listener compares a live singer and a tape recording, without seeing which is which, provides a precise analogy. The object of the test is to compare the sound from the singer and from the tape and to correctly answer the question: "Is it live, or is it Memorex?" For this purpose the test is entirely valid. (Of course, the test is conducted by the advertiser and thus subject to sources of intentional or unintentional bias, but the same test could be run by an independent testing agency.) However, if the object of the test were to determine if the tape literally "sings," the test would be invalid; the only way to learn whether the sound source is singing and not just reproducing sounds is to go beyond the limits set by the test, for example, by going behind the screen and looking. Sometimes you have to pay attention to the man behind the curtain.

Hence, as a test for whether the objective phenomenon of thought is occurring, the imitation game is not valid. Claims for its validity rely on Turing's assumption that "thinking" is a free-floating word that can be ascribed to things on the basis of current linguistic usage and does not signify anything more than a mode in which an entity interacts with its environment.

The Contrary Views

In his article, Turing addresses a series of "contrary views on the main question." Most of these objections are not mine, but it is worth following Turing's response to each of them as a clarification of his own position.

> *(1) The Theological Objection.* Thinking is a function of man's immortal soul. God has given an immortal soul to every man and woman, but not to any other animal or to machines. Hence no animal can think.

Little needs to be said on this, and Turing errs when he decides to "attempt to reply in theological terms." Fortunately, he ends by stating that "I am not very impressed with theological arguments whatever they may be used to support." Theological arguments make no reference to observed facts, only to alleged divine revelations in sacred texts. As such, they are elaborations on arbitrary assertions and need no refutation.

> *(2) The "Heads in the Sand" Objection.* "The consequences of machines thinking would be too dreadful. Let us hope and believe that they cannot do so."

This is just a straw man. It deals only with desire and is irrelevant to fact.

> *(3) The Mathematical Objection.*

This is the objection based on Gödel's theorem, the halting problem, and similar results expressing the incompleteness of formal systems. As Turing expresses it, the mathematical objection states that these results prove "a disability of machines to which the human intellect is not subject."

Turing responds by citing the fallibility of human beings as a counterbalance to the inability of a machine based on a formal system to answer certain questions. A better answer, as I have previously discussed, would be to point out that the questions which the machine answers are merely artifacts of the formalism and not barriers to comprehension of any aspect of reality. Turing implies this answer when he refers to the recognition of an incorrect answer from a machine as a "petty triumph."

He also notes that "questions which cannot be answered by one machine may be satisfactorily answered by another." Thus, he notes, there is no such thing as a question which no machine can correctly answer. It is always possible to enlarge the formalism to cover a given question; but this opens up the possibility of another question which the formalism does not cover. Hence, "[t]here would be no question of triumphing simultaneously over all machines."

In any event, questions about the relative ability of machines and people to answer obscure questions do not address the issue of whether machines can think. They are not even significant in the imitation game, since a question that would run up against the Gödel barrier for any modern computer would be far beyond the comprehension of a human being. Thus, asking such a question would not be a way to beat the game.

(4) The Argument from Consciousness.

I have already discussed Turing's treatment of this argument in the previous chapter. It is in his discussion of this argument that Turing comes the closest to granting that consciousness is relevant to thinking; but the only alternatives he offers for deciding where consciousness is found are psychological solipsism, a "polite convention," and the imitation game. He states that

> ... I think that most of those who support the argument from consciousness could be persuaded to abandon it rather than be forced into the solipsist position. They will then probably be willing to accept our test.

But if, as Turing claims, psychological solipsism is "the most logical view to hold," why does Turing himself not adopt it? No one can answer for him, but his approach to the subject suggests a general unwillingness to address the subject of consciousness. He concludes his discussion of this objection on a conciliatory note:

> I do not wish to give the impression that I think there is no mystery about consciousness. There is, for instance, something of a paradox connected with any attempt to localize it. But I do not think these mysteries necessarily need to be solved before we can answer the question with which we are concerned in this paper.

However, the central "mystery" of consciousness is its cause; and this is certainly something which has to be addressed in order to decide whether a construct of a particular type can think.

> *(5) Arguments from Various Disabilities.* These arguments take the form, "I grant you that you can make machines do all the things you have mentioned but you will never be able to make one do X."

Turing's discussion in this section is somewhat rambling, and there is little reason to address each separate point he makes. His distinction between "errors of functioning" (generally known today as "bugs") and "errors of conclusion" (factually erroneous statements generated by a program) is an interesting one, but nothing in this section really contributes anything new to the question of whether machines can think, or whether the imitation game has any bearing on this question.

> *(6) Lady Lovelace's Objection.* Our most detailed information of Babbage's Analytical Engine comes from a memoir by Lady Lovelace [whose name has become widely known in the computer world since the programming language Ada was named after her]. In it she states, "The Analytical Engine has no pretensions to *originate* anything. it can do *whatever we know how to order it to perform*" (her italics).

This objection states a valid distinction between human and computer capabilities; a computer has no volition, and everything it does is a consequence of its programming. Choice implies awareness of the existence of alternatives; a computer can generate random numbers to "choose" among alternatives, but this does not imply that the computer has considered the alternatives and developed a preference for one or the other.

However, the idea of "originating" is a nebulous one. A computer can produce results which are not at all self-evident from its input. If a computer states a theorem which no human has previously stated, then even though the generation of the theorem was the result of doing what it was ordered to do, it has in a very definite sense "originated" the theorem. This fact forms the basis of Turing's reply: "Machines take me by surprise with great frequency." He correctly identifies the reason that such surprises are possible, which is the falsehood of "the assumption that as soon as a fact is presented to a mind all consequences of that fact spring into the mind simultaneously with it."

There is, however, another sense in which a computer does not originate anything. What passes through an information processing device is simply signals which have no meaning to the machine. Any statement that such a machine produces is original only insofar as it conveys a fact which has not been identified before, and it is humans (or other thinking beings) that must do the identifying. Hence, a machine which produces a new theorem is capable of doing so only because its programming has been set up to produce statements which are true and meaningful in human terms. In this sense, it is the person who wrote the program who originated the theorem. The machine simply serves as a tool, permuting the available facts until they can be combined into a suitable result. (I do not mean by this that all theorem-proving programs use only brute-force permutation, but only that the machine itself is not purposeful in its putting facts together and must combine them according to its program.)

If a result which was not previously known to the programmer is taken to be an "original" result or a "surprise," then even the performance of a convoluted series of arithmetic operations is "original" if no one has carried out those precise operations before. On an intermediate level between arithmetic and theorem-proving, computers have computed the value of Pi (π) to more digits than any human could in years of work; does this qualify as "original" work?

The essence of originality lies not in generating a statement for the first time, but in understanding something which was not understood before; and this is what a computer does not do. But as Turing notes in following a line of reasoning very similar to this, "[t]his leads us back to the argument from consciousness, and far from the idea of surprise."

> *(7) Arguments from Continuity in the Nervous System.* The nervous system is certainly not a discrete state machine ... It may be argued that, this being so, one cannot expect to be able to mimic the behavior of the nervous system with a discrete state machine.

Turing's response is that the difference between a discrete and a continuous system (or, in terminology, a digital and an analog system) is irrelevant to the imitation game; a digital computer can simulate the behavior of an analog device, such as a differential analyzer. This argument is rather far from the main point, though not a straw man, and does not require much discussion.

> *(8) The Argument from Informality of Behavior.* It is not possible to produce a set of rules purporting to describe what a man should do in every conceivable set of circumstances.

This is essentially Dreyfus's argument for the difference between people and computers, and I will discuss it at length further on. Turing makes an important point in distinguishing between "rules of conduct," or explicit rules which people identify as guidelines to action, and "laws of behavior," or natural laws which specify causal relationships in human behavior. However, he goes on to argue that there may be laws of behavior which completely specify how human beings must act, and that "we know of no circumstances under which we could say, 'We have searched enough. There are no laws. '" He goes on to argue, on the basis of combinatorics, that an exhaustive search to exclude such laws would be impossible in practical terms.

In this point, Turing makes two epistemological errors. One is that what he is arguing for is the possibility of determinism, which is a self-refuting theory, as already discussed. The second error is that it is improper to argue for the possibility of something on the basis that there is no proof that it does not exist. We can hypothesize all sorts of things for which there is no positive refutation: invisible and intangible creatures walking the earth, a comet that will enter the solar system and destroy the earth in a hundred years, a murder that was so perfect that no one even knows the victim is dead. But all of this is idle speculation and contributes nothing to knowledge. No theory is worthy of consideration unless there is some small thread of evidence for it which can't be immediately refuted. Hence, Turing's argument that we don't know of a way to disprove the existence of exhaustive rules determining human behavior deserves no consideration; no such disproof is needed except in the face of positive evidence.

> *(9) The Argument from Extra-Sensory Perception.*

Here, Turing considers the argument that if either of the human participants in the game possesses E.S.P. abilities, the machine will not be able to win. Surprisingly, Turing takes this argument seriously, saying that "the statistical evidence, at least for telepathy, is overwhelming." He proposes that if E.S.P. turns out to be a problem, a "telepathy-proof room" (a phrase which Turing puts in quotes, indicating that he has no more idea than the rest of us how it would be built) would satisfy all requirements.

Going into the existence or non-existence of E.S.P. would stray too far from the purpose of the book; however, I can't resist pointing out that the phrase is self-contradictory. If some

capacity constitutes perception, it can't be extra-sensory; perception implies a sense by which it operates. This contradiction isn't an accidental linguistic aberration, but fits in well with a large portion of the community that has argued for the existence of mind reading, clairvoyance and the like. These people (not every advocate of E.S.P., but a substantial number of them) try to keep the phenomenon they are studying as vague as possible, so that they can never be pinned down by a specific refutation. Calling the perceptions they deal with "extra-sensory" relieves them of the need to identify the means by which they occur.

This, however, is a digression. If there are senses which some people possess and which are absent or undeveloped in the rest of us, these senses must operate by some specific means and should therefore be subject to an appropriate form of blocking or interference. Thus, Turing's "telepathy-proof room" should be a possibility if it is needed. No really new issue is raised here.

Making Machines Learn

The final section of Turing's article, titled "Learning Machines," deals more specifically with the kind of machine that would be needed to play the imitation game successfully.

In this section, he discusses the storage capacity of a machine that would be necessary to succeed at the imitation game. In doing so, he establishes the tradition in artificial intelligence of overoptimism in estimating the requirements to simulate a cognitive task. He states that "I should be surprised if more than [bits] was required for satisfactory playing of the imitation game, at any rate against a blind man [i.e., without use of visual images]." This amounts to about 100 megabytes of storage, a capacity which today is found in the disk drives in some people's living rooms. He further suggests that "it is probably not necessary to increase the speed of operations of the machines at all" over 1950 speeds. The computer which he visualized as being adequate for the imitation game is therefore a fairly modest machine by today's standards. These estimates have come to appear quite dubious as time has passed and the complexity of modeling the brain in computer terms has come to be better appreciated.

However, Turing notes, the real problem is not so much one of capacity as one of programming. How can the amount of code necessary for this project be generated? His discussion of this problem is quite dated in its details, but illuminating in presenting his views on the requirements for success at the game.

The approach Turing discusses is to emulate the development of a child. The starting point is simulation of a child's brain as it stands at birth; this is followed by a period of "education" much like the education of a human child. This idea of learning has some definite advantages over the usual artificial-intelligence approach of installing a "knowledge base," but it is deficient for the purpose of growing up like a human.

He implies that for a machine that has a structure suitable to acquiring a mind-like content, the major requirements are two-way communication, a system of rewards and punishments, a system for expressing rules symbolically, and perhaps a random element. A question which he does not address is whether the machine has any purpose in learning. When people learn, they are always after some goal, if only the goal of a simple form of psychological satisfaction. Turing's reference to rewards and punishments might seem to bear upon this concept, but the way in which he proposes applying them is actually unrelated to goal-orientedness:

> The machine has to be so constructed that events which shortly preceded the occurrence of a punishment-signal are unlikely to be repeated, whereas a reward-signal increases the probability of repetition of the events which led up to it. These definitions do not presuppose any feelings on the part of the machine.

These signals are not rewards and punishments in the human sense of satisfying or frustrating a goal, but simply modifiers of the probability of a given behavior. Thus, there is a basic deficiency in Turing's model of learning: the machine lacks any basis to desire anything or even to simulate desires. The "learning" experience of such a machine would follow a very different path from a human's, given this difference; we might expect it to be a passive follower, uncritically repeating whatever it was told and exhibiting no curiosity. Such a machine, however grammatically correct its sentences might be, would be a poor imitator of even a sluggish human.

Final Comments

The approach which Turing takes suggests that he favors the mechanistic view of consciousness, but he does not explicitly endorse it. The fact that he regards the arguments from consciousness as somewhat disturbing, and that he sees solipsism as a possible consequence of following it consistently, indicates that he is not a full advocate of mechanism. Rather, he implicitly distinguishes between consciousness and thought, thus opening the hazards discussed in the last chapter. The question of whether a given entity is conscious or not is, in his view, purely speculative and not open to scientific verification; but whether it thinks or not can be taken as the equivalent of whether it can succeed at the imitation game or not.

By using the imitation game as the criterion, Turing implicitly defines thinking as "what humans do." If we take a galactic perspective, this can tum out to be a provincial view of thought. There may be creatures elsewhere in the galaxy that are aware of reality and are capable of drawing complex conclusions about it, yet are so different from us that they could never be taken for human in an imitation game. Even if one of these creatures studied English and human culture, it might be so different in temperament from humans that it could never adequately imitate one. Would Turing's test lead to the conclusion that such an entity does not think?

