1. WHAT IS ARTIFICIAL INTELLIGENCE?

One of the fascinating aspects of the field of artificial intelligence (AI) is that the precise nature of its subject matter turns out to be surprisingly difficult to define. The problem, of course, has two parts, since securing an adequate grasp of the nature of the artificial would do only as long as we were already in possession of a suitable understanding of the idea of intelligence. What is supposed to be "artificial" about artificial intelligence, no doubt, has to do with its origins and mode of creation in arising as a product of human contrivance and ingenuity rather than as a result of natural (especially biological or evolutionary) influence. Things that are *artificially intelligent*, in other words, differ from those that are *naturally intelligent* as artifacts that possess special properties ordinarily possessed by non-artifacts. So these are things that have a certain property (intelligence) as a result of a certain process (because they were created, designed, or manufactured in this way).

These artifacts, of course, are of a certain special kind in the sense that they are commonly thought of as being *machines*. If machines are simply things that are capable of performing work, then, since human beings are capable of performing work, human beings turn out to be machines, too. If anyone wants to know whether or not there are such things as intelligent machines, therefore, the answer seems obvious so long as human beings are (at least, some of the time) *intelligent*. A more difficult question, whose answer is less evident, however, arises from distinguishing between animate and inanimate machines. Human beings, after all, may result from human contrivance and ingenuity, but they are clearly biological in their origin. The issue thus becomes whether or not *inanimate* machines, as opposed to human beings, are capable of possessing a certain special property that human beings are supposed to display – if not always, at least on certain special occasions.

The problem revolves about the identification or the definition of what is meant to be "intelligent" about artificial intelligence. The dictionary may seem to be an appropriate place to begin. *Webster's New World Dictionary* (1988), for example, defines "intelligence" as "a) the ability to learn or understand from experience; ability to acquire and retain knowledge; mental ability; b) the ability to respond quickly and successfully to a new situation; use of the faculty of reason in solving problems, directing

J. H. Fetzer, *Artificial Intelligence: Its Scope and Limits*

conduct, etc., effectively; c) in *psychology*, measured success in using these abilities to perform certain tasks". If the concept of intelligence is to be applicable (at least, in principle) to inanimate machines, however, it must not be the case that inanimate machines could not possibly be intelligent as "a matter of definition".

Matters of definition, of course, are often thought to be arbitrary or capricious, capable of resolution merely by mutual agreement: "The question is who is to be master, that's all!" But much more hinges upon the meaning that we assign to the words that we use when we address specific problems and search for suitable answers. In this case, the establishment of an appropriate definition for "intelligence" becomes equivalent to a theory of the nature of intelligence as a phenomenon encountered in the world. Hence, for example, any analysis that has the effect of *excluding* human beings from the class of intelligent things, on the one hand, or that has the effect of *including* tables and chairs within the class of intelligent things, on the other, would thereby display its own inadequacy. By reflecting upon "exemplars" of the property in question – instances in which its presence or absence and other features these things possess are not in doubt – it should be possible to achieve conceptual clarification, if not to secure complete agreement, on the crucial features of intelligence itself, which amounts to a prototype theory.

Even if we assume that human beings *are* among the things that happen to be intelligent, the problem remains of isolating those specific aspects of human existence that are supposed to be "intelligent". Since human beings exhibit anger, jealousy, and rage, it might be asked if inanimate machines would be intelligent if they could exhibit anger, jealousy, and rage. Notice, the point is not the trivial one that we are likely to mistake anger, jealousy, and rage for "intelligence" as a matter of fact, but rather the subtle one that *unless we already know that anger, jealousy, and rage are not aspects of "intelligence"*, we are not able to rule them out. If human beings are the best exemplars of the property in question, yet are widely observed to exhibit anger, jealousy, and rage, then by what standard, yardstick, or criterion can we tell that these are *not* some of the features characteristic of intelligence? And, in fact, there are situations, circumstances, and conditions under which anger, jealousy and rage would be intelligent. So how can we possibly tell?

While these definitions may seem helpful with respect to human beings, they remain problematic in application to machines. If we take for granted that human beings do possess the ability to learn or understand from

experience, for example, that in turn invites elucidation of precisely what happens when something – it could be anything at all – "learns" or "understands" from experience. So unpacking "intelligence" now becomes unpacking "learning or understanding from experience". This may appear to be a doubtful advance, especially since whether or not inanimate machines can learn or understand from experience appears to be no less troublesome than the question whether or not inanimate machines ever have the capacity to be intelligent. Similar reflections obtain with respect to the ability to acquire and to retain knowledge and with respect to mental ability in general: whether inanimate machines can acquire and retain knowledge (or whether inanimate machines can have minds) appears to offer few benefits in understanding "intelligence" – unless we already understand the nature of knowledge or of learning or of mentality, in which case theoretical progress might yet be possible, after all.

CRITERIA OF INTELLIGENCE

If we possessed the practical ability to sort things out, to determine of any specific thing whether or not it ought to be included within the class of intelligent things, of course, it might not matter if we were to be less than completely successful in devising a theory of intelligence. In this case, perhaps a *criterion* could serve in lieu of a *definition*, where the criterion functions as a (usually reliable, but not therefore infallible) evidential indicator for deciding, in a given case, whether or not that case is an instance of this property. Indeed, several tests of this general type have been proposed, among which the most famous is that devised by Alan Turing, which is now known as "the Turing Test". If the Turing Test – or another criterion – were to prove to be a usually reliable evidential indicator of the presence or the absence of intelligence, that might be enough (even if it were not infallible).

The Turing Test. Turing (1950) offered a suggestion that has turned out to have widespread appeal, especially because it offers the hope of resolving these difficult problems at a single stroke. The test he proposed is a special case of what is known as "the imitation game", in which two contestants are pitted against one another, but in opposite ways, with respect to some third party. An example might be the separation of a male and a female behind a curtain, where the third party attempts to guess the sex of the contestants, based strictly upon their answers to questions he poses. The

female might attempt to lead the third party into the realization that she is female, while the male would endeavor to deceive him into believing that he is female as well. The means of communication between parties, of course, would have to be one that would not reveal the sex of the communicant for the game to be a success. What is most important is that almost any property might do.

Turing recognized that an especially interesting application of the imitation game would pit an inanimate machine against a human being, where the property in question was not their sex but their intelligence. He reasoned that if, under the appropriate conditions (including, therefore, a suitable means of communication, such as by teletype machine), a third party could distinguish between the human being and the inanimate machine no better than he could distinguish between a male and a female on the basis of their answers to questions he might pose, as before, then it would be reasonable to conclude that the parties were equal with regard to the property in question. The criterion that might be used to empirically ascertain whether or not a machine possesses intelligence, in other words, would be its capacity to fool the third party – a human being – into believing that it is human, too. The result would depend upon its success in inducing a specific false belief.

The Chinese Room. Indeed, from its conception around 1950, the Turing Test remained largely unchallenged as a suitable criterion for machine intelligence until confronted by the scenario that has come to be known as "the Chinese Room", posed by John Searle [(1980) and (1984)]. Searle offered his argument as a counterexample to certain claims about *machine understanding* that were then being advanced by Roger Schank on behalf of "scripts" (with fascinating features that we shall explore in Chapter 7). But the same case tends to undermine the plausibility of the Turing Test with respect to machine intelligence, precisely because it tends to undermine the plausibility of the imitation game within which it assumes its significance. The Searle counterexample thus provides an indirect argument against the Turing Test, but one that – as we shall discover – can receive considerable reinforcement.

The argument Searle fashioned is a thought experiment of the following kind. Imagine that someone (Searle himself, in particular) is locked into an enclosed room with one entry through which Chinese symbols are now and then sent into the room and one exit through which (the same or different) Chinese symbols could be sent out. Thus, if the occupant of that

room, who is fluent in English but ignorant of Chinese, had in his possession a book of instructions, written in English, directing that certain specific Chinese characters should be sent out when certain other Chinese characters happen to be sent in, then if that person were to act in accordance with those instructions, it might appear to those outside the room as though, contrary to the hypothesis, he understands Chinese. While this special arrangement might fool observers into believing he understands Chinese, however, he does not.

Indeed, Searle strengthened his argument by suggesting that the characters sent into the room might be called "input", the characters out "output", and the book of instructions "a program"; yet however impressive its performance might appear, it would not be the case that an input-program-output system of this kind actually understood Chinese. Even though its publicly observable inputs and outputs might be indistinguishable from those of another input-program-output system that actually understood Chinese (say, the publicly observable inputs and outputs of a fluent Chinese interpreter), any inference to the conclusion that this system actually understood Chinese would be completely mistaken, since that conclusion would be untrue.

Looking back upon the imitation game itself from this point of view, moreover, it is striking to realize that, even in the first instance, if a male were successful in convincing a third party that he were female, that obviously would not change his sex! Analogously, if an input-program-output system of the kind Searle envisions were to fool outside observers into the belief that it understood Chinese, that obviously would not mean that this system actually understood Chinese! And – most importantly – if an input-program-output system were to display behavior that was indistinguishable from the behavior displayed by a human being, that would not dictate that that system really possessed intelligence! There is, after all, a fundamental difference between appearance and reality. The net impact of Searle's case thus appears to be a drastic reduction in the plausibility of the Turing Test.

The Korean Room. Some students of AI have drawn different conclusions from the premises described. William J. Rapaport, especially, takes a different tack, suggesting that the Turing Test can be entertained as a tacit reflection of suitable conditions for the possession of intelligence. Rapaport seeks to dispurse the damage wrought by the Chinese Room argument with a variation of his own, which is intended to demonstrate that

Searle's argument may be less conclusive than it initially appears. He proposes the scenario of a Korean professor who is an authority on Shakespeare, even though he does not know the English language. Having studied Shakespeare's work in excellent Korean translations, he has overcome what for others appeared to be an insuperable obstacle to master the plays. His articles on the Bard, which have been translated into English on his behalf, have met with great acclaim. He is viewed as an expert scholar [Rapaport (1988), pp. 114-115].

The point of this fanciful account is that the Korean professor stands to Shakespeare's plays as the-man-in-the-room stands to Chinese. Yet, as Rapaport himself insists, it would surely be a mistaken inference to conclude that the Korean professor does not understand *something;* and, by parallel reasoning, it would similarly be a mistaken inference to conclude that the-man-in-the-room understands *nothing*. The difficulty encountered at this juncture, of course, is unpacking precisely what it is that this Korean professor is supposed to understand, since even Rapaport does not deny that, whatever it may be, he certainly does not understand *English*. By parallel reasoning, once again, however, it seems to follow that for the-man-in-the-room, whatever he is supposed to understand, it is certainly not *Chinese*.

Even more important, from a theoretical point of view, however, is the perspective Rapaport brings to bear upon the Turing Test itself. Correctly observing that Turing (1950) rejected the question, "Can machines think?", in favor of the more behavioristic question, "Can a machine convince a human to believe that it (the computer) is a human?", Rapaport infers that,

> To be able to do that, the computer must be able to understand natural language. So, understanding natural language is a necessary condition for passing the Turing Test, and to that extent, at least, it is a mark of intelligence. I think, by the way, that it is also a sufficient condition. [Rapaport (1988), p. 83]

Whether or not the capacity to understand natural language is a necessary condition for the possession of intelligence, of course, invites controversy in the case of non-human animals, many of which display behavioral patterns that certainly *seem* to be intelligent. (Ask any owner of a cat!) But even if understanding natural language is merely a sufficient condition, rather than both necessary and sufficient for intelligence, it raises intriguing questions.