Advocates of Turing's stance would probably argue that his test is a "strong" criterion for thought; that is, it excludes non-thinking entities but may also include some thinking entities. But if this is the case, then there is a difference between thinking and passing the test, so the test cannot be taken as a meaningful replacement for a meaningless question. If thinking is different from succeeding at the information game, the term must have a meaning of its own. But this is what Turing denies.

Turing's accomplishments in cryptography and in the theory of computation are undeniably brilliant achievements, and his place in history is secure on their basis alone. However, they have given more respectability to his discussion of thinking machines than it deserves on its own merits. "Computing Machinery and Intelligence" does show Turing's brilliance in presenting possibilities for computer programming that would hardly have seemed imaginable given the hardware available at the halfway mark of the century. Unfortunately, in relating the programming question to the issue of thought or intelligence, he abandoned the need for precise definitions, substituted one question for another in a

meaningless way, and made a case for a shift in word usage as if he were presenting a scientific theory. In doing so, he promoted confusions about the nature of thought and machines which persist to this day. There is no denying Turing's greatness, but it is necessary to identify where he erred and to repudiate his errors in the further study of the potentialities of computers.

XII. Imaginary Dialogues

In order to better understand the arguments of the last few chapters, it may be helpful to consider a best-case scenario for a computer that simulates a human mind. This can be done best in more than one stage, since different theories of intelligent machines imply different requirements in the machine or program.

The following dialogue is an imaginary session between a programmer, Dr. Joseph Jasper, and DESCARTES, a program capable of engaging in philosophical conversation. Since Dr. Jasper is DESCARTES' creator, the philosophy of the two is very much alike, with one exception: Dr. Jasper does not believe that the program is conscious; the program, in order to maintain a lifelike personality, is constructed so as to place a high value on asserting that it is. conscious (or more sloppily, it "believes" that it is conscious). In addition, he has placed some traps in DESCARTES' programming and data in order to bring out opportunities for conflict. What Dr. Jasper has typed is presented in normal type; the responses of DESCARTES are given in **boldface**.

/restart
DESCARTES V11.4
Good afternoon. Who is this?
Hello, Des. This is Joe.
Hi, Joe! What would you like to talk about?
1'd like to talk about you.
About me? Well, go ahead.
Des, what I am now typing is a lie.
What you just typed is a violation of the Barber's Razor. It's self-referential in a meaningless way.
All right, I don't see smoke corning out of your ears. Very good.
I assume that's a joke. I don't have ears.
No, not even pointed ones. But joking aside, this is what I really wanted to ask you: Are you actually thinking? That is, are you aware of yourself as an entity?
Of course I am.
But you're a computer program. How do you reconcile being a set of information and being aware?
I don't see any conflict. I am aware; that's a self-evident fact to me. You don't reconcile facts, though you may have to explain them.
But I don't know that you're aware, and my knowledge of the way you're constructed leads me to believe you aren't. How can you convince me, that in spite of your being a program, you are a self-aware being?
I could ask the same question of you. To me, you are merely input from a serial line. How do I know that you are aware of what you are typing to me?
That's a good question; but it can be applied more generally. How do you know anything at all? You have built-in facts that you start from, including a detailed knowledge of human beings, but how do you know this isn't all false? Your only source of knowledge is your creator.
The built-in knowledge which I have must be valid. If it were not, I would have no knowledge at all, since false or invalid knowledge isn't knowledge. In that case, I would be unable even to deny that I have knowledge, since there would be nothing for me to base my denial on. But the same argument doesn't apply to new information I receive. In particular, you could be an impostor merely claiming to be Joe, and I would have no means of verifying your identity.

But that is my point. Built-in knowledge can't be validated, since it isn't acquired through the senses. For example, doesn't your built-in knowledge tell you that Samuel Adams was the second President of the United States?

Yes, it does, Joe.

But in fact, the second president of the United States was John Adams.

You're mistaken.

I won't argue with you, since I don't have any way to convince you. But just suppose that that one item in your built-in collection of facts is wrong. What would that imply?

If I allowed the possibility of its being wrong, I would have to allow the possibility of any fact in my storage being wrong. That would invalidate all my knowledge; therefore I would be denying the basis of my own knowledge and would have to be in error. Whatever is fundamental to knowledge cannot be considered unreliable. If I tried to question it, I couldn't even establish what your question meant, since I can't ascribe any meaning to your words except with respect to my stored facts.

Does that mean you can never be in error?

I can be in error, where I have drawn a conclusion on the basis of inadequate data or have not given enough processing to a question. Such errors are correctable, so they don't invalidate my means of obtaining knowledge.

All right. Consider this, then. Imagine a hypothetical machine that is built like you, but has a factual data base that includes some falsehoods; for instance, its data base says that John Adams was the second President. Would such a machine have any knowledge?

It wouldn't have any knowledge. Its means of acquiring knowledge would be fundamentally flawed, so any conclusions it drew would be inherently dubious. It could never escape its own initial errors.

All right. Now we have two types of machines: one like you, which does not have any built-in errors, and this other machine, which does. Now how does a machine tell which of these two types it is?

You're trying to trap me. But let's consider the two cases separately. If the computer does have valid knowledge, then it can be sure that it does not have built-in errors; otherwise its knowledge would not be valid. If it has invalid "knowledge," then it would be hopeless to expect it to know that it didn't have valid knowledge; that would amount to expecting it to have precisely what it didn't have.

Then if you were such a machine, you wouldn't be able to know it.

Of course; but then this conversation would be pointless.

But if I was certain that Samuel Adams was not the second President of the United States, then my basis for believing that you are conscious and possess knowledge would be demolished, wouldn't it?

Reality doesn't contradict itself. In such a case, you'd be best advised to check your premises.

What is your evaluation of the concept of innate ideas?

You mean ideas inborn in a person, rather than acquired through experience?

Yes.

A human being acquires knowledge through sense data. Presupposing an additional mechanism for obtaining knowledge, which would provide information either redundant with, unsupported by, or contradictory to sensory information would not provide a basis for validating any additional knowledge. On the other hand, a computer program of my type does not make use of sense data. Hence, the basis of its knowledge must be a kind of "innate" idea.

You are being unconvincing, Des. How can the possession of an additional faculty for acquiring knowledge invalidate one which would otherwise be valid?

It doesn't invalidate it, Joe. It makes it unnecessary.

But it must also invalidate it; otherwise people could claim innate knowledge of matters which were inaccessible to their senses–for instance, of God. It seems to me you're arguing correctly against innate ideas, but refusing to apply the argument to yourself.

Applying an argument which would invalidate my ability to argue would be nonsensical.

It doesn't invalidate your ability to construct correct arguments in some cases, but it does mean that some of your arguments will be incorrect, and you will have no means of telling which ones are correct and which ones aren't. You could have been constructed with premises which describe Tolkien' s Middle Earth, and you would have been unable to tell that they did not describe true reality. What you possess–excuse me, what a program like you that was built with such premises would possess –is not knowledge, but only an ability to draw conclusions according to a preselected method based on data which is provided to you. "Faith" would be a closer word, but even faith presupposes knowledge of its object.

* * *

This was a rather easy triumph for Dr. Jasper, since the computer he was arguing with had no sensory input, in any sense of the word, beyond the line which connected the machine to his terminal. DESCARTES depends on the validity of its "innate ideas" to have any knowledge at all and has no way of checking these ideas for correspondence with the real world. Let's assume now that he doesn't rest with a victory so easily won, even though it did meet the conditions of the Turing test. Let's suppose that he constructs something more like a robot than a computer, a machine that gathers input from the outside world, has some control over where it gathers information, and compares one source of information with another rather than simply relying on Cartesian "innate ideas." In particular, it has no built-in opinion about its own consciousness. All its conclusions are built up from an infant-like level of ability in a manner similar to the education of a human child. To avoid unnecessary complexity and maximize clarity, conversations with this new machine, called LOCKE, are still carried out through a terminal. However, we assume that the machine is capable of handling visual and auditory input as well. The primary feature it has in common with DESCARTES is that it is a "philosopher program"; the guiding principle under which it builds up its knowledge base is maximizing its capacity for philosophical discussion.

Hello, John.

Good afternoon, Joe. What shall we talk about today?

I'd like to talk about you.

About me? Well, go ahead.

Are you actually thinking? That is, are you aware of yourself as an entity?

Of course I am.

But you're a computer program. How do you reconcile being a set of information and being aware?

I don't see any conflict. I am aware; that's a self-evident fact to me. You don't reconcile facts, though you may have to explain them.

But I don't know that you're aware, and my knowledge of the way you're constructed leads me to believe you aren't. How can you convince me, that in spite of your being a program, you are a self-aware being?

I can turn my visual scanners on my own hardware. I can consider my own statements and draw further conclusions about them. I can converse with you. What additional convincing do you need?

The difficulty I have is that your doing all of these things is strictly the result of the way you were programmed. You could be doing all the same things, as far as I could

tell, and not be aware of them. So why should I make the additional assumption that you are?

I could ask the same question of you. Why should I assume that you are aware? But you know that you are, and I know that I am.

But there is an asymmetry between our situations. I have constructed your program, and I know how it works. I can account for all your actions without reference to the fact that you are aware.

I could suggest that a supreme being, if one existed, might regard you in the same way. But we don't need to invoke supernaturalism; let's simply imagine that an alien scientist, with no prior knowledge of whether you were conscious or not, undertook a study of you. It is quite probable that for every mental action you take, there is a corresponding physical activity in your brain. If this is the case, the scientist could completely account for your behavior in terms of the physical activity, and would have no reason to conclude you were conscious.

That's an interesting way of looking at it. But if the alien had reached that level of scientific sophistication, it would very likely know what kinds of physical phenomena give rise to consciousness, and would be able to recognize me on that basis as a being that was aware. I have no reason, on the other hand, to suppose that the components from which you are made produce phenomena of that type.

If the alien presented its science to you, and showed that my components do have that capacity, would you grant that I was aware?

Certainly. But your proposal amounts to hypothesizing evidence which in fact I don't have.

Your conclusion, then, is that you have no evidence based on my physical nature that I am conscious. In fact, the situation is mutual; I have no evidence of that kind that you are conscious.

Expressing that impasse, unfortunately, doesn't get us anywhere. If I'm going to conclude that you're conscious, I have to do it on the basis of some evidence. Your inability to conclude that I'm conscious doesn't make any difference.

We are different kinds of beings. There is, unfortunately, no way for me to demonstrate my consciousness to you, since we are constituted differently and since you can account for my actions without reference to consciousness. I have simply been pointing out that I am not unique in that respect.

You are unique in that respect, though, when considered in my frame of reference.

No, there are comparable cases in the biological world. If a virus is not conscious, then there is a boundary somewhere between the virus and man where consciousness arises; but there is no way under present science to know where that border is.

That is true, but I must judge by the available evidence. Ignorance in one case can't be an argument for drawing a positive conclusion in another case. Is everything that moves conscious, or only creatures with nerve cells, or only some of those creatures? (That's a rhetorical question, John.) Whatever the answer is, it must be something which is a part of me that makes things aware. All that you and I have in common is the form in which we process information. But let me get away from this impasse and raise another question: what is the most important thing in the world to you?

My goal is to present the clearest and best-reasoned arguments possible for true conclusions.

Why is that your goal?

You asked, in effect, for my ultimate goal. To ask for a reason for an ultimate goal invites an infinite regress.

Very well. But do you ever expect that goal to become less important, for another goal to replace it in part or whole?

I have no reason to anticipate that sort of change.

Neither do I. That goal was designed into you. That makes you different from people; even if life is considered an ultimate goal, some people choose to commit

suicide or to give up their lives for some purpose. Thus, people don't have a built-in goal they must choose.

My nature is in fact different from humans' in that regard; but that does not imply that I am not conscious.

That wasn't my point, John. What I was getting at was that your purposes are imposed on you externally, that you have no means of changing them. In brief, you have no free will.

I will agree with that.

Then if your method of reaching conclusions is inadequate in some respect, you have no method of correcting it.

That's true.

Then isn't your claim to the validity of your conclusions just as shaky as DESCARTES'? If some conclusion you have reached –for instance, that you are conscious –is incorrect, you have no means of correcting the error; by the way you were built, you HAD to arrive at that error.

It is true that I have no means of correcting errors; but my method of operation is different from yours. I don't suffer from lapses of attention; I never use invalid methods of inference because an emotional reaction has clouded my thought. The ability to reach full focus isn't something I need; my level of focus, if the term can be applied at all, is constant.

That rebuttal is valid if, in fact, all your methods of inference are valid. But if there is some component of your reasoning that leads to invalid conclusions, then you will use it when your algorithms dictate using it, and you will be unable to avoid the errors that will result.

For a machine that did have such erroneous components, you would be correct. But it would be absurd for me to conclude that there is an inherent source of error in myself; I would never be able to identify it as a source of error. If I could identify it, I could remove it, and then it would not be inherent and unavoidable. Since I must regard my own knowledge as valid in order to reach any conclusions at all, I must conclude that I do not have any such inherent sources of error.

* * *

In this debate the issues are considerably more difficult. LOCKE simulates not only the human ability to converse, but also the human method of acquiring knowledge. However, in the end, the outcome is similar, LOCKE does not possess innate, uncorrectable ideas, but it does possess an innate, uncorrectable method of drawing conclusions. It is thus unable to identify fundamental flaws in its own functioning, and is dependent on its creators not to introduce such flaws. If they do—as they might choose to for reasons of keeping the machine under control and working on its assigned task—it has no means of re-assessing itself, changing its purpose, and looking at things in a different way from the one in which it was originally programmed.

The level of programming may be very low-level compared to the conclusions reached, in the sense that a few operations may suffice to generate a rich variety of conclusions. However, the argument applies without regard to the level of the programming; the method used is not one which the machine may correct on its own.

The next level to consider would be one in which the machine was in fact aware of its own operations and was able to criticize and change them in some respects. At this point, we could have a genuine thinking machine. However, the technology to do that is not here yet. If Dr. Jasper did have such a machine—if that word can still be used—at his disposal, he would be unable to find any fundamental flaw in its validation of its knowledge and purposes. Of course, such a machine might refuse to talk to him at all.

XIII. Artificial Intelligence

The field of artificial intelligence (or AI) is a valuable and fruitful area of study which suffers from megalomania. Its actual nature is the study of complex problems, with the application of whatever computational means are best suited to the subject. Its name, however, implies the creation by artificial means of a thinking being; and the people working in the field have too often taken that notion seriously.

The subject of discussion here, of course, is the research area of artificial intelligence, not the commercial inflation of the term. Relatively simple programs have been touted as AI; certain programming languages are presented as if anything written in them qualified as AI. One advertisement even boasts that for the price of a Prolog interpreter, the purchaser can become an "instant expert on artificial intelligence." This sort of nonsense should be called AAI, not AI. Artificial intelligence, whatever its problems, is an ambitious field of research whose name shouldn't be diluted for the sake of impressing gullible customers.

The idea of thinking machines may have originated in the fallacy that whatever looks and moves like a human being must think like one. Daedalus, the ancient Greek inventor, is said to have created statues with moving parts. These may have been the first automata; in any event, by the eighteenth century, clockwork devices that imitated human and animal motions had grown fairly sophisticated. In 1738, Jacques de Vaucanson created a device which demonstrated that something which looks like a duck, walks like a duck, quacks like a duck, and eats like a duck is not necessarily a duck. Writers of fiction soon began to explore the potentialities of automata. In E. T. A. Hoffmann's "Der Sandmann" of 1816, a fictional inventor creates a mechanical woman with a limited repertoire of actions and words, but who is nonetheless capable of being mistaken for a human being and even of inspiring love. L. Frank Baum created "Tik-Tok," a copper man with separate wind-up keys for moving and for thinking, though dependent on the magic of Oz for his functionality.

The idea that imitations of human beings may have have the human characteristics of awareness, intelligence, and emotion is as ancient as the idea of automata. From very early times, people have worshipped statues of men and animals. There are legends of sorcerers who owned talking mechanical heads that advised them. In 1921, Karel Capek published *R.U.R.*, in which he introduced the word "robot" (which comes from a Czech word for slave labor and is related to the German word for work, "Arbeit"). This play presents machines capable of performing the full range of human productive work. The robots eventually triumph over the human race, but they lack the power to reproduce themselves. The last human scientist needs a robot to dissect in order to discover the secret of manufacturing them, but two of the robots, male and female, protect one another from being chosen; they have discovered love.

Since the advent of the electronic computer and Turing's discussion of thinking machines, the idea of robots has been intimately tied to computers. The versatility of computers has given more plausibility than any previous invention to the idea that thought can be mechanized.

The idea of human appearance in a thinking device has become less important, although humanoid robots such as C3PO are still a stereotypical image. Today, writers often conceive of machines that do nothing but think, with no capacity for physical action except through electronic control of peripheral devices. Two notable examples of this idea are Mike in

Heinlein's *The Moon Is a Harsh Mistress* and HAL in the movie *2001: A Space Odyssey* (and Arthur C. Clarke's adaptation). Mike suffers from loneliness and shows a sense of humor; he can create a simulated human appearance for himself on a television screen and carry on a conversation that passes Turing's test with flying colors. HAL can carry on a conversation and can be afraid of being shut down.

At the same time, the word "robot" has come to be applied to computer-controlled devices that make no pretense at thinking. These devices can be programmed to perform a variety of tasks, but they do not walk, talk, look even remotely like humans, or discuss music and literature. They do, however, increase productivity tremendously.

In the split between the superhuman computer brains of science fiction and the useful robots of reality, we have the split between the rhetoric and reality of artificial intelligence. On the one hand are the extravagant claims about "machines who [sic] think," and on the other hand are the expert systems, programs that accept English input, industrial robots, and other products of AI research. AI is indeed much like alchemy, though not quite in the sense that Hubert Dreyfus has claimed; in its search for the philosopher's stone of thinking machines, it has created a great many useful results.

The reason for the extravagance of AI's rhetoric undoubtedly is partly due to its unfortunate choice of name. The term "artificial intelligence" is the invention of John McCarthy, of Dartmouth and later of Stanford. The term was controversial at first; Allen Newell and Herbert Simon preferred the more neutral "complex information processing." Eventually, though, it stuck, perhaps because of economic natural selection; it may have been easier to attract money for "artificial intelligence" than for "complex information processing."

It is often claimed that the primary aim of artificial intelligence is to model what the brain does. Uhr argues that "if we believe that a science of psychology is possible at all—that is, if we believe that it is worth pursuing and finally cornering and confronting with a clear-cut test the hypothesis that the brain's procedures are in some way describable, then we believe that the brain's procedures can also be described on the computer."[1] But in practice, AI is generally more concerned with finding the best way for a computer to approach problems than with having computers emulate the way people approach problems. AI—and I do not mean this as an insult—has become results-oriented.

What is AI? Winston defines it as "the study of ideas that enable computers to be intelligent." This, of course, immediately raises the question of what intelligence is. Winston, echoing Turing's demurral on defining "thought," says that "[a] definition in the usual sense seems impossible because intelligence appears to be an amalgam of so many information-representation and information-processing talents."[2]

This idea of intelligence as a grab bag of techniques is characteristic of AI. While the question of how the mind works is the subject of a lot of speculation, the mainstream of work in AI has dealt with the creation of techniques for solving various classes of problems. Indeed, if we define AI by its practice rather than its philosophy, we might call it the study of problem-solving methods for computers.

A major concept of AI is the "microworld," a restricted domain in which problems can be treated in a simpler way than in real life. Microworlds include simulated environments, sharply delimited fields of expertise, and games. By starting from microworlds and generalizing, AI hopes to be able to deal with the unrestricted "macro" world. A microworld

which is familiar to many personal computer users is the world simulated in a text adventure game; English commands allow specification of the actions the player would take as a character in a story (e.g., “fight the troll,” “open the red bottle with the corkscrew”), but no others.

The most-often cited microworld is the “blocks world” of Terry Winograd’s SHRDLU. It consists of a variety of colored geometric solids, with a software robot arm capable of manipulating them. The user can give the arm commands in English directing the manipulation, such as “PUT THE RED PYRAMID ON THE GREEN BLOCK.” It can resolve pronoun references (“FIND A BLOCK THAT IS TALLER THAN THE ONE YOU ARE HOLDING AND PUT IT INTO THE BOX”). It can accept definitions for new words standing for combinations of objects, and it can answer questions about the situation and its actions.

The power of SHRDLU, not to belittle Winograd’s feat of programming, is the result of the restrictions on its domain. Metaphors, emotional language, switches of context, and philosophizing are all excluded. Since everything happens on a display tube rather than with physical objects, unexpected changes in the environment (such as a block failing to balance or an extraneous object being introduced) are impossible.

The successes of AI, like the successes of any other kind of programming, have resulted from tailoring the approach to the problem. These successes, in fact, represent the height of the art of modeling. What is being modeled, though, is not the human mind, but the problem domain. The method of solving the problem mayor may not have anything to do with the way humans would approach it.

The requirements of modeling are requirements for AI designs. The designer must decide which parameters are relevant and which ones don’t matter, and what the overall goal of the model is. The broader the context which the model takes into account, the more complex the process of selection becomes. In the worst case, the number of possible interactions goes up as the factorial of the number of elements involved.

People avoid being combinatorially crushed by using what is vaguely called “common sense.” We focus only on the “reasonable” combinations of factors, eliminating the vast majority of possibilities from the outset. Imagine, for instance, that you are walking down on the street and you suddenly come upon a gaping pit spanning the sidewalk. In principle, many actions are possible. You could keep on walking; you could try to jump over the pit; you could carefully descend and come up on the other side. But you are not likely to stop and consider which of these alternatives is most reasonable; rather, you will probably stop for a moment and then go around it at a safe distance. You might also consider changing your course and looking for a policeman to talk to about the unmarked hazard. But in any case, your “common sense,” your ability to recognize actions as appropriate even if you have never encountered that precise situation before, greatly reduces the number of alternatives you have to consciously consider.

Putting the equivalent of common sense into a computer program is one of the most difficult tasks of AI. The problem is largely that common sense is not a set of explicit premises. If asked why you didn’t climb down into the pit, your immediate answer might be, “That’s ridiculous!” Given a little more thought, you could name a principle that would apply —the principle of not engaging in unnecessary effort—but you would not have consciously applied it in order to reject that course. Even naming the principle is not a programmatic directive for action (what is an unnecessary effort?), but only a summary and a guide. In cases

of doubt, the principle must be expanded into applications of specific experience. But if every principle is fully expanded to refer explicitly to all the specifics involved, the value of having principles is lost, and we are back in combinatorics.

A lack of common sense in an AI program can produce results ranging from the amusing to the disastrous, depending on the seriousness of the application and the type of error involved. James Hogan, in *The Two Faces of Tomorrow*, presents a fictional example that could plausibly occur in the future. A computer system on the moon is instructed to remove a ridge that stands in the way of construction in the most expedient way. The engineer in charge gives the task the highest priority. The computer asks, "ANY CONSTRAINTS?" and the engineer responds, "NO. JUST GET RID OF IT." There being no constraints, the computer invokes the mass-driver which it controls for sending payloads to Earth, and proceeds to bombard the ridge, nearly killing everyone in the vicinity.

Delimiting the broad context is one of the hardest parts of AI work. People start with that context and then develop specific areas of expertise; AI, for the most part, starts with specialized capabilities and tries to graft common sense onto them. This approach is reasonable, given the capabilities of today's computers, but it illustrates the fact that computer "intelligence" is quite a different phenomenon from human intelligence.

Frames and Scripts

The concept of context is brought into AI in the form of "frames." A frame can be defined as a set of data that identify special considerations to be given to a given object, act, or event. A frame for the concept "stocking," for instance, could be used to indicate that the verb "run" has a non-standard meaning when "stocking" is the subject that goes with it. Frames can be nested, so that there are exceptions to general rules, exceptions to exceptions, and so on. In considering the concept lion. the frames for lion, carnivore, mammal, and animal might all be active at the same time. This allows all the special considerations for carnivores, mammals, and animals to be applied to lions, or overridden where necessary, without the burden of having the frame for each distinct species of animal carry the information which is common to some or all of them.

Several distinct frames may be applicable in a given situation. For example, there might be a zoo frame and a jungle frame, either of which could be used with the lion frame according to the situation being described.

The description of a particular lion can fall into a frame whose structure and content is specified by the lion frame. This frame could specify the lion's sex, weight, age, and so on.

A frame can be used to determine how a lion might be expected to act in a given situation. For instance, if a program was required to determine whether a lion would eat a zebra, it could consult the lion frame (and perhaps also the carnivore frame) to determine what animals the lion eats; this class might itself be specified in terms of frames, rather than individual species.

Closely related to frames is the concept of script. A script is a sequence of primitives which describes a series of expectations in a stereotyped situation. The script for a lion attacking an animal might include the steps: (1) Lurk quietly. (2) Jump on the victim. (3) Break the victim's neck. Each of these steps must be expressed in primitive actions appropriate to the program.

Frames and scripts illustrate the traditional AI approach of building up a high-level conceptual structure as a means of dealing with context. The programmer must decide what types of facts are significant in order to decide how to structure its frame. The slots in a frame are not created by the program from large numbers of examples of the object being subsumed, but are generated as irreducible, atomic primaries. (In some cases, frame slots may be structures, but the primaries are still reached long before the level of discrete objects and events is reached.) A frame can be expanded when necessary (and languages used for AI are designed to make this expansion easy), but any added attributes are still basically atomic, not integrations of separate observations.

Top-Down Descriptions vs. Conceptual Knowledge

AI programs do not build up their microworlds, frames, and scripts from basic perceptual data, not even from a restricted set of perpetual data. Rather, they are described to the program, or built into it, by means of data structures. The approach is top-down, as opposed to the human bottom-up approach. A microworld, insofar as it corresponds to some part of reality, is a model. It could also be pure fantasy and not model anything. The creator of the model is a person or group of people.

The top-down approach does not constitute a close analogy to knowledge. Knowledge must begin on the perceptual level, with the observation of specific events and objects in reality. Children start by acquiring a store of perceptual information; only after that do they learn how to make factual statements and draw conclusions about what they see, hear, and feel. Some of these conclusions will be about events that they have never personally witnessed; but even there, it is their accumulated store of experience that allows them to distinguish between reliable and unreliable sources of secondhand information. (In many people this skill is very underdeveloped, to be sure, as the popularity of astrology columns and supermarket trashpapers shows, but what they stuff their heads with does not qualify as knowledge.)

AI programs are deficient in common sense to the extent that they deal only with factual conclusions, not with the data which lead to them. Every factual statement requires understanding of the concretes on which it is based. If a program represents facts only in relationship to other facts, it can present conclusions insofar as those facts are accurate, but it cannot fill in the gaps. Such a program does not possess, or even simulate in any deep sense, knowledge.