Rapaport could argue, for example, that the appearance-of-understanding qualifies as understanding with regard to understanding the plays, even though the appearance-of-being-of-female-sex does not qualify as being-of-female-sex. The issue might then become whether this really is an instance in which appearance *is* reality [a position that Rapaport (1988) and Shapiro and Rapaport (1988) have proposed]. Such a defense, however, although ingenious, possesses very little *prima facie* plausibility. That it might appear to outsiders as though the-man-in-the-Korean-room understands Shakespeare's plays, after all, is no more contrary to the hypothesis that he does *not* understand Shakespeare's plays than that it might appear to outsiders as though the-man-in-the-Chinese-room understands Chinese is contary to the hypothesis that he does *not* understand Chinese. Hence, as a criterion of intelligence, the Turing Test still seems to be an unreliable or a poor one.

Notice that, although Rapaport views his conception as essentially the same as Turing's, there are several reasons for doubt. Turing's Test, in particular, concerns whether or not a machine *could fool a human into thinking that it is human too*. So its success hinges upon whether or not a belief could be induced by the behavior of a machine without concern for whether or not that machine actually understands anything at all. Rapaport's Test, by comparison, concerns whether or not a machine *could really understand natural language*. Rapaport's Test depends upon whether or not a machine actually understands a natural language without concern for whether or not it might use its linguistic ability to induce beliefs in any human being. The Turing criterion concerns belief states and the conditions under which they might justifiably be acquired, while Rapaport's concerns the kind of thing a machine happens to be apart from any beliefs that we might form about it.

While Rapaport wants to interpret his test as a criterion of intelligence as-detected-by-the-Turing-Test itself, moreover, the benefits that may be derived from its introduction tend to depend on presuming that the Turing Test functions as a criterion of intelligence, whereas Rapaport's Test serves as something akin to a definition. Notice, especially, that if the ability to understand natural language were both necessary and sufficient to possess intelligence, then anything that were unable to understand natural language could not be intelligent, while anything that were intelligent could not fail to understand natural language. And this remains the case whether Rapaport intended it as a clarification of the Turing Test or not. Consequently, I shall interpret Rapaport's Test as proposing a definition

of the property under consideration rather than merely positing a criterion of its presence.

VARIETIES OF INTELLIGENCE

There are at least three reasons for appreciating the point of view that Rapaport's Test supplies. The first concerns the focus it provides upon the kind of thing that something happens to be rather than the beliefs that we fallible human beings might form about it. Questions about what kinds of things there happen to be fall within the domain of *ontology* and are often referred to as "ontic" issues. Questions about the beliefs that we, as fallible human beings, happen to adopt, by contrast, fall either within the domain of *psychology* and are frequently referred to as "cognitive" matters (when they concern the beliefs we do adopt) or within the domain of *epistemology* and are referred to as "epistemic" questions (when they concern the beliefs we ought to adopt). The differences between them are immense: Turing's Test is an epistemic criterion, for example, while Rapaport's is an ontic definition instead.

If Rapaport's Test takes us from a criterion to a possible definition of intelligence, it also has the virtue of focusing attention upon the pivotal role of language. It is widely assumed that there is a fundamental relationship between language and thought, especially when the suggestion is proposed that all thinking takes place in language. Without language, there could be no thought, if this suggestion is correct. Thus, by accenting the importance of language with respect to the nature of intelligence, such a test invites us to consider an ability that may very plausibly be supposed to be far more essential to the nature of intelligence than anything like the ability to fool a human being into embracing a false belief. And, indeed, whether or not Rapaport's Test ultimately survives as a definition of intelligence, we shall discover ample grounds for regarding the ability to understand natural language as a crucial factor in distinguishing natural and artificial intelligence.

The third reason for appreciating Rapaport's Test is much more subtle, reflecting as it does the importance of implied *success* in such descriptions as "the ability *to learn and understand* from experience, ability *to acquire and retain* knowledge, ability *to respond quickly and successfully* to new situations", and the like. All of these characterizations, including Rapaport's conception of the ability *to understand* natural language, imply suc-

cess in undertaking a certain activity or in exercising a specific faculty or ability. Notice, in particular, that *attempting* to learn, *trying* to acquire and retain, *wanting* to respond quickly – but *without* success – would not ordinarily be entertained as exemplifications of intelligence. Indeed, from this point of view, there is an important difference between "things that can do things successfully" and "things that can do things by using their minds". There are lots of things, after all, that can do things by using their minds, where those things are stupid, foolish, or otherwise lacking in intellectual merits.

Notice that an automated mechanism, say, a device for capping bottles as they emerge from a production line, might perform its intended function quite successfully and very dependably without raising any questions as to whether or not it possessed intelligence. In fact, all sides – every party – to these disputes would be inclined to agree that this machine (call it a robot, if you will) does not possess any intelligence at all. Yet the successful way in which it performs its task would certainly merit the accolade implied by reference to this entity as an "intelligent machine". No doubt, there is room for debate, even in simple cases like this one, depending upon the features of mechanisms of this kind. Nevertheless, later on, when we have occasion to reflect upon what we have discovered in the meanwhile, it may turn out to be worthwhile to consider just what examples such as this have to tell us.

Rationality. Taking a cue from the last of these issues, yet another way the problem of intelligence might be addressed would be as a synonym for rationality. The idea of rationality, however, is highly ambiguous, insofar as there appear to be at least three varieties of rationality, which, in turn, suggests the possibility that there might be three corresponding varieties of intelligence. The first of these concerns *the rationality of ends*, meaning the rationality of choosing or selecting specific goals, aims, or objectives as worthy of pursuit. While we often take it for granted that anyone can try to do anything they want – whether or not it is illegal, immoral, or fattening – if someone seriously, rather than wishfully, wanted to discover a number which is both even and odd (where these terms mean what they ordinarily mean within the theory of numbers), then anyone aware of this state of affairs might think something was amiss, precisely because it is logically impossible, i.e., strictly inconsistent, for any number to be both odd and even.

Hence, a necessary condition for the rationality of ends appears to be

that the attainment of that aim, objective, or goal must not be a logical impossibility. That this requirement is clearly not sufficient follows from reflection upon another class of cases, including individuals who sincerely rather than wistfully want to be in two places at the same time, to consume food without exercising and still lose weight, etc. (How familiar they look!) In these cases, what is wrong is not that the situations described are logically inconsistent; these are logically possible states of affairs that might have been possible if only the world had been different in certain respects. For these scenarios are fanciful precisely because their realization could occur only by violating natural laws (concerning the locations of physical things, the relations between food consumption and weight loss, etc.). It appears as though a rational goal must be a physical as well as a logical possibility.