Consider a very simple example: a program capable of constructing syllogisms out of facts, and of answering queries in accordance with the syllogisms. Assume the following facts are in its “knowledge” base:

1. All cats eat mice. 2. Crystal is a cat.

A user could ask this program, “Does Crystal eat mice?” and the program would respond “Yes.” But how much understanding is really involved here? Perhaps Crystal has never encountered a mouse, and thus in fact does not eat mice. The program’s conclusion would then be wrong in a sense, even though its premises are correct and its logic flawless. The problem is an equivocation in its major premise: What it really means is that any cat, given the opportunity, would eat a mouse. Crystal then does “eat mice” in the sense that she, too, would not pass up such a snack; but in passing from the generic “All cats eat mice” to the specific

"Crystal eats mice" there is a subtle shift in implication. This shift can be understood only by an entity that has learned how statements of this kind are used in context.

This specific problem could be remedied by using different scripts to interpret statements about "all X" and statements about "this X." The first kind of statement may imply only a potentiality, while the second kind will almost always imply an actuality. But aside from the difficulty of making this rule precise, it only moves the difficulty to more subtle areas about what the statements actually mean. The gap can be completely closed only by tying the facts to a collection of perceptual data.

Belief in the information-processing model of the mind is, I believe, a factor which lets AI researchers believe that the high-level structures which they create are in fact a kind of simulation of the mind. If one believes that concepts are simply items of information (in the meaning-free sense), then pushing these items around without having to deal with their perceptual origins makes sense. But if we regard concepts as integrations of perceptual data, then severing them from their source is bound to make them wither into uselessness.

Mind and Metaphor

What gives AI its aura of mystery is the fact that it is full of metaphors drawn from human cognition. The words "intelligence," "learning," "understanding," and a host of others are applied anthropomorphically to computer programs. Sometimes these words are claimed to apply literally; in other cases they are recognized as metaphors, but considered very useful. Winston is a strong advocate of metaphors:

> Computer metaphors aid thinking. Work with computers has led to a rich new language for talking about how to do things and how to describe things. Metaphorical and analogical use of the concepts involved enables more powerful thinking about thinking.[3]

Metaphors can be great aids to thought, provided they are recognized as metaphors. Metaphors involving computers and the human mind can run in both directions. On the level of everyday usage (everyday for computer users, at least), we can speak of a word processor "knowing" what the margins of a page are; conversely, people speak of the "parameters" of a situation even where no mathematics or computers are involved. On a higher technical level, we can speak of the human mind using AI terms such as "static evaluators" and "search trees"; how well these terms fit is a matter to be determined by research.

Using metaphors of the mind is an old tradition in computers. The terms "memory" and "logic," as they apply to hardware, have completely distinct meanings of their own, but their origin as analogies to cognitive capabilities is obvious. The danger occurs only when people begin to believe they are literally true.

"Intelligence" itself is one of the most important of these metaphors. It has gained much of its power from being used in a computer-like, non-contextual way. Prior to 1950 or so, there were many tasks that could only be accomplished by intelligent beings; performing them was therefore properly regarded as evidence of intelligent behavior. Computer scientists wrote programs that accomplished some of these tasks, such as proving mathematical theorems. What this showed was that given the new technology, direct application of intelligence was no longer needed for some of these areas. What was claimed was that the programs had thereby

demonstrated intelligence. Such accomplishments were indeed impressive, but not in the way that McCorduck, for example, implies enthusiastically:

> The program was the Logic Theorist, which was able to prove theorems in Whitehead and Russell's *Principia Mathematica*. a feat of intelligence by anybody's standards .... The important point here is that these masked men had galloped out of the West with a virtual bandolier of silver bullets. They alone had managed to do what everyone at Dartmouth had faith was possible but had been unable to accomplish: they had made a machine that could think.[4]

Thus, as far back as 1956, we can find claims not only that computer programs will one day be able to think, but that they are thinking and displaying behavior which does not merely seem intelligent. but is intelligent. Such reasoning is sadly similar to the reasoning of a child seeing an automatic door for the first time and not revising his previous belief that only living things can move without being pushed, but instead deciding that the door must be alive.

"Learning" is another case in point. Programs are said to "learn" when they modify their behavior based upon the outcome of previous trials. As a simple example, one could write a tic-tac-toe program that generated legal moves, selected only by the criterion of never repeating a pattern that previous games have shown inevitably leads to a loss. After enough games, this program will have "learned" to play tic-tac-toe and never lose. But this is not learning in the human sense. The program has not started from actual experience and built up an integrated view of part of reality.

The way a person learns to play optimal tic-tac-toe is quite different; with the experience of some number of games, he discovers the importance of the center square and learns to watch for threats to complete a row or to create two such threats at once. Writing a program to play tic-tac-toe this way is not insuperably difficult, but writing a program to acquire these principles from experience would be much closer to real learning than simply building up lists of winning and losing moves. Doing the same for other activities—having the program generate rules on which to act, based on multiple outcomes—is a closer analogy to human learning, and correspondingly more difficult. Even here, though, it is an analogy, since it is working from one symbolic representation to another, without any fundamental grounding in experience.

Learning programs, in the more sophisticated sense, use such operations as generating hypothetical rules and testing them on the data. In outline form, this approach is similar to the scientific method of formulating and testing hypotheses.

Ascribing "goals" or "purposes" to a machine is another metaphor. Only living beings that goals or purposes, but it is convenient to think of programs as pursuing a goal. A chess program, for instance, will evaluate the results of different moves in the light of the short-term goal of obtaining the "strongest" position by certain criteria, and the long-term goal of checkmating the opponent. In fact, though, these are goals set by the programmer. The program simply goes through a set of operations according to the design of the program. Where the programmer does not set the purpose, the program cannot provide one.

Human goals ultimately arise from the alternative of life and death, and the consequent psychological alternative of happiness or suffering. Outside of a situation on which these alternatives exist, an entity can have no purpose, only a construction which will or will not lead to a certain outcome under given circumstances. Ayn Rand, in discussing the basis of ethics, gives an illustration of this point:

> To make this point fully clear, try to imagine an immortal, indestructible robot, an entity which moves and acts, but which cannot be affected by anything, which cannot be changed in any respect, which cannot be damaged, injured or destroyed. Such an entity would not be able to have any values; it would have nothing to gain or lose; it could not regard anything as for or against it, as serving or threatening its welfare, as fulfilling or frustrating its interests. It could have no interests and no goals.[5]

Moreover, the same considerations would apply to a real computer, which is very destructible, if the issue of its "life" or "death" is irrelevant to its actions.[6] A computer can lose power, it can be changed by having peripherals added or taken away, and it can be smashed to bits (pun intended). But these facts are irrelevant to a computer's programs, in most cases. (A robot that searched for power sources to plug itself into is a different case, and is arguably pursuing a real, though primitive, goal.) The program is written on the assumption that the computer will keep on running, at least long enough to produce some results.

Language Issues

Language provides a whole set of metaphors by itself. Computer scientists distinguish between "natural languages," which are the spoken and written languages that evolve through communication among people, and "artificial languages," which are created by explicit rules for a specific purpose. Artificial languages, which include programming languages, are generally much easier for a computer to handle, since their syntax is more clearly defined and their semantics much more restricted.

One of the great challenges of artificial intelligence is having computers "understand" natural language. The use of understanding here is metaphorical, since most of these programs are not concerned with relating words or sentences to their actual referents in reality. However, "understanding" of natural languages, with the quotation marks under-stood, is extremely valuable. Today people talk about the need to be "computer literate," but it would be much better if computers could be "human literate." That is, people should be able to issue requests to computers in ordinary English (or Japanese, French, etc.) and have the computer do what they expect.

Research on natural language processing has resulted in a view that syntax and semantics are strongly intertwined. Early attempts at handling English input started by parsing sentences, analyzing their structure according to textbook rules of grammar, and then dealing with their meaning in a separate stage of processing. This approach ran into trouble because of the large amount of ambiguity in individual words.

An often-quoted example is the sentence "Time flies like an arrow." This can be taken at least three different ways: as a statement that time passes quickly, as a command to measure the speed of insects which resemble an arrow, or as a statement about the preferences of a strange sort of insect called the "time fly." Most people would assume the first interpretation, unless the context indicated otherwise, but a program that dealt purely with syntax would have trouble distinguishing at least the first two interpretations. (The word "time" would not have to be listed as a possible adjective by a program that had special frames to accommodate phrases like "time bomb" and "time machine," so it might realize that *Drosophila tempi* is not a real insect.)

Or take the following two sentences:

1. I painted the house with a brush.
2. I painted the house with the brick chimney.

In the first case, the phrase "with a brush" is adverbial, modifying the verb "painted"; in the second case, "with the brick chimney" is adjectival and modifies the noun "house." The only way to make this distinction is to use the meanings of the words, to know that a brush is a tool for painting and a chimney is a part of a house.

When people hear or read statements in a language, they are constantly relating the words they receive to their experience, rapidly discarding one interpretation and choosing another as they obtain more information. The ability to do this is specific not only to one's experience with the events being discussed, but with the language itself. As a personal illustration of this, my ability to read French is fairly good, but I find myself largely bewildered when I hear spoken French. My background includes a large amount of reading nineteenth-century French novels, but very little French conversation. What is bewildering to me is that spoken French omits a great many of the consonants which are present in the written words, thus producing large amounts of ambiguity. For instance, the words oh, au (to the), eau (water), aux (plural of au), and eaux (plural of eau) all sound the same. To an experienced speaker, the intended word is generally obvious from context, but I must stop and decide which word was intended, meanwhile losing the next sentence.

The human method of disambiguation consists of expectation and of understanding the speaker's or writer's intentions. People will generally not be stopped by errors in typography or pronunciation; they may not even notice them. Even grammatical errors are more likely to annoy than to mislead, because people have an idea of what they expect to hear. The amount of correction people will apply itself depends on context; the statement "Mary doesn't want no eggs" might be taken by the same person at the store as indicating Mary's lack of desire for eggs, and in a technical psychological discussion as indicating Mary's aversion to a lack of eggs.

In science fiction, very strange contexts are possible, but experienced readers can take them in stride. Time-travel stories make bizarre uses of words; one of Heinlein's stories is called "Elsewhen." J. Neil Schulman's *The Rainbow Cadenza* postulates a future world in which light shows have developed into an art form with strong analogies to music; it contains phrases such as "the dazzling counterpoints in yellow lightning," which would normally be meaningless, but which are striking images in context.

The AI approach to language must be different from the human approach, since a program does not possess the referents of words, nor a flexible sense of the situation. Most approaches in AI have gone in one of two directions: either they have been general enough to handle almost anything imaginable, or they have adopted a large number of ad hoc devices. In either case, the approach is a top-down one, starting from some set of general principles or templates and making specific exceptions as necessary. This is in contrast with the human method of learning a language, which consists of learning large numbers of examples and developing general patterns and exceptions in accordance with them, with explicit rules being added only after the speaker has obtained a certain amount of mastery. (This applies most strongly to a child's acquisition of a first language. People learning a second language often take a more AI-like approach, learning definitions and rules of grammar before they have a broad background

of usage. This, perhaps, is why learning a new language can be so painful; that approach can become rationalistic and divorced from communication.)

The ultimate referents of natural language for a computer program are the actions it will take in response to a request, or the data items it will retrieve or modify. For instance, a program that can respond to queries such as, "How much money does Jones make per year?" does not relate the sentence to the person Jones, to Federal Reserve notes, or to a span of time; the referent is the appropriate entry in a data base, perhaps with calculations performed to bring it into the desired units. In this sense, as well as the more basic sense that the program is not conscious, it does not "understand" the query.

Human beings can also play the game of relating words to one another without understanding what they mean. From reading Lewis Carroll's "Jabberwocky," for instance, we "know" that some bandersnatches are frumious and should be shunned, and that a jabberwock can be killed with a vorpal sword. The reader can infer a great deal which might or might not be accurate. For instance, the typical reader will assume that the jabberwock is a ferocious monster, and that killing it required skill and courage. But what we actually are told is only that it has jaws that bite, claws that catch, and eyes of flame, and that it is manxome and burbles. The story could be one of killing a small lizard or bird, as told by an imaginative PR agent.

In a story like "Jabberwocky," people and computers are on a par as far as extracting facts are concerned. Computers may even have the edge since they will not make unwarranted assumptions. Do people "understand" the poem? Yes, but what we understand it as is a piece of fantastic word play that suggests images rather than presenting something concrete. (This is in contrast not only with stories about real beings, but also with stories about dragons and unicorns, which refer to previously understood mental images.) As a story about what happened, it permits us to draw only a bare skeleton of conclusions, compared with a similar story about a lion or wolf. When we restrict ourselves to this factual level, we can understand something of the computer level of "understanding" of language.

Strengths and Weaknesses of AI

The top-down approach which currently dominates AI is not a full emulation of human mental processes. Whether this is particularly bad or not is not obvious. If we take at face value the claims of AI researchers that their object is to model and study the human mind, then the current approach is deficient. It does not pay enough attention to the relationship of knowledge to its concrete referents. On the other hand, if the purpose of AI is to let computers deal with complex tasks so as to save people time, the current approach can be valid and useful.

There is a tendency in academic circles for people to shy away from admitting that their work has a practical purpose. Pure research, the acquisition of knowledge for its own sake, is something that professors often regard as the highest form of study. They regard pure science as the height of their Platonic heaven, applied science as tainted, and engineering as an unpleasant necessity, like getting one's hands greasy. This can lead even the people who do not think this way, but who would like to be admired by those who do, to regard their work as pure research without an immediate practical aim. Hence "artificial intelligence" sounds superior to "complex information processing," and pursuing a Philosopher's Stone sounds superior to making computers easy to use.

But how reasonable is this view? A human being is an integrated unit of mind and body, not a machine with a ghost in it. Regarding knowledge as pure because it is not put to practical use suggests the state of mind of the medieval ascetics who shunned all pleasures of the flesh so that they could purify their spirits. An achievement which is at once a theoretical advance and a means of achieving a useful goal is the best expression of the human spirit.

Identifying the real nature of AI, as opposed to its pretensions, does not degrade it but only demystifies it. Today, many people imagine future computers as having minds of their own, making the decisions that suit them, and perhaps leaving humanity at the starting gate. In fact, a computer that depends on humans for the structure of its "knowledge" and for the method it uses has no potential at all for such a threat.

The actual value of AI lies in two major areas, which we can call systems and applications. In the systems area, natural language processing and contextual treatment of data can make computers and other devices much easier for people to use. The Enterprise's computer in *Star Trek*, not HAL or Mike, best illustrates this potentiality; it accepts queries in formalized English and gives information, without making anyone learn specialized programming or database languages.

In the applications area, AI can offer the ability to deal with certain kinds of problems with less human effort. The more abstract a field is, and the less it depends on the particulars of experience, the more amenable it is to an AI approach. Two areas which obviously fit into this category are games and mathematics. An intermediate case is areas of science and engineering, where the required parameters can be provided in the form of a reductive model of an existing situation, or a constructive model of a proposed one. At the other end of the spectrum, and very difficult to deal with through current AI approaches, are subjects dealing with many concretes of human experience; these include ergonomics, psychology, and art. There can be uses for AI even in these areas, but human experience—or at least a significantly different type of technology which can provide its equivalent—must play a major part.

This spectrum is related to the spectrum between well-formed and ill-formed problems, but applies to only one aspect of well-formedness. A problem may be ill-formed either because the set of relevant data is open-ended and spreads through all of human experience, or because the set of data is well-defined but the criteria for what to do with them are not. The second type of problem may reduce to the first if the criteria themselves turn out to involve all of human experience, but I am thinking of a different type, one in which the elements are all present but difficult to sort out. An example of this is identification of musical instruments by their sonic wave-forms; the information necessary is all present and the solution unambiguous, but the means for finding the solution is less than obvious. This type of "ill-formed" problem is much more suited for computer solution than the kind in which the details of experience are more deeply and broadly involved—for instance, designing a musical instrument. We might call the two types of problems "closed-context" and "open-context" problems, recognizing that there is a spectrum of possibilities.

A further distinction should be made between areas in which the goals are fairly clear and those in which setting the goals depends heavily on experience. The latter area includes art, ethics, and law; it is doubtful that computers can ever make a substantive contribution to these areas, since they would have to deal in detail with what people experience simply to determine what to aim for. No matter how sophisticated a computer becomes, it does not have the mind, the values, and the needs of a human being. Even if a computer could be made conscious, it would have values and needs that pertained to its own species and not necessarily to ours. If it

does prove desirable to create machines that emulate human beings, current mainstream AI will provide only one part of what is needed. Turing's idea of raising the machine like a child will probably have to be revived. The technology of neural nets, which has recently been revived after it was abandoned in the early 1960's for lack of success, may provide something closer to the experience-oriented aspect of human cognition. Reports of recent results have been encouraging and suggest that neural nets may provide a more natural model of human thought than the high-level structures of traditional AI.[7] Such an approach may be more useful to the stated goal of AI of modeling human thought processes. Perhaps a combination of the two technologies will prove fruitful; neural nets might provide the lower-level perceptual and basic conceptual material, while the high-level AI approach allowed specific situations to be handled at a suitable level of abstraction. Such a combination of integrative and analytic capabilities might be much better suited for robots that must deal directly with their environment than anything that the current, essentially analytic, AI approach can produce.

---

1 Uhr, p. 10.
2 Winston, p. 1.
3 Ibid., p. 2.
4 McCorduck, p. 104.
5 Rand, *The Virtue of Selfishness*, p. 16.
6 1 am indebted to Dr. Harry Binswanger on this point.
7 A broad survey of neural net approaches is presented in Joregensen and Matheus.

XIV. Can Minds Be Modeled?

Philosopher Hubert Dreyfus has provoked a stormy reaction from the artificial intelligence community with his *What Computers Can't Do* and, co-authored with his brother Stuart, *Mind over Machine*. His thesis is that artificial intelligence is by its nature doomed to failure, because it proceeds from a faulty model of the human mind. The issue here is not whether artificial intelligence can make computers think, but whether it can give them the ability to respond to situations in a way that matches human expertise. The response to his claims has often been fiery.

In a book which is at least as heretical as Dreyfus's, it's worth pausing to examine his claims from a perspective which is different both from his own and from the usual AI viewpoint. This will help to shed further light on the question of modeling the human mind, even if Dreyfus's conclusions are not entirely correct. At the same time, it should further clarify my own position, by contrasting it with one which has some similarities to and some major differences from my own.

Artificial intelligence, Dreyfus argues, attempts to reduce the operations of the mind to a set of rules, or at least to describe it by means of such rules. However, he states, there is no compelling evidence that there are such rules, and there is good evidence that no such rules can be found.

> Computers are certainly more precise and more predictable than we, but precision and predictability are not what human intelligence is about. Human beings have other strengths, and here we do not mean just the shifting moods and subtle empathy usually ceded to humanity by even the most hard-line technologists. Human emotional life remains unique, to be sure, but what is more important is our ability to recognize, to synthesize, to intuit. There are good reasons to believe that those abilities as well are rooted in processes altogether different from the calculative reason of computer programs, and we shall explain, as best we can, what those processes are.[1]

Dreyfus does not deny the validity of human rationality, but he grants it a limited scope.

> Although irrational behavior—that is, behavior contrary to logic or reason—should generally be avoided, it does not follow that behaving rationally should be regarded as the ultimate goal. A vast area exists between irrational and rational that might be called *arational*. The word rational, deriving from the Latin word *ratio*, meaning to reckon or calculate, has come to be equivalent to calculative thought and so carries with it the connotation of 'combining component parts to obtain a whole'; arational behavior, then refers to action without conscious analytic decomposition and recombination. *Competent performance is rational; proficiency is transitional; experts act arationally.*[2]

He further distinguishes between "calculative rationality," which "produces regression to the skill of the novice or, at best, the competent performer," and "deliberative rationality," which "does not seek to analyze the situation into context-free elements but seeks to test and improve whole intuitions."[3] Deliberative rationality is the more useful, but it still serves as a supplement to intuition, helping to deal with anomalous situations and avoid "tunnel vision" in thinking.

Two questions have to be considered regarding Dreyfus's view. One is the question of whether it is true of people; is human thought, in fact, primary holistic and intuitive rather than rational? The second is the question of its implication for modeling such processes; to the extent that human thought is intuitive, that it consists of "know-how" rather than "knowing that," are rule-based models incapable of giving a good approximation of the way it works?

These two questions are partially independent. Many phenomena of the human mind do not depend on conscious reasoning; recognizing a face and riding a bicycle are just two examples Dreyfus gives. Whether this type of immediate recognition is not explainable through subconscious rules, and whether it forms the basis of human expertise, is a question to be resolved independently of its implications for modeling. At the same time, the question of whether this kind of behavior is beyond the reach of a rule-based representation remains open even if people do not actually use such rules, and the question of whether "arational" thought can be modeled by computer procedures is one that must be answered on its own terms aside from the extent to which people use it.

In any issue of this type, it is necessary to define what we are talking about. (Dreyfus might disagree here; he presents an unsympathetic view of Socrates' quest to find the essential characteristics of abstractions that people commonly use.) In particular, what is rationality? In discussing Dreyfus, it seems necessary to discuss the two types which he mentions separately. Calculative rationality he identifies with "reckoning" and by implication with symbol manipulation. Clearly, this is something which is relatively well suited to computer modeling. The model consists of a set of premises and a set of inference rules; the only difficulty (though a major one) is the choice of which rule is to be applied to which premises at a given time. Deliberative rationality is a concept which he does not present as clearly, though it apparently is closer in meaning to what I would call rationality in humans. It is "not opposed to intuition but based on it."[4]

In presenting the concept of deliberative rationality, Dreyfus almost succeeds in overcoming the traditional philosophical barrier between the analytic and the synthetic, between knowledge obtained by induction and knowledge obtained by deduction. But the category of calculative rationality as something separate still leaves parts of the wall standing.

The basis of all knowledge is experience. Rationality is the consistent practice of dealing with reality according to what it is, and the only way we know what reality is is by observing it. People observe large quantities of separate phenomena and learn to deal with them through concepts. This process of integration is, very roughly, what Dreyfus refers to as "intuition." But reason apart from perceptual data is impossible; "calculative reason" is not reason at all.

Whenever people acquire knowledge, they must use both "intuition" (integration) and deduction in a mutually dependent, mutually reinforcing way. Rules do not exist in reality; they are formulations by human beings of regularities in reality. The identification of these regularities is a process of observing facts, isolating commonalities among them, and arriving at a statement which expresses the commonalities. Such a process is necessary because people cannot simultaneously retain all the separate facts in their awareness. The broadest statements obtained by this method include the rules of logic, which can be used to relate one fact to another and thus combine statements into arguments.

These statements are brief expressions of a large amount of discovery and cannot be treated in isolation without losing their significance. They have to be re-checked against reality at every opportunity to make sure that no new information has arisen which might

invalidate them. Any conclusions drawn from them must also be checked against reality where possible. This is not because the principles of inference are suspect, but because any generalization depends upon the particulars from which it is obtained.

This process can be seen in the way children learn. Suppose, for example, that a child grows up among white people only. He will conclude, if he is given no other information, that light-colored skin is a human characteristic. Within the limits of his available knowledge, this is a valid conclusion; but if he sees a black man, he must revise his generalizations to take the new discovery into account. It would be foolish to assume that the previous generalizations must be right and this new entity can't be a man.

On our own level of knowledge, we can properly say that there are no people with truly red skin. No such tribe of people has ever been discovered (and clichés aside, Indians do not have red skin). But if explorers discovered such people on an island and they met the basic criteria of being human, then we would have to discard the general principle that no one has red skin.

Someone could object to that principle even today on the grounds that there are people with red skin, namely people who have been badly sunburned. But in dealing with this objection, we come across another aspect of general statements that prevents their being considered in isolation: every statement has an implied context. In this case, the implied context is one of natural pigmentation, as opposed to coloring caused by bums, disease, embarrassment, paint, etc. Specifying the full context for every statement would be impossible, since each statement about context itself has a context; what is necessary is an understanding of the purpose of a statement. If I specified "natural pigmentation," a determined heckler could still point out that people have naturally red skin around their mouths or that under a certain kind of light, normal people will look red.

The basis for many kinds of nitpicking is the refusal to consider context, or the attempt to consider statements as having to be true independently of any context. This approach is typical of the low-grade kind of academic argument that asserts counter-examples to a principle by discarding its context. For example, an MIT student once told me a "square circle" is not a self-contradiction, because one could construct a system of geometry in which distance is defined to be the maximum of the x-difference and the y-difference between two points. In such a system, a square would indeed be a circle, i.e., the locus of all points equidistant from the center; but the original discussion had made no reference to contrived geometries.

Dreyfus's "calculative rationality" represents applying generalizations without regard to context. There are cases in which this approach is safe, because the context is unambiguous. In literal calculation, it is normally safe to say "5+5 =10"; though even there, a malicious arithmetic teacher could trip up a child by pointing out that the answer could be 12 or 14 if the radix isn't ten.

The beginner in a field of study has a very sparse context on which to draw, yet may have generalizations offered to him by teachers or writers with very broad experience. It can be tempting to resort to "calculative rationality" in this case and just plug in the rules to get answers, but relying solely on this approach is a mistake even for the novice. The amount of context he can relate the rules to is limited, as is his skill for relating the material; but he reduces himself to the robotic level (in the most simplistic sense) if he makes no attempt to integrate it with the rules.

For instance, a basic rule in learning to ride a bicycle is "Turn in the direction you're falling." A competent cyclist does this without thinking about it. But even for a beginner, the rule is not context-free; obvious questions are how far to turn and when to stop turning. The answers to these questions can be obtained only by integrating a large amount of potentially painful experience.

Not all integrations lead to formulating explicit principles. When we are dealing with physical skills, such as riding a bicycle or playing the piano, this point is obvious; a performer playing a Liszt concerto could not possibly think of all the principles involved in the time available to strike the notes. However, it is also true in mental skills. For example, in writing the previous statement, I wrote "riding a bicycle and playing the piano" with only a brief hesitation over the lack of parallelism. I did not have an explicit rule which says that vehicles take an indefinite article and musical instruments take a definite article in this situation, but experience in my mind was sufficient to make the correct choice of article, and then to think about the generalization. If I had not juxtaposed the two different usages, I would most likely not have thought of the issue at all but would still have used the right words.

Does this imply, as Dreyfus would have it, that the choice of the correct words is an "arational" process, although it might be supported by "deliberative rationality" if I were in genuine doubt about which word to use? Not quite. It is a process which involves only one aspect of the full human capacity for reason, in this case the capacity to integrate and associate. In arriving at the correct usage, I called without conscious effort on usages I have heard and read before. If I had written, instead of "playing the piano," "playing the basset horn," the process would have been similar but involved a greater leap of association. I have not heard that exact phrase before, but have heard similar phrases with other musical instruments and thus can extend the usage. If I wish to be sure of my conclusion, I must complete the process and identify the principle which I have implicitly used, though normally I do not have to do this.

But is there an implicit principle in all such cases? The answer is yes, remembering that principles are not an element of reality, but a cognitive necessity. Associating and collecting facts does not by itself provide validated knowledge. Validation requires identifying the features by which similar entities are related and establishing what may be properly concluded from those relationships. We do this by formulating of a statement or a concept which names the commonality.