Suppose, however, that someone wanted to secure a state of affairs whose attainment was neither logically nor physically impossible, perhaps by being the first man to climb Mt. Everest, the second man to marry Elizabeth Taylor, or something such. In cases of this kind, the pursuit of these goals would be pointless and without purpose, not because it would be logically impossible (in relation to a particular language) or because it would be physically impossible (in relation to the world's own laws), but because it would be historically impossible (in relation to the history of the world up to now). One cannot do (for the first time) something that has already been done, even though whether or not it can be done (for the first time) is a function of the historical past rather than of logic or of physical laws.

There are other types of rationality, of course, including, in particular, *rationality of action* and *rationality of belief* [cf. Hempel (1962)]. Rationality of action involves choosing means that are appropriate to attain one's ends, where means are "appropriate" in relation to one's ethics, abilities, capabilities, and opportunities. It would be irrational in this sense for a skinny vegetarian with no other source of support to enter a hamburger eating contest for a $5.00 prize because he aspires to travel to Europe on the Concorde. Rationality of belief, by comparison, involves accepting all and only those beliefs that are adequately supported by the available relevant evidence, where the form this evidence could take might be perceptual, inductive, or deductive, as we shall subsequently ascertain. It would be irrational in this sense for someone who has seen (direct and indirect) evidence that man has landed on the moon to fail to believe that man has landed on the moon, unless the total available evidence were to override

that belief. These are matters we shall discuss again in Chapters 4 and 5.

Intentionality. On the basis of these reflections, the evidence tends to suggest that rationality (in any of its senses) does not provide a promising avenue toward understanding "intelligence" in the sense appropriate to AI. But the reason this is so should not be overlooked, since it represents (what appears to be) one of the fundamental differences between human beings and inanimate machines. For rationality of ends, rationality of action, and rationality of belief tacity presuppose the existence of agents, organisms, or entities whose behavior results (at least in part) from the causal interplay of motives and beliefs, where *motives* are wants and desires of a system with *beliefs* that might possibly be true. But these conditions are not very likely to be ones that an inanimate machine should be expected to satisfy.

Notice that the rationality of ends, for example, involves choosing or selecting appropriate goals, aims, or objectives as worthy of pursuit, where the propriety of those choices from a subjective point of view is a function of the range of available alternatives, the means that might be employed in their pursuit, and the morality of adopting such means to attain those ends. One view of the most defensible conception of decision-making activities by rational human beings involves envisioning them in terms of preference relations between alternative states of affairs (or "payoffs"), which represent – explicitly or implicitly – expected utilities (in some appropriate sense) as a function of expectations and desirabilities. [Eells (1982) reflects some recent philosophical work in this area, while Cohen and Perrault (1979), Allen and Perrault (1980), and Halpern (1986) exemplify some approaches in AI.]

Important theoretical questions arise here between different accounts of decision-making, including the respective merits of optimizing, satisficing, and cost-benefit decision policies. For our purpose, however, what is significant about all this is that these are extremely implausible questions to ask in relation to any machine. One can always inquire as to the principles of design that entered into a machine's construction, of course, but it is simply silly to raise questions such as whether the bottle-capping robot described above would prefer to do something else instead. Even if we surprisingly frequently indulge ourselves by ascribing human properties to inanimate things – suggesting, for example, that our lawn mower might not want to start (as though it were driven by motives and desires) – we

accept these practices as metaphorical vestiges of anthropomorphic tendencies of the past. They are nothing more than a very convenient manner of speaking.

As though he could convert vice into virtue, Daniel Dennett (1971) has endorsed "the intentional stance", which occurs when beliefs and desires *are* ascribed to inanimate things, such as chess-playing machines. Thus,

> Lingering doubts about whether the chess-playing computer *really* has beliefs and desires are misplaced; for the definition of intentional systems I have given does not say that intentional systems *really* have beliefs and desires, but that one can explain and predict their behavior by *ascribing* beliefs and desires to them.... The decision to adopt the strategy is pragmatic, and is not intrinsically right or wrong. [Dennett (1971), pp. 224-225]

The principal advantage of adopting the intentional stance, in Dennett's estimation, is that "it is much easier to decide whether a machine can be an intentional system than it is to decide whether a machine can *really* think, or be conscious, or morally responsible" [Dennett (1971), p. 235]. On that score, there can be no doubt. But the fundamental questions still remain.

It should be obvious that the intentional stance, at best, offers (what we may think of as) a practical technique to adopt in attempting to explain and to predict the behavior of inanimate things. In emphasizing the pragmatic dividends of embracing such an attitude, however, Dennett implicitly endorses the position known as *instrumentalism*, which maintains that we need not believe in the existence of entities and properties that lie beyond our observational capacity or our perceptual range but may treat them instead as, say, "convenient fictions". An alternative position, known as *realism*, denies that imperceptible and unobservable entities and properties are therefore any the less real. While the evidence for their existence must be indirect in character, when it supports theories that posit their existence, it supports belief in their existence as well. Dennett's account appears to be a special case of instrumentalism regarding the existence of mental entities.

Some questions, of course, are hard to answer. In suggesting that the intentional stance poses a question that is easier to answer than "whether a machine can *really* think", Dennett has changed the subject. We are not looking for an answer to the question, "Can it be beneficial to regard some inanimate machines as though they were thinking things?" The answer to that question, I am willing to agree, is almost certainly, "Yes". Even here,

however, the benefits are less vast than Dennett promises them to be. If machines really lack beliefs and desires, "explanations" ascribing to them properties that they do not have surely cannot be adequate – they cannot even be true! So long as truth is a condition for the adequacy of an explanation, the most that the intentional stance can provide is predictive utility.

Moreover, it should also be apparent that Dennett has taken two steps backward in embracing an epistemic criterion in lieu of an ontic definition. If our objective is to understand whether or not machines can have minds, his position does not help us. But it can be useful to come to the realization of how widespread the practice has become of taking for granted a certain mode of speech whose consequences we would tend to deny, if only they were made explicit. And it calls back to mind the Turing Test itself insofar as it hinges upon the question of whether or not some digital machine might be able to fool a human being into believing that it is human too. For surely it makes sense to talk about a machine "fooling" a human being only if that effect has been achieved deliberately (or on purpose), i.e., as a causal consequence of that machine's own motives and beliefs. Otherwise, the best that can be said is that a human might mistake the behavior of this machine for that of a human being but not that he has been "deceived" by such a thing.

Humanity. A more candid and less question-begging response to this entire problem complex has been advanced by Eugene Charniak and Drew McDermott. In their impressive survey of AI, Charniak and McDermott suggest, "Artificial intelligence is the study of mental faculties through the use of computational models" [Charniak and McDermott (1985), p. 6]. The assumption underlying this conception is that, at some suitable level, the way in which the brain functions is the same as the way in which certain computational systems – digital machines, in particular – also function. Yet they also maintain, "The ultimate goal of AI research (which we are very far from achieving) is to build a person, or, more humbly, an animal" [Charniak and McDermott (1985), p. 7]. Science fiction addicts might rejoice and theologians may cringe, but there, stated as baldly as could be, is what to many appears to be a fantasy: AI is attempting to create an artificial human being!

For the moment, let us defer consideration of the theoretical tenability of this conception and instead attempt to appreciate the various ingredients whose availability would be required to make its realization even

remotely plausible. Specifically, the situation can be schematized in terms of a simple stimulus-process-response framework or model, where stimuli are "inputs", responses are "outputs", and processes – for now – may be viewed as "black boxes". The precise content of these black boxes is absolutely crucial to AI as Charniak and McDermott describe it, since if, for example, what is going on inside a black box when it generates outputs from inputs is not a computational process, then AI is – to that extent, at least – a failure. If the reason the black box is not a computational process is that AI research has not yet captured the way in which humans operate, then it may just be a failure in practice; but if AI research has captured the way in which humans operate and the process is not computational, then it has to be a failure in principle.

Hence, the following diagram suggests the model that tends to be tacitly assumed within AI when human beings are compared with digital machines:

	Human Beings:	*Digital Machines*:
Domain:	Stimuli	Inputs
Function:	Processes	Programs
Range:	Responses	Outputs

Fig. 1. The Basic Model.

Notice, in order for this analogy to be correct (or well-founded), it has to be the case that human beings can be characterized by processes that are properly viewed as functions from (the domain) stimuli to (the range) responses and that digital machines can be characterized as well by programs that are properly viewed as functions from inputs to outputs. The parallel between human beings and digital machines need not hold in every respect for such an analogy to be acceptable, however, since inanimate machines and human beings are otherwise very different kinds of things in innumerable respects.

As a consequence, there is a tendency to focus upon those processes that appear to be most likely to be unaffected by the differences between animate and inanimate things, where the subset of processes on which the most attention has focused are those that involve the processing of data, information, and knowledge. Indeed, the processes of logical and mathematical reasoning ("deductive inference"), of natural and technical language utilization ("linguistic capability"), and of visual and sensory information acquisition ("perceptual inference") have tended to receive

considerable attention. Digital machines may not ingest, digest, or otherwise process food, water, and spirits as humans do, but these (presumably non-computational) processes can be set aside in favor of those "cognitive" processes they happen to share.

This more restricted focus suggests that AI is not so much attempting to create an artificial "human being" as an artificial "thinking thing", a prospect that theologians may find less menacing. In particular, Charniak and McDermott [(1985), p. 7] advance a sketch of what they think AI is really all about:

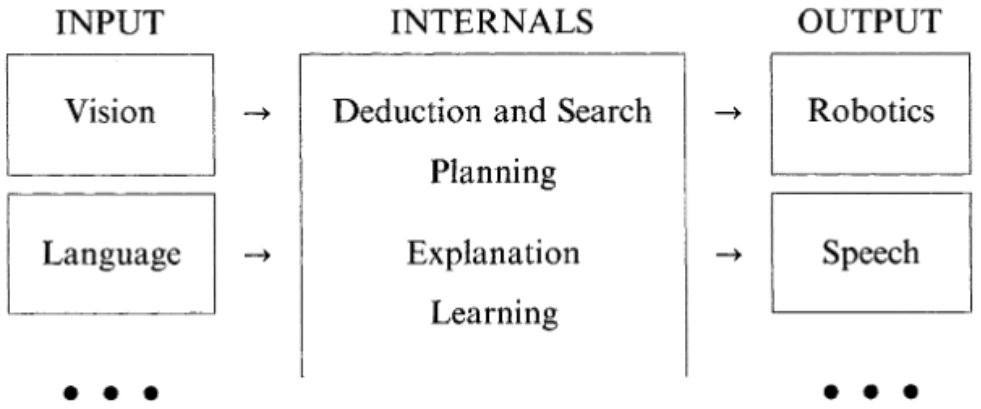

Fig. 2. Mental Faculties.

Here, "INTERNALS" stands for processes going on inside a human being (or inside a digital machine) and the sets of dots "..." and the open box imply that there may be more going on inside and outside the black box than has been portrayed. But bear in mind that, since digital machines are what AI researchers have to work with, any cognitive process that is not a computational process cannot be a computer process, but it may exist, nevertheless.

A great deal of excitement has been created by the claim that all cognitive processes are computational [see, for example, Pylyshyn (1984)]. This contention, of course, might be true or it might be false – a question whose answer we are going to discover. What is already clear, however, is that a description of AI as "the study of mental faculties through the use of computational models" harbors an important equivocation, since these "models" themselves might be of one or another kind, namely: there are models that *simulate* (by effecting the right functions from inputs to outputs) and models that *replicate* (by effecting the right functions by means of the very same – or similar – processes). Simply put, this means that

machines might be able to do some of the things that humans can do (add, subtract, etc.), but they may or may not be doing those things the same way that humans do them.