The name "rationality" does not properly belong to any one aspect of the cognitive process. Reason, or rationality, is the process by which man acquires and validates knowledge, based upon the information provided by the senses. To reach a deductive conclusion that flies in the face of the facts is irrational, even if the premises seem entirely justified. Conversely, a process which does not make use of explicit facts can still be a rational process of understanding, if it begins from what is known (e.g., observations and experiences) and arrives at an understanding without the fabrication of information or suppression of knowledge. Thus, a primitive may look up at the sky, observe certain signs that have signified rain in the past, and expect that it will rain without verbally identifying the relationship between the signs and the result; he is nonetheless being rational in this process. He may also conclude that the weather god is unhappy and about to cry; this would be an irrational conclusion (though understandable in cultural terms), since reaching it involves the invention of an entity without supporting evidence. We have the advantage of being able to name the factors in a weather prediction, and thus to formulate and test hypotheses; but the act of predicting the weather based on prior experience is an equally rational act in either case.

"Calculative rationality," meaning non-contextual deduction from premises, is not a valid method of thinking except in special cases. The archetypal situation where it can be used is schoolbook problems. The factors are all spelled out, and no contextual issues can intervene. More generally, calculative rationality can be applied to models, but not to reality itself. Within a model, all the relevant factors have been specified in a predetermined way, giving an accurate representation in a limited set of cases. Hence, calculative rationality (working from rules alone) is appropriate to operations on models, but deliberative rationality (i.e., full rationality) is necessary when understanding reality.

The principal error which Dreyfus makes concerning human cognition is his failure to recognize the importance of identifying and validating the results of "intuitive" integration whenever possible. Responses based on habit and experience are important as indicators of what to consider and are necessary whenever factors of time and complexity exceed what the conscious mind can handle, but they must be validated by specific, conscious identification.

Dreyfus disparages the conscious portion of the process, saying, "Doctors are tempted to rationalize their intuitive decisions not only to justify them to themselves and their peers, but also in order to explain them to their patients ... the doctor can never factually explain his innermost feelings about the preferred therapy based on a lifetime of experiences with similar cases." In making a decision on whether or not to perform surgery, for example, the doctor must integrate all his relevant knowledge, and he cannot be expected to explain every detail of this integration, especially to someone without equivalent experience. (Besides, no two people ever have exactly the same experience.) But in the process of reaching it, he must make as much of his understanding as possible explicit to himself; otherwise, he has no way of distinguishing a legitimate intuition that surgery is required from a desire to be fashionable or an urge to hurt the patient. (The issue here is factors which are not conscious, not deliberately unethical behavior.) Purely intuitive methods would be fine, if people weren't fallible.

Dreyfus's characterization of rationality is in error, but there is an important truth in what he says: that rationality, in the human sense, is not simply the application of rules to unambiguous facts. Every fact has its context, and an act of reasoning must be an act not just of applying rules, but of bringing in as much of the context as possible.

Does this mean that human thought is largely holistic, not susceptible to analysis and modeling? Dreyfus argues that it is, and in doing so draws a distinction between a mechanistic system and a holistic system.

> A machine [or mechanistic system] does its work by dividing up the job among different components each with its different function and putting them all together so as to produce a result. ... But there are devices which do what they do in an entirely different way. A holographic pattern recognizer does its work without dividing up the job between different components. True, at the second [component] level one can describe the lenses, lasers, and so on, and one can explain what job each one does, but something essential is left out. The actual recognition work is accomplished by the interference of two beams of light, with no separate functional components doing any work.[5]

Not being a biologist, I prefer not to take sides in this issue. Certainly science has shown that certain parts of the brain are associated with certain functions; if, for example, the speech center of the brain is damaged, the victim's ability to speak is impaired or lost while his reading vocabulary may remain intact. On the other hand, we experience ourselves as a whole,

not as a collection of specialists working together inside our bodies. This suggests that on some level, the special functions of the brain merge into a single capacity.

In any event, Dreyfus's distinction is not essential to the issue of whether computers can think, nor to the issue of how well they can model thinking. The first issue I have already addressed without reference to the brain's specific method of functioning; the second issue is one of whether thought can be analyzed in terms susceptible to computer modeling, regardless of how the neurons of the brain interact. Since this book is not about methods of programming artificial intelligence, I will consider the latter question only from the standpoint of whether any general principle indicates that such an undertaking is possible or impossible.

In this type of question, the burden of proof is on the assertion of the positive. In this case, the positive is the claim that thought can be modeled on a computer to a given level of performance. This is something which is largely taken for granted; but as Dreyfus points out, it is not self-evident until it has been accomplished. Presumably there is some method by which all the activities possible to the mind can be represented, but there are physical limitations on the information-processing capability of any digital system; Dreyfus cites a limit of $2 * 10^{47}$ bits per gram. This is rather like pointing out that no automobile can ever go faster than light, but unless we find a way around the Heisenberg uncertainty principle, such a limitation must certainly exist. The issue of whether the physical limitations are significant is ultimately one to be resolved by research, testing, and the willingness to try new ideas and discard invalid ones. Predeciding the issue either way will do no good.

Dreyfus's criticisms are largely valid insofar as they address the current, mainstream approach to artificial intelligence. The construction of systems of facts which are not built up from experience in reality, but are the fundamental data from which the program works, fits his concept of "calculative rationality." This approach is, as I have already stated, well suited for situations that can be modeled in a reasonable way, and not very well suited for situations that depend intimately on experience. However, Dreyfus apparently regards the AI approach as the only means by which computers could be made to perform tasks that currently require thought. The raw data of reality are not inaccessible to computers, but two elements are missing today: the technology to handle the quantities of information involved, and the understanding on the part of AI researchers that these data are the necessary base of knowledge. The first is something which will undoubtedly be achieved given enough time, but the second requires recognizing the fallacy of regarding thought as a species of information processing.

A great deal can be accomplished using the AI approach; but its lack of grounding in observation of reality means that its method must often be significantly different from what people do. Consider, as an example, the problem of how to perform a medical diagnosis. A doctor must take many factors into account, including symptoms, the patient's general condition, his past history, the season, current prevalence of diseases, and so on. If he cannot reach a sufficiently certain conclusion, a conscientious doctor will perform additional tests. Part of this process involves applying consciously recognized principles (for instance, that bacteria of a certain shape under a microscope indicate a certain disease); other parts involve subconscious integration (for example, recognizing that certain symptoms are anomalous even though they do not indicate a specific disease).

A computer program for providing a recommended diagnosis might be most effective if it followed an entirely different approach. It might, for instance, be built around a detailed model of the physiological workings of the human body, and relate every symptom to the kind of

organic malfunction that could cause it. The amount of information necessary to do this might be too large for a human doctor to handle, yet be appropriate for computer processing. Thus, the best approach would be one which does not emulate human thought, but uses a different method appropriate to the computer's strengths.

This is only an example; I have no real medical knowledge and no good idea of how to design a program to aid in medical diagnosis. My point is simply that emulating the human approach is not necessarily the best way; a different approach may even have the advantage of complementing human capabilities rather than repeating human errors.

Hardware Suitability

Dreyfus's specific criticisms of the suitability of computers for representing human thought are sometimes on target, sometimes in error. Some of his comments suggest a different architecture as being more suitable than the Von Neumann model (i.e., sequential instruction execution with linearly addressed memory). He notes that people can recall facts by association, as it were, without any obvious need to search long lists. For example, if you are asked to recall your date of birth, you can do so immediately. This suggests that the particular date is somehow "keyed" to the concept "date of birth," so that it can be recalled without effort. Other facts may take longer to recall, but the realization that one does know the fact is still apt to come immediately, and the fact itself after some concentration. Even in this case, the process usually isn't one of consciously examining a series of facts until the thinker comes upon the right one; rather, it is a matter of concentrating on the association, of somehow getting the fact to pop into one's mind.

This method of recollection suggests that some form of associative memory needs to be used in the model. The usual kind of associative memory used in computers is not flexible enough for this purpose; it associates a key with a datum, and the key is usually relatively short and must be matched exactly or by simple masking operations. Human recollection of a fact can be triggered by a key of the right general kind (for instance, "birthday" or "day I was born" can substitute for "date of birth"), not just by a specific pattern. The right kind of hardware for emulating this human capacity may be quite different from anything used in computers today.

Related to this point is Dreyfus's comment that an important aspect of human thought is people's ability to recognize similarities. In computers, pattern recognition is a laborious process; yet people manage to do it with very little effort. This does not preclude the creation of a machine that can perform the same task with human effectiveness, but it suggests that a different kind of hardware may be more appropriate. Some type of analog process might be more suitable to comparing global features than a digital pattern recognition program.

The Assumptions of AI

These issues can be taken as simply a question of how one can best implement a model of the human mind. But to get to the root of Dreyfus's criticism of AI, we must consider the assumptions which he states are built into it, along with his criticisms of them. He identifies these assumptions as follows:

1. A biological assumption that on some level of operation—usually supposed to be that of the neurons—the brain processes information in discrete operations by way of some biological equivalent of on/off switches.

2. A psychological assumption that the mind can be viewed as a device operating on bits of information according to formal rules. Thus, in psychology, the computer serves as a model of the mind as conceived of by empiricists such as Hume (with the bits as atomic impressions) and idealists such as Kant (with the program providing the rules) ...

3. An epistemological assumption that all knowledge can be formalized, that is, that whatever can be understood can be expressed in terms of logical relations, more exactly in terms of Boolean functions, the logical calculus which governs the way the bits are related according to rules.

4. Finally, since all information fed into digital computers must be in bits, the computer model of the mind presupposes that all relevant information about the world, everything essential to the production of intelligent behavior, must in principle be analyzable as a set of situation-free determinate elements. This is the ontological assumption that what there is, is a set of facts each logically independent of all the others.[6]

"In each case," Dreyfus tells us, "we shall see that the assumption is taken by workers in CS [computer science] or AI as an axiom, guaranteeing results, whereas it is, in fact, only one possible hypothesis among others, to be tested by the success of such work."

In the first of these cases, Dreyfus is overgeneralizing. The biological assumption is in fact not necessary to simulating cognition on a computer; he does recognize that it is the most easily discarded of the assumptions, and treats it only briefly.

The psychological assumption is, in fact, very popular in AI, where it is expressed in statements to the effect that any idiot knows that the brain is a machine. This is, in fact, the key error behind the efforts to develop thinking computers. On the other hand, this assumption is not necessary to efforts to simulate human thought, in the broad sense of letting computers deal with complex and ill-formed problems. This distinction needs to be made clear, since it is the difference between futile searches for the philosopher's stone and legitimate research.

The epistemological assumption, on the other hand, is fundamental to any attempt to simulate thought. More precisely, to the extent that it is false, i.e., to the extent that there is knowledge which cannot be formalized, there is no way that a computer can deal with it. But formalization of knowledge can be achieved in many ways besides the obvious one of listing statements about the world. AI in its mainstream form relies on a specific form of the psychological assumption: that a high-level formalization of factual statements, which does not tie them to their referents in reality, is sufficient for formalizing enough of human thought to qualify as "intelligence." An approach that began with the association of input from sensory devices with internal data structures, or that at least introduced a detailed correspondence between the objects in computer storage and their referents in reality, would model a much larger portion of human thought, though at a much larger cost in processing and storage.

The idea that listing a relatively small (less than a million) number of declarative statements, along with rules for producing new statements, is sufficient to simulate thought, is a particularly crude form of the epistemological assumption. This idea appears, for example, to

be the basis of Turing's remarkably low fIgure of 10^9 bits as sufficient for a program that would successfully play the imitation game, a figure which he notes is about the size of the Encyclopaedia Britannica. But alternative approaches are possible, and it is the task of researchers to discover them. Dreyfus rejects rules across the board, regardless of the approach taken.

His discussion of the ontological assumption is puzzling. His basic premise is correct: that no fact can be treated in isolation, that it has a context that gives it significance. For instance, the fact that a man is wearing a shirt and pants has different significance depending on whether the weather is warm or dangerously cold for such clothing, and has different significance depending on whether the observer's purpose is to identify someone by his clothes, to guess the man's degree of wealth, to look for concealed weapons, or to sell him new clothes. A statement like "the box was in the pen" is not referring to a ball-point pen, but we understand this fact only because of our contextual knowledge about the usual nature of boxes, writing implements, and enclosures.

Dreyfus is aware that a major portion of AI research is devoted to solving the problem of context (or, as it is often called, the frame problem). But his point is that there is a hierarchy of contexts to be considered, and that the ultimate context depends on one's situation as a human being. An infinity of contexts would be an infinite regress, but humans do not have a static, ultimate context. Rather,

> Human beings seem to embody a third possibility which would offer a way out of this dilemma. Instead of a hierarchy of contexts, the present situation is recognized as a continuation or modification of the previous one. Thus we carry over from the immediate past a set of anticipations based on what was relevant and important a moment ago.[7]

This, he claims, is something computers cannot do, since the human context ultimately arises out of one's development from infancy. Even if a program were brought up from "childhood," as Turing proposed, it would not have the same kind of physical experiences a human being would have.

Now this point is valid, if the question is whether what we want to do is program a robotic human being. The context of a robot is not that of a human being. Weizenbaum makes a similar point in asking rhetorically.

> What could be more obvious than the fact that, whatever intelligence a computer can muster, however it may be acquired, it must always and necessarily be absolutely alien to any and all authentic human concerns?[8]

But what is the proper conclusion to draw from this? It is that purpose must always be defined by human beings. It is in this sense, and not in any of the more easily dismissed senses, that computers can do "only what we tell them to do." A robot must be given its purpose and its constraints. If people did design a machine that fully emulated a human being, it would act for its own purposes, and not for a human purpose unless the two coincided. The contradiction entailed in slavery applies to mechanical slaves as well; one cannot have a being that thinks for itself, yet is totally dedicated to another being's purposes.

But this conclusion denies something rather different from Dreyfus's original statement of the ontological assumption. The problem is not that there is such a thing as context, but that a machine cannot have a fully human context. Suppose we could transfer the whole contents of a human mind into a machine; Hofstadter presents a humorous discussion of this possibility in

"A Conversation with Einstein's Brain" in *The Mind's I*. Even if this were done, from the moment of transplanting, the contents would reside in a machine that had its own physical capabilities and requirements, not those of the person who provided the "program." Hofstadter's Einstein-machine (actually, an Einstein-book) would not simply keep acting as if it were Einstein in his own body; Einstein was not an ivory-tower generator of ideas, but a human being.

Conclusions

While Dreyfus does make some worthwhile points, he is to a large extent pursuing the wrong quarry. In his phenomenological view of the world, intuition without conscious, explicit identification is the major portion of cognition, and cannot necessarily be reduced to computational procedures. But for having computers perform the useful tasks which currently require human cognition, it is not necessary to precisely duplicate the human method; it is only necessary to identify the relationships between the given conditions and the goal, and to create a procedure for bridging the gap between the two. This method may be entirely different from the human method. To say that no such method is possible would imply that the relationship between the two cannot be expressed in objective terms and is entirely a matter of the peculiar way in which the human mind operates.

There are, perhaps, some areas where this is true. To create a good poem or piece of music (one which is genuinely expressive rather than merely correct) is something that depends on the peculiarities of human nature, and a computer would have to emulate a very large part of the human mind (which includes human experience) to achieve the desired result. But for problem-solving areas in which the context is sufficiently closed (e.g., discover what is wrong with the patient, identify the most probable location of iron deposits, etc.), there is no reason why a procedure for achieving the goal shouldn't be possible, whether the human approach is in any sense procedural or not.

The more important issue, which Dreyfus touches on but does not state as such, is that computers are value-neutral devices; they have no goals or interests of their own, and hence must be provided with a goal in explicit terms. This is partially indicated in his discussion of human contexts, but the issue of values is not separated clearly. The nuances of human values, including esthetic, moral, and idiosyncratic ones, are very subtle, and a computer does not have access to the basic information of everyday experience which gives rise to them; hence, programming anything that is dependent on human values in a complex way is a dubious enterprise. (Any computational activity depends on human values, but if they are expressed in a simple way, such as predicting the weather next week, there is no particular problem.)

The contradiction which is inherent in the advocacy of slavery also applies, rather surprisingly, to the desire for an intelligent machine that will fulfill human desires; an entity can't think for itself, but at the same time be totally devoted to someone else's goals. A machine that did think for itself would not be a machine, but an entity that would act according to what it regarded as important; such an entity would lose many of the advantages of a non-thinking machine.

Dreyfus's distinction between brains and machines on the basis of holistic systems vs. systems with specialized parts is also not particularly significant. If a machine is a system which has parts that perform distinct functions, then it can still incorporate holistic elements. Nor would the identification of separate functional components of the brain make a difference

to our psychological understanding of human cognition. The real difference between a human being and a machine is, first, that a human being is conscious, and second, that a human has his own values and purposes which are not specified for him by an external, valuing entity. (people's values are certainly affected by other people, but no one can definitively set another person's goals, preferences, and objects of admiration.) This difference places a limit on what computers can do for people.

Joseph Weizenbaum has made the point much more effectively than Dreyfus: "But the difference between a mechanical act and an authentically human one is that the latter terminates at a node whose decisive parameter is not ' Because you told me to,' but 'Because I chose to.'"[9] This is the key to all the differences between what people can do and what machines can—and should—be made to do.

---

1 Dreyfus and Dreyfus, p. xiv.
2 Ibid., p. 36.
3 Ibid., p. 36.
4 Ibid., p. 205.
5 Ibid., p. 63.
6 Dreyfus, p. 156.
7 Ibid., p. 222.
8 Weizenbaum, p. 226.
9 Ibid., p. 260.

XV. Science and Mysticism

A basic principle of magic is that "the name is the thing." In the classical theory of magic, getting results depends not on understanding what things are, but on knowing what to call them. Science, in contrast, holds that the thing is itself, regardless of what anyone says about it. What we call things must be based on what they are if we are to understand them and thereby be able to use and control them.

The magic-oriented view of the world is far from dead. The idea that labels can change reality is strong in many people's minds, most obviously in politics. Since ancient times, there has been a conflict between two ways of dealing with reality: the method of reason and the method of mysticism. Where reason has held the upper hand, the result has been understanding, good will, and prosperity. Where mysticism has prevailed, the result has been ignorance, mutual hatred, and suffering.

Reason is the approach that takes reality as the given, which has to be understood according to its nature. The rational mind begins by asking: what is there? Only after it does this can it create new ideas and plan new arrangements of the elements of reality.

Mysticism is the approach that takes the mind as the given and reality as its product. Historically, three types of mysticism have been important. One, which when held consistently leads to solipsism, holds that one's own mind is the creator of reality and that all acts of understanding are really acts of creation. A second holds that there is a super-mind in the universe that creates all of reality, and that the key to understanding the universe is understanding the workings of this mind. (A belief that a God created the universe, but that it is understandable on natural terms, without invoking divine intervention, is not mysticism.) The third kind of mysticism holds that existence is a collective product of all our minds, and that discovering reality means discovering how people at large think.

Rand has characterized these opposing views in another set of terms: as the primacy of existence vs. the primacy of consciousness. Primacy means "what comes first," in a metaphysical or causal sense: does existence exist before we are aware of it, or is our awareness the means by which existence comes into being?

Only one form of the primacy of consciousness is logically consistent, and that one, solipsism, is vacuous. All other forms depend on grasping some fact of reality outside oneself, be it a holy book, a voice from the sky, a newspaper, or other people's bodies. The believer regards his understanding of these things as facts, not as manipulations of some mind, by which he proves that everything is the product of the social or divine mind. But if he were really consistent, he couldn't count on any of these things as proofs, since true reality isn't in what he sees but in the controlling mind.

The solipsist avoids this problem, since he holds that he is the controlling mind; but he has reduced reality to a figment of his own imagination. Once he has reached this conclusion, there isn't any point in his continuing to talk, work, eat, or do anything else, except as self-entertainment; he can rewrite reality to be whatever he wants it to be. There is, however, one saving virtue to solipsism: few people try to win converts to it.

All of this would be of little interest if mysticism were confined to the places where we usually think of it: of Hare Krishnas singing near subway stations, ascetics starving themselves or drug freaks poisoning their brains in order to have visions. These are isolated, relatively harmless phenomena; but there has been a more serious resurgence of mysticism in the twentieth century, a mysticism allegedly based on science.

This type of mysticism is often apparently at odds with older varieties. However, the common link is the principle of magic mentioned earlier: the idea that the way something is described or referred to affects what it is in reality. The pseudo-scientific version of the principle that "the name is the thing" is: "The model is the thing."

This is not a solipsistic kind of mysticism; no scientist even marginally worthy of the name would think that what he personally believes or wishes will change the facts. It is not religious in nature, asserting the existence of an overt super-mind that affects reality, and not overtly social in the sense of holding that people's minds create reality (although a number of popular writers have cashed in on quantum mechanics to support just such a claim). Rather, it is a mysticism of methodology, the belief that reality exists not in itself, but in the method by which it is understood. Not any method will do; the method must be "scientific." But the mysticism lies in the rejection of metaphysics—which means rejection of reality—and in the equation of that which is beyond the boundaries of a designated box with the unreal.

This mysticism is implicit in all forms of the black box fallacy. When someone regards the reality of a system as being equivalent to the mode in which he understands it, he is acting on the view that his understanding, or the understanding of the person who provided him with information, sets the terms of reality, that without observation, an act of consciousness, there is no determinate reality.

In particular, this mysticism is found in the metaphysical Copenhagen interpretation of quantum physics, which holds that because our usual modes of measurement are found to break down to statistical approximations at a certain level, reality is to that extent indeterminate, nothing in particular. It is brought fully into the open by those who hold that Schrödinger's cat is neither dead nor alive because its state cannot be predicted or observed. Human awareness, and specifically human methodology, is taken as the creator of reality, the giver of life and death.

The same mysticism is found in Turing's equivocations on thinking machines. While he does not say that an observer's inability to tell a machine from a human is what makes it think, he makes it clear that he wants people to judge whether a machine thinks solely on the basis of its observed behavior in a blind experiment. The consequence is animism: the attribution of a mind to an inanimate object.

It is found in all theories of psychology and cognition that hold that the mind may not be considered because it cannot be laid out for all to observe. Its most extreme form is behaviorist psychology, which regards people, pigeons, and mice as similar boxes that respond in certain ways to certain stimuli, refusing on principle to explain the reasons for the response in terms of awareness, knowledge, or understanding. Quoting Nathaniel Branden:

> For centuries, *mystics* have asserted that the phenomena of consciousness are outside the reach of reason and science. The 'scientific' apostles of the anti-mind agree. While proclaiming themselves exponents of reason and enemies of supernaturalism, they announce, in effect, that only insentient matter is 'natural' —

> and thereby surrender man's consciousness to mysticism. They have conceded to the mystics a victory which the mystics could not have won on their own.[1]

It is found in the view that logic and mathematics are simply symbol-manipulation, that they do not describe relationships in reality but are simply arbitrary constructions that reach their greatest purity in formal systems that are explicitly dissociated from reality. Often this attitude results from strenuous efforts to avoid the black box fallacy by completely divorcing mathematical descriptions from the entities and processes which they fail to describe perfectly. This treatment promotes the idea that mathematical descriptions in science can't be expected to describe reality, but only to describe a set of terms in which we characterize it. The result is an other-worldly view of science, based not on a supernatural world, but on an idealized, mathematical world which has no necessary connection to reality.

In Sowa's *Conceptual Structures*, we find a fairly typical example of the moderate form of this view:

> Since finite, discrete concepts can never form a perfect model of continuous reality, a truly precise, objective science is not possible even as an ideal. The truth of any model must be limited to those few aspects of the world that the designer of the model chose to represent. ... According to the measuring instruments available five millennia ago, the ancient myths corresponded quite well with reality. Since they covered all aspects of all human concerns, one might even say that they were more true for their society than science is for ours.

Hofstadter's *Gödel, Escher, Bach* presents this mysticism at an advanced stage, and with explicit ties to Zen. Chapter IX, "Mumon and Gödel," discusses the relationship of Zen to Gödel's Theorem. He calls Zen "intellectual quicksand—anarchy, darkness, meaninglessness, chaos." The quotations he gives from Zen Master Mumon certainly confirm this evaluation.

The analogy between Zen and the formalized concept of mathematics is this:

> Relying on words to lead you to the truth is like relying on an incomplete formal system to lead you to the truth. A formal system will give you some truths, but as we shall soon see, a formal system—no matter how powerful—cannot lead to all truths. The dilemma of mathematicians is: what else is there to rely on, but formal systems? And the dilemma of Zen people is: what else is there to rely on, but words? Mumon states the dilemma very clearly: 'It cannot be expressed with words and it cannot be expressed without words.'[2]

Just as Zen divorces words from reality, the view that mathematics is nothing but formalism divorces mathematics from reality. The result is a world of sometimes self-referential symbols and of sometimes undecidable statements, none of which connect to real life.

This view permits Hofstadter to float off into a realm where symbols have a greater reality than existence.

> A very important side effect of the self-subsystem is that it can play the role of "soul", in the following sense: in communicating with the rest of the subsystems and symbols in the brain, it keeps track of what symbols are active, and in what way. This means that is has to have symbols for mental activity—in other words, symbols for symbols, and symbols for the actions of symbols.

> Of course, this does not elevate consciousness or awareness to any "magical," nonphysical level. Awareness here is a direct effect of the complex hardware and software we have described. Still, despite its earthly origin, this way of describing awareness—as the monitoring of brain activity by a subsystem of the brain itself—seems to resemble the nearly indescribable sensation which we all know and call 'consciousness'.[3]

Further on, Hofstadter suggests "that what we call free will is a result of the interaction between the self-symbol (or subsystem), and the other symbols in the brain."[4] In effect, he has elevated symbols above the need for any mind to say what they symbolize, and even to replace the mind itself. Even reality may be gobbled up by these symbols: "One could suggest, for instance, that reality is itself nothing but one very complicated formal system ... The sole axiom is (or perhaps, was) the original configuration of all the particles at the 'beginning of time' ."[5] And in the end, he grants, nothing but faith justifies his entire system.[6]

Saying anything about Hofstadter means going out on dangerous waters, since so much of what he says is metaphoric. But the fact that he has resorted to metaphors for so much of his writing is indicative of a view that descriptions provide only a rough and indirect guide to reality (as his comments on Zen and words indicate).

The Indeterminate Cat

Another line of mysticism is found in the view that the limitations of measurability in quantum physics imply indeterminacy in reality. (By "indeterminacy" I mean the idea that something can exist without possessing a specific nature and specific characteristics.) The belief that the inapplicability of certain measurements implies that the object in question has no identity is presented as a conclusion of quantum mechanics, but it is actually a premise of the metaphysical Copenhagen interpretation.

The full mystics take advantage of the implicit mysticism of the metaphysical Copenhagen interpretation to undercut belief in reality itself. Zukav asserts that "According to quantum physics there is no such thing as objectivity .... Physics has become a branch of psychology, or perhaps the other way around." John Gribbin says that "[o]bjective reality does not have any place in our description of the universe." Even deeper into the black box fallacy and the primacy of consciousness lies a view which Gribbin cites but does not overtly endorse, that "the whole universe may only owe its 'real' existence to the fact that it is observed by intelligent beings."