THE AIM OF THIS INQUIRY

Models that simulate and models that replicate, as they are intended to be understood here, achieve those effects independently of consideration for the material (or "medium") by means of which they are attained. Another category of models, therefore, consists of those that *emulate* (by effecting the right functions by means of the same – or similar – processes implemented within the same medium). A relation of emulation between systems of different kinds thus entails that they be constituted of similar material (or components), which, in the case of humans, would involve flesh, blood, and nerves. Indeed, since digital machines are made of electronic components while human beings (leaving bionic men and women to one side) are not, it would be an inappropriate imposition to insist upon emulation as AI's goal.

Such a requirement, after all, would have the consequence of begging the question with regard to the fundamental issue of whether machines can have minds or not, since it would imply that only human beings are capable of emulating the cognitive processes of other human beings. Almost no one, however, challenges the capacity for AI to achieve the objective of simulating the cognitive processes of human beings across a broad range of species of reasoning, including, for example, number crunching. It would be odd, indeed, to doubt whether digital machines can add, subtract, etc., in the mode of simulation, since even hand-held calculators can generate the right output from the right input. The crucial question concerns whether machines could ever be constructed that would add, subtract, etc., in the mode of replication.

There are those, such as Fred Dretske (1985), who go so far as to deny that computers can add, subtract, etc., after all. But in adopting his position, Dretske is not suggesting that entering the sequence "2", "+", "2" with a handheld calculator could not produce "4" as its result. On the contrary, Dretske would concede that computers can simulate human reasoning while denying that that they can replicate human reasoning processes. The difficulty – and it is, of course, rather grave – becomes that of establishing a suitable evidential warrant for arriving at a conclusion

concerning whether digital machines might or might not be able to replicate the cognitive processes of human beings. This question, moreover, is not one of technology, since there are few who would want to claim that current machines are capable of realizing such a goal today. Properly conceived, the question is one of physical possibility.

From this perspective, the purpose of this inquiry is to assess the field of artificial intelligence with respect to its scope and limits. Thus, one of the most pressing issues that we shall confront is whether or not machines have the potential for replicating human cognitive processes. Since this question can only be answered if we have some idea of the nature of human cognitive processes themselves, the character of mentality itself has to be addressed. Thus, the next two chapters are devoted to investigating the nature of mentality with respect to humans, other animals, and machines. The pursuit of this issue leads to an elaboration of the conception of minds as semiotic (or "sign-using") systems and to a comparison of this conception with the physical symbol system conception advanced by Alan Newell and Herbert Simon.

The arguments presented in Chapter 2 indicate that digital machines can qualify as symbol systems in Newell and Simon's sense, but that symbol systems of this kind are not semiotic systems. Even though these machines are capable of symbol processing in Newell and Simon's sense, they do not therefore qualify as the possessors of mentality. These issues are carried further in Chapter 3, moreover, where computational, representational, and dispositional conceptions are explored. Additional reasons are advanced for doubting that computers as currently conceived could possibly satisfy the desiderata required of thinking things. The possibility that machines of a different design (implementing parallel processing by means of connectionist architecture, for example) might have minds, however, cannot be entirely ruled out.

Chapter 2 emphasizes the nature of mentality, while Chapter 3 concerns the nature of language. The position developed in these chapters maintains that mentality involves a triadic relation between signs, what they stand for, and sign users; that computational conceptions reduce language and mentality to the manipulation of possibly meaningless marks; that even representational conceptions cannot salvage the standard conception, because they cannot account for the meaning of the primitive elements of language; and that only a pragmatical conception – that emphasizes the interrelations between signs, what they stand for, and sign users – can provide an adequate resolution of these difficulties. Thus, the symbol

system conception may be suitable for digital machines but does not reflect the mentality of human beings.

Fortunately, the potential benefits of AI are not restricted to those that would be available if digital machines could replicate the mental processes of human beings. The representation of knowledge, the development of expert systems, the simulation and modeling of natural and of behavioral phenomena, and the rapid processing of vast quantities of data, information, and knowledge are destined to assume a major role in shaping the future of our species. In order to comprehend the full range of AI capabilities, therefore, it is indispensable to acquire an understanding of the nature of knowledge. Chapter 4 provides an introduction to the theory of knowledge and aims at the compact presentation of the crucial distinctions essential to this domain.

Perhaps the most important distinction that arises here, moreover, concerns the difference between natural and artificial language; for the results of the preceding chapters, which undermine the prospects for digital machines to understand natural language, reinforce their potential for securing enormous benefits through the utilization of artificial language. The level of discussion found in this chapter may be too detailed for beginners and not detailed enough for experts, but it affords a point of departure for an analysis of "knowledge" within the context of AI as presented in Chapter 5. The differences between ordinary and scientific knowledge are explored, including the nature of common sense and of defeasible reasoning. The implications of this approach for the frame problem are given initial consideration.

Chapter 6 focuses upon the nature of expert systems, with special concern for epistemic aspects of their construction and utilization. Even though expert systems are not in fashion with AI researchers preoccupied with the character of cognitive processes, they afford an excellent illustration of the extent to which AI can produce valuable products within the scope of simulation. Indeed, I think that expert systems will prove to be one of the most enduring legacies of the AI revolution. Chapter 7 continues this exploration by applying the distinctions introduced in previous chapters to understanding the distinctive characteristics of some of the most widely used modes of knowledge representation – semantic networks, predicate calculi, and scripts and frames – with special concern for their varied strengths and weaknesses.

Chapter 8 changes the cadence by taking a look at the epistemic foundations of programs themselves by exploring distinctions between validity

and soundness, algorithms and programs, abstract entities and physical systems, and pure and applied mathematics. Some readers may want to bypass this chapter, but others will appreciate its general significance for understanding the epistemic limitations of computers and their programs. Chapter 9, finally, returns to the problems of cognition with which the book began, turning once more to the differences between humans and machines concerning the nature of mentality. This summation of the situation is then extended to an exploration of the relationship between bodies and minds and the nature of our knowledge of other minds, thus completing an investigation of the three great problems of the philosophy of mind as they arise within the context of AI.

The conception that this work is intended to convey, therefore, is that the scope of AI should not be limited by concern for the problem of replication. Even if digital machines are restricted to the utilization of artificial languages that may be significant for the users of those systems but not for use by those systems themselves, the contributions that AI products can extend to the enhancement of human existence are very important, indeed. Not the least of the benefits that this book ought to provide, I should add, is a more systematic appreciation for the influence which presuppositions – especially with respect to methodology – can exert upon our theoretical investigations. Thus, it ought to be emphasized that the conception of mentality advocated here is strongly non-behavioristic, non-extensional, and non-reductionistic, where these notions are to be understood as the following passages explain.

Behaviorism. Various conceptions concerning the character of scientific language have pivoted about the empirical testability or the observational accessibility of hypotheses and theories. A distinction between three kinds of non-logical terms has often been drawn, where predicates that designate (or "stand for") properties are classified on the basis of whether the properties they stand for are observational, dispositional, or theoretical in kind. Richard Rudner (1966), for example, has proposed this classification scheme:

(D1) observational predicates =df predicates that stand for observable properties of observable entities;

(D2) dispositional predicates =df predicates that stand for unobservable properties of observable entities; and,

(D3) theoretical predicates =df predicates that stand for unobservable properties of unobservable entities;

where every predicate should fall into one and only one of these categories.

But a long-standing debate has raged over the merits of predicates making reference to unobservable properties within the context of empirical inquiries. Practically all parties to this dispute tend to acknowledge the fundamental function fulfilled by observational language, which is used to describe the contents of (more or less) direct experience. The problem thus arises over the status of unobservable properties, where much of the war has been waged over the scientific standing and scientific significance of dispositional predicates, which are intended to ascribe to things habits or tendencies to display specific kinds of behavior under specific conditions. The participants in this debate have included Gilbert Ryle in philosophy and B. F. Skinner in psychology, but its origin traces back to David Hume. [See, for example, Carnap (1936-37), Ryle (1949), and Skinner (1953).]

Since dispositions are tendencies to display specific kinds of behavior under specific conditions, it looks as though it might be possible to settle the debate over the status of dispositional language, at least, by establishing that these unobservable properties are explicitly definable by means of observational predicates and truth-functional logical connectives, such as the "if ___ then . . ." conditional understood as the *material* conditional. The predicate "x is hungry" can then be defined in the following fashion:

(D4) x is hungry at time t =df if x is given food at time t, then x eats that food by time t^*;

where t^* equals t plus some suitable (specific) interval, permitting time for that behavior to be displayed. The promise of this approach in offering a satisfactory solution to these problems of meaning and significance was so great that it became a cornerstone for the position that is known as *behaviorism,* which endeavors to eliminate appeals to unobservables by strict adherence to this methodology [see, for example, Block (1980)].

The imposition of these rigid epistemic requirements upon scientific definitions has proven highly problematical, not least of all because the material conditional, although well-defined, does not possess the logical

properties appropriate to dispositional predicates. In particular, since a material conditional is true when either its antecedent is false or its consequent is true, when "x is hungry" is defined by (D4), its definiens will be satisfied by anything x at any time t that it is not being given food as well as by anything that is given food at that time and eats that food during the specified interval. If a leather chair or a delicate antique does not happen to be given food at a certain time t, then it turns out to be hungry at that time, as a consequence of the form adopted for proper definitions.

The problem for leather chairs and delicate antiques, however, can be disposed of by restricting the class of kinds of things that might or might not be hungry to animate things, so long as their possession of this property could be ascertained on the basis of experiential grounds alone. For the underlying rationale for behaviorism's emphasis on *stimuli* and *responses* arises from their construction as publicly observable antecedents and as publicly observable consequents. But the problem turned out to be even-more difficult than that, since there appear to be other factors besides its availability that influence whether or not a human being, for example, is disposed to eat when presented with food. If a person were participating in a religious fast, scrupulously adhering to a rigid diet, or otherwise morally inhibited from consuming food, he might very well not eat food when it was presented, even though he was indeed hungry. To cope with problems of this kind, however, it would be necessary to assume the existence of other unobservable properties of x, which tends to defeat the program.

Extensionality. There were skillful efforts to overcome these difficulties, some of the most ingenious of which were undertaken by Rudolf Carnap in (1936-37). Yet even Carnap's ingenuity was not enough to salvage the behaviorist platform, for at least two different reasons. The first has already been implied above, namely: that the causal factors that tend to influence human behavior cannot be reduced to those that are accessible to observation in the form of publicly observable stimuli and publicly observable responses. If we tend to accept ordinary "folk" psychology, a more adequate inventory of the inner states of human beings would include not only motives and beliefs but also ethics, abilities, capabilities, and opportunities. But the complex causal interaction of factors of these kinds renders their presence or absence subject to indirect modes of inference rather than to direct observation. The epistemic resources that behaviorism embraced were inadequate to cope with the complexities of the subject of behavior.

Even more striking than the limitations that behaviorism imposed on the non-logical terms that it wanted to employ were those discovered to attend its logical resources, in particular. Coping with what would happen if x were to be given food (whether or not x ever is) and with what would have happened had x been given food (when x had not) requires the use of *subjunctive* conditionals and *counterfactual* conditionals – where the latter are subjunctives with false antecedents – in intensional logic. Even Carnap eventually arrived at the conclusion that the definition of dispositional predicates could only be accomplished by reaching beyond the resources of material conditionals and of extensional logic. Thus, the behaviorist program for the elimination of theoretical and dispositional language could not be sustained. [For an illuminating survey, see Hempel (1965), pp. 101-122.]