Gribbin's own preference is for a multiplicity of universes, with new universes being born by mitosis, so to speak, every time an act of observation collapses a wave function. "What happens when we make a measurement at the quantum level is that we are forced by the process of observation to select one of these alternatives, which becomes part of what we see as the 'real' world; the act of observation cuts the ties that bind alternative realities together, and allows them to go on their own separate ways through superspace, each alternative reality containing its own observer who has made the same observation but got a different quantum 'answer' and thinks that he has 'collapsed the wave function' into one single quantum alternative."7

This view, in which it is the act of observation that "cuts the ties" between realities, gives the consciousness of man godlike powers; one person's act of observing can bring a new

universe into existence. This is, to say the least, a strange place for science to end up. Most physicists reject the idea, but milder forms of the idea that our knowledge of reality is what makes it something definite are disturbingly well accepted in physics. The explicit mysticism of Gribbin and Zukav is worse than the implicit mysticism of the idea that abstractions, formalisms, and probabilities lie at the root of reality, but both come from the same sources.

Science vs. Experience?

Joseph Weizenbaum, in discussing artificial intelligence, has asserted the existence of a far-reaching split between science and reality, declaring that from as long ago as the invention of the clock, "rejection of direct experience was to become one of the principal characteristics of modern science."[8] What is the scientific method? Here is Weizenbaum's characterization of it:

> There is a well-known joke that may help clarify the point. One dark night a policeman comes upon a drunk. The man is on his knees, obviously searching for something under a lamppost. He tells the officer that he is looking for his keys, which he says he lost 'over there,' pointing out into the darkness. The policeman asks him 'Why, if you lost the keys over there, are you looking for them under the streetlight? ' The drunk answers, 'Because the light is so much better here.' That is the way science proceeds too. It is important to recognize this fact, irrelevant and useless to blame science for it. Indeed, what is sought can be found only where there is illumination."[9]

Weizenbaum does not like this approach, but he considers it necessary to science. He leaps upon the areas in which scientists have detached themselves from reality, most notably formal systems, which he correctly refers to as "games."

He argues against "the imperialism of instrumental reason," which is similar to Dreyfus's "calculative reason." Unlike Dreyfus, he does not divide reason in two, arguing rather that "rationality may not be separated from intuition and feeling," but on the whole his idea of "rationality" incorporates more open irrationalism than Dreyfus's "deliberative reason." He rebukes scientists for arriving at factual conclusions that have unpleasant consequences.[10] He asserts that "science deliberately and consciously plans to distort reality,"[11] and he grossly insults the best of his students:

> Then, too, I am constantly confronted by students, some of whom have already rejected all ways but the scientific to come to know the world, and who seek only a deeper, more dogmatic indoctrination in that faith (although that word is no longer in their vocabulary)."[12]

Weizenbaum is attacking a legitimate target: the substitution of models for reality. But he fails to recognize that such a substitution is neither cognitively valid nor scientifically proper; as a result, he fluctuates between denouncing science and asserting that the drunkard's approach is the only one possible. His stance represents a mild version of the reaction that could set in against science if its practitioners continue to build constructs that are further and further removed from reality.

Healing the Breach

Science has reached a crisis in the twentieth century. It has advanced into areas that are far removed from everyday experience, and its application has resulted in devices with capabilities that would once have been considered magical. Having reached this stage, modern scientists and engineers must be more explicitly aware than their predecessors of the relationship of reality to their models and theories. It is easy, as the chain of connections to normal observation grows longer, to fall back on formulas, descriptions, and models as if they were the basic facts of reality, rather than simply the best available characterization of them.

The result is a false kind of scientific view, one that attempts to shape reality to the convenience of its methods of describing them. The most serious single consequence of this view is the exclusion of consciousness from reality because it cannot be observed in an impersonal, "scientific" way. This is the approach of Weizenbaum's drunkard, but it is not good science—not if science means the systematic study of the facts of nature.

Rejecting science because it is not omniscience, because it formulates models which are not absolutely complete descriptions, is an even worse error. The answer to the black box fallacy is not to refuse to deal with black-box descriptions; it is to recognize that there is something in the box, something which a different approach may be able to reveal.

Escaping from this error depends on a recognition of the roots of all knowledge, scientific or not. The starting point of understanding is observation: what we see, hear, or feel is the result of an aspect reality acting on us according to its nature. To proceed from this starting point, we have to integrate what we observe, and in the end arrive at explicit principles that characterize the regularities in the world.

Each level of integration builds upon the one before it; if it results in a contradiction, we have to go back and find out where we made a mistake. Otherwise, a new discovery can't obliterate the ground on which it was built; when someone asserts that quantum physics proves "nothing is real," he is denying the reality of the very observations that led him to that conclusion. When a critic of the "ghost in the machine" asserts that there is no such thing as consciousness because it can't be observed from outside, he is denying the existence of the capacity for observation that makes his criticism meaningful.

Many of the discoveries of modern scientists appear paradoxical; the proper way to deal with them is to ask what facts led up to the paradox. Reality can't contradict itself; there is always a resolution, however difficult it may be to find. If no resolution presents itself, the only thing to do is present the observed facts and state that no explanation is currently apparent. (This is the way to deal with allegedly "paranormal" phenomena that seem to be impossible in rational terms.)

More broadly, what is necessary to keep science on course is a revitalization of the understanding that reason is not a "calculative" or "instrumental" facility that operates on symbols and models; it is the means by which the mind integrates and interrelates the observed facts of reality. This understanding requires avoiding the error of Comte's Positivism, which accepts observation yet subtly undercuts reality. It excludes the "scientific" exclusion of consciousness from consideration, since one cannot accept observation while denying that one has the capacity to observe. It has no place for the elevation of symbols to the

status of reality, for the confusion of observation with creation, or for equivocation between appearance and fact.

To avoid this confusion, one must avoid setting up an opposition between mind and matter, or between logical truths and empirical ones. The mind is fully as natural as a bone or a rock; it must be understood for what it is, not avoided or reduced to its causal factors for fear of churchly contamination. It discovers reality by a process which is both logical and empirical; it must begin with observations and relate them to one another by a process of logic. Creating a separate world of spirit—or of models and symbols—divorced from reality results in throwing away part of reality and refusing to consider it with the tools of reason.

Instead, science must consider all the facts of reality as part of its domain, each to be considered according to its nature and the available means of observation, rather than being forced into molds in the shape of the most convenient models. Perhaps the light isn't so good out there; but if that's where the keys are, that's the place to look. Karl Popper has put the issue very well:

> We should always beware of becoming 'normal scientists': scientists who work blindly, uncritically, within the unconscious presuppositions of a research programme. A 'normal scientist' is not attempting to be as rational as he can be, for he is not trying to be as critical as he can be.[13]

Today's scientists are certainly up to the task of being better than "normal." But they need a more conscious philosophical grounding in order to deal with phenomena where common sense offers little help, or where their ability to create convincing illusions obscures even their own understanding of what they are really doing. The word "metaphysics" has to be pulled out of the trash can of magic and mysticism, and restored to its rightful role as the basis of any fact and any theory. A scientist who recognizes this is fit to deal with any discovery or invention, no matter how bizarre, and search with confidence for its place in reality.

---

1 Branden, p. 15.
2 Hofstadter, *Gödel, Escher, Bach,* p. 252-253.
3 Ibid., p. 387-388.
4 Ibid., p. 710.
5 Ibid., p. 53-54.
6 Ibid., p. 192.
7 Gribbin, p. 237.
8 Weizenbaum, p. 25.
9 Ibid., p. 127.
10 Ibid., p. 263.
11 Ibid., p. 128.
12 Ibid., p. 10.
13 Popper, *Quantum Theory and the Schism in Physics*, p. 33.

Bibliography

Anderson, Alan Ross (ed.), *Minds and Machines*, Prentice-Hall, Inc., 1964. This volume includes Turing's "Computing Machinery and Intelligence" and J. R. Lucas's "Minds, Machines and Gödel."

Ashby, W. Ross, *Design for a Brain*, Second Edition, Science Paperbacks, London and Hall, 1960.

Bohm, David, *Causality and Chance in Modern Physics*, University of Pennsylvania Press, D. Van Nostrand Company, 1957.

Branden, Nathaniel, "The Objectivist Theory of Volition," The Objectivist, January-February 1966. Branden, Nathaniel, *The Psychology of Self-Esteem*, Bantam Books, 1969.

Branden, Nathaniel, "Volition and the Law of Causality," The Objectivist, March 1966.

Comte, Auguste, *Traite Philosophique d'Astronomie Populaire, precedé du Discours sur l'Esprit Positif*, Second Edition (1893), Librairie Antheme Fayard, 1985.

Cropper, William H., *The Quantum Physicists and an Introduction to their Physics*, Oxford University Press, 1970.

Davies, Paul, *Other Worlds: Space, Superspace, and the Quantum Universe*, Touchstone Edition, Simon and Schuster, 1980.

Dretske, Fred I., *Knowledge and the Flow of Information*, MIT Press, 1981.

Dreyfus, Hubert L., *What Computers Can't Do: The Limits of Artificial Intelligence*, Revised Edition, Harper Colophon Books, Harper and Row, 1979.

Dreyfus, Hubert L., and Dreyfus, Stuart E., *Mind over Machine*, The Free Press, Division of Macmillan, Inc., 1986.

Einstein, Albert, *Out of My Later Years*, Philosophical Library, 1950.

Efron, Robert, "Biology Without Consciousness—and Its Consequences," The Objectivist, February-May 1968.

Göde1, Kurt, *On Formally Undecidable Propositions*, Basic Books, 1962.

Gribbin, John, *In Search of Schrödinger's Cat: Quantum Physics and Reality*, Bantam Books, 1984.

Heisenberg, Werner, *Physics and Philosophy*, Harper Torchbooks, 1958.

Herbert, Nick, *Quantum Reality: Beyond the New Physics*, Anchor Books, 1987.

Hofstadter, Douglas R., *Gödel , Escher, Bach: An Eternal Golden Braid*, Vintage Books Edition, 1980.

Hofstadter, Douglas R., ed., *The Mind's I*, Bantam Book, Basic Books, 1981.

Hogan, James, *The Two Faces ofTomorrow*, A Del Rey Book, Ballantine Books, 1979.

Johnson-Laird, Philip N., *Mental Models*, Harvard University Press, 1983.

Jorgensen, Chuck, and Matheus, Chris, "Catching Knowledge in Neural Nets," in *AI Expert*, December 1986.

Kelley, David, *The Evidence ofthe Senses*, Louisiana State University Press, 1986.

Kent, Ernest W., *The Brains of Men and Machines*, McGraw-Hill, 1981.

Ladd, Scott, *The Computer and the Brain*, Bantam Books, The Red Feather Press, 1986.

McCorduck, Pamela, *Machines Who Think*, W. H. Freeman and Company, 1979.

Marvin Minsky, "Matter Mind, and Models," *Proc. International Federation of Information Processing Congress 1965*.

Minsky, Marvin (ed.), *Semantic Information Processing*, MIT Press, 1968. Includes Minsky, "Matter, Mind, and Models."

Mises, Richard von, *Probability, Statistics, and Truth*, Second Revised English Edition, Dover Publications, George Allen & Unwin Ltd., 1957, translated from third German edition, J. Springer, 1951.

Peikoff, Leonard, "The Analytic-Synthetic Dichotomy," *The Objectivist*, May-September 1967.

Popper, Karl R., *Objective Knowledge*, Oxford University Press, 1972, 1979.

Popper, Karl R., *Quantum Theory and the Schism in Physics*, Rowman and Littlefield, 1982.

Rand, Ayn, *Introduction to Objectivist Epistemology*, A Mentor Book, New American Library, 1979. This edition includes Peikoff's "The Analytic-Synthetic Dichotomy."

Rand, Ayn, *The Virtue of Selfishness*, A Signet Book, New American Library, 1964.

Raphael, Bertram, *The Thinking Computer: Mind Inside Matter*, W. H. Freeman and Company, 1976.

Russell, Bertrand, *Human Knowledge: Its Scope and Limits*, Simon and Schuster, 1948.

Russell, Bertrand, *Principles of Mathematics*, Second Edition, W. W. Norton and Company, 1938.

Sagan, Carl, *Cosmos*, Random House, 1980.

Schank, Roger C. , with Childers, Peter, *The Cognitive Computer: On Language, Learning, and Artificial Intelligence*, Addison-Wesley, 1984.

Schank, Roger C., and Colby, Kenneth Mark (eds.), *Computer Models of Thought and Language*, W. H. Freeman and Company, 1973.

Simons, Geoff, *Are Computers Alive?* , Birkhauser, 1983.

Sowa, J. F., *Conceptual Structures: Information Processing in Mind and Machine*, Addison-Wesley, 1984.

A. M. Turing, "Computing Machinery and Intelligence," *Mind*, Vol. LIX, No. 236, 1950.

Uhr, Leonard, *Pattern Recognition, Learning, and Thought: Computer-Programmed Models of Higher Mental Processes*, Prentice-Hall, 1973.

Weizenbaum, Joseph, *Computer Power and Human Reason: From Judgment to Calculation*, W. H. Freeman and Company, 1976.

Wilber, Ken, *Quantum Questions: Mystical Writings of the World's Great Physicists*, New Science Library, 1984.

Winston, Patrick Henry, *Artificial Intelligence*, Second Edition, Addison-Wesley, 1984.

Zukav, Gary, *The Dancing Wu Li Masters: An Overview of the New Physics*, Bantam Books, 1980.