In its full dimensions, therefore, the behaviorist position encountered three distinct problems. These can be formalized by utilizing '$\_\_\_\rightarrow \ldots$' as the material-conditional sign and '$\_\_\_\Rightarrow \ldots$' as the subjunctive-conditional sign. Then let "Sxt" stand for x's exposure to stimulus S at time t, "Rxt^*" stand for x's display of response R at time t^*, and "$F1, F2, \ldots, Fm$" stand for other properties (which are not necessarily directly accessible). For defining dispositional predicates, (1) "$Sxt \rightarrow Rxt^*$" is logically flawed because of the limitations of the material conditional; (2) "$Sxt \Rightarrow Rxt^*$" is empirically flawed because of the restriction to observable causal factors; yet, (3) "$(Sxt \,\&\, F1xt \,\&\, \ldots \,\&\, Fmxt) \Rightarrow Rxt^*$", which at least appears to be headed in the right direction, is not compatible with the behaviorist program. Even the subjunctive turns out to not be strong enough to fulfill its intended role and ultimately requires replacement by a causal conditional. [For more discussion of this and related issues, see Fetzer (1978), (1985a).]

What often tends to be overlooked about the material conditional is that its use presumes no special kind of connection between its antecedent and its consequent. They do not have to be related by virtue of meaning, or of causality, or of lawfulness, or of any consideration other than merely the *truth values* of those sentences alone. Hence, whenever the antecedent of a material conditional happens to be false, that entire conditional will be true no matter how unrelated or disconnected the subjects of its antecedent and its consequent may be: "If nine were even, Reagan would be President" and "If nine were even, Reagan would not be President" are both trivially true if interpreted as material conditionals, because both their antecedents are false!

While logicians have long since grown accustomed to this result, many of those laboring in other fields find this mystifying, with suitable justification. Ordinary language, educated intuition, and common sense are violated by a failure to distinguish between conditionals with false antecedents that are true and conditionals with false antecedents that are false. But extensional logic does not permit them to be distinguished, since *all* material conditionals with false antecedents are true. The need for stronger modes of conditionality should therefore be apparent, where the resources provided by *intensional* (or "non-truth-functional") logic far exceed those of *extensional* (or "truth-functional") logic, where the use of a subjunctive conditional, for example, entails but is not entailed by a corresponding material conditional. Indeed, the only case for which the truth value of a subjunctive is determined by the truth value of the correponding material conditional is when that material conditional happens to be false [cf. Fetzer and Nute (1979), (1980)].

The use of material conditionals, moreover, implies that an outcome response will occur whenever the input stimulus occurs, which will not happen when the relationship between them is one of probability instead. A behavioristic approach that accommodates probabilistic relationships by invoking the notion of "probability of response", therefore, has been elaborated in the work of B. F. Skinner [(1953), (1972)]. From Skinner's point of view, a disposition like hunger should be characterized, not as an invariable tendency to eat whenever food is presented, but as a high probability:

(D5) x is hungry at time t $=$df the probability that x will eat food by time t^* when given food at t is high.

Since "probability" must be interpreted as an extensional relative frequency rather than as an intensional causal propensity in order to preserve its "scientific respectability", however, this account fall prey to the same problems that undermine material conditionals, a point to which we return in Part II.

Reductionism. These distinctions, as I have hinted above, are meant to provide an introduction to some issues that lie just beneath the surface of the problems that are the principal subjects of our discussion in the following. For I shall argue that the wrong conclusions have sometimes been embraced because of mistaken commitments to untenable methodologies, especially those of behaviorism and of extensionality. The point, there-

fore, is to provide enough background to appreciate the reasons why these methodologies are open to dispute without attempting to exhaust the problems that they raise, while suggesting resources for further exploration. Thus, the last of these that I want to mention here is also the most general, for both of the others can be viewed as special cases of the program known as *reductionism*.

My reason for saying so is that, when behaviorism as it has been defined functions as a standard of acceptability for research in psychology, it serves as an attempt to reduce the unobservable to the observable within the field of psychology. And when extensionality as it has been defined functions as a standard of acceptability for scientific language, it serves as an attempt to reduce the intensional to the extensional within all fields of science. For, in a similar fashion, reductionism can be viewed as any effort to reduce what is complex to what is simple, largely driven by the spirit of Occam's Razor, a methodological maxim asserting that entities should not be multiplied beyond necessity: "It is in vain to do by many what could be done by fewer!" [On Occam's Razor and its significance, cf. Smart (1984) and Fetzer (1984).]

Occam's Razor thus suggests that we ought to prefer simpler over more complex theories, which provides leverage in moving toward less complex theories. But reductionism, thus understood, suffers from one critical flaw. For surely we must prefer simpler to complex theories only when they are adequate. Behaviorism, after all, provides a simpler account of the nature of psychological properties than its theoretical alternatives, but it remains hopelessly inadequate, nevertheless. Extensionality, likewise, provides a simpler account of the nature of scientific language than its theoretical alternatives, but it remains hopelessly inadequate, nevertheless. Reductionism is an appealing conception, but it is not necessarily always justifiable.

If we want to define dispositional predicates like "hunger", which are properties of observable entities such as human beings, the appropriate place to begin would be with material conditionals. Once we have discovered the consequence that people turn out to be hungry merely because they are not being subjected to the appropriate test, however, it becomes imperative to consider alternative constructions. And if there simply is no simple relationship between stimulus and response for human beings because they are complex causal systems whose behavior is affected by inner states that are not amenable to direct observation, then we ought to embrace a more sophisticated methodology that is equal to the task that we

would impose upon it. And if there should be no simple way to relate the mental to the physical, that is something that we shall have to accept.

Perhaps a more appropriate conception of Occam's advice would be to construe his concern as a drive for "elegance", in the sense of striving to derive maximal sets of consequences from minimal sets of assumptions in order to optimize, say, explanatory power in science or theoretical significance in philosophy. For, if this were the case, then simpler theories would only be preferred theories when they provide an acceptable account of the phenomena with which they are intended to deal. When matters are complex, Occam's Razor would no longer have us embrace simple theories, but only those that are as simple as they can be while remaining adequate to cope with the demands of a complex world. The problems we confront in dealing with the scope and limits of AI are complex. Perhaps in place of simple theories we ought to be looking for elegant theories instead.