Artificial Minds: Fact or Fantasy?

David Davenport

Bilkent University
Department of Computer Eng. & Information Science
Ankara 06533 - TURKEY

Technical Report BU-CEIS-93-17

Artificial Minds: Fact or Fantasy?

David Davenport
Computer Eng. & Information Sciences Dept.,
Bilkent University,
Ankara 06533 Turkey

email: david@bilkent.edu.tr
DAVID@TRBILUN

Abstract: Trying to fathom the inner workings of the human mind has been one of the main preoccupation's of philosophy and psychology. The advent of the electronic digital computer provided both a new tool and a new perspective for this quest, and spawned the research field known today as artificial intelligence (A.I.). Yet, is an artificial intelligence a real possibility, or is it simply a myth, unreachable in principle? Would an A.I. be conscious, would it literally be an artificial mind, or is there something intensely human about minds which preclude our constructing one? Despite years of research, we still have no real answer to these most basic of questions. Speculation and controversy abound, rekindled of late by the resurgence of the connectionist paradigm. This paper examines some of the key arguments in the debate and suggests a new theory based on "inscriptors" may offer plausible solutions.

Keywords: Artificial Intelligence, Chinese room experiment, Turing test, symbol grounding, Inscriptors.

Introduction

Philosophers have long pondered over the inner workings of the human mind; just how is it possible that each of us can come to view and understand the world around us? Lacking the necessary conceptual apparatus, early attempts to answer this question proved rather unsatisfactory. On the Dualist view, for example, "the mind... is composed not of physical material obeying physical laws but of soul-stuff, or 'spooky' stuff, and it operates according to principles unique to spooky stuff" (Churchland & Sejnowski, 1989). In Plato's time, taking mental phenomena (experiences, beliefs, etc.) to be non-physical, was perhaps natural, afterall, what physical entities appear, combine and disappear with such ease. Later Materialist philosophers insisted that the functioning of the mind must be due to purely physical phenomena. However, it is only recently, with

developments in formal logic and mathematics that the necessary conceptual apparatus have become available. Moreover, the advent of electronics, and of the electronic digital computer, have provided the technical means which may make it possible to actually construct an artificial intelligence. The stage appears set, "real" answers might now be forthcoming.

Of Symbols

Early digital computers were frequently referred to as "electronic brains" (Martin, 1993), so it was quite natural for people to ask whether such machines really were intelligent or whether they were merely imitations of the genuine thing. Can machines REALLY think? Turing took up this question in his classic 1950 paper (Turing, 1950). He proposed the, now (in)famous, Turing test. In effect he side-stepped the real question and suggested instead that if we could not tell the difference in responses between a human and a machine, then (presumably) the machine was indeed "thinking". The test (game) was cleverly designed so that "irrelevant" factors, such as appearance, physical abilities, accuracy, speed etc., could be masked as far as possible. For this reason, the participants are kept in separate rooms and all communication is done via teletypes. Furthermore, the machine is allowed to "lie" in order to convince the investigator that it is indeed human. Is there any reason to suspect that some (future) computer could not pass this test? Turing showed that the more obvious arguments which might be advanced to show that no computer could ever pass the test, were at least suspect, if not outright mistaken. Interestingly, the only real difficulty which Turing foresaw was the possibility that extra sensory perception, ESP, (the evidence for which he saw as "overwhelming") might upset the balance. This potential problem could presumably be alleviated by placing the competitors in "telepathy-proof" rooms. (However, in such a case there remains an uneasy feeling that there might actually be more to thinking, and minds, than meets-the-eye!)

In fact, independent of the potential difficulties posed by ESP, there are two real problems with Turing's test. The first, as we shall see again later, is that it ignores completely the physical faculties, e.g. vision, touch, motion, etc., which enable us to interact with, and recognise and discriminate things in the world, such that any agent lacking linguistic abilities (including animals and perfectly intelligent but mute people) would be excluded from the test. This may be interpreted, incorrectly, as meaning that speechless agents have no ability to think, whereas in fact it simply indicates the need for a more refined test, similar perhaps, to Harnad's Total Turing Test (Harnad, 1991). The second problem with Turing's test is that it is blatantly behavioural. The behavioural stance is unfortunate partly because it tells us nothing about what thinking is or how thinking entities function, and partly because it leaves the way open for fakes. There is a possibility, admittedly remote due to the open ended nature of the

test, that it may be passed by a machine which quite clearly is not "thinking". An indication of this came with Weizenbaum's 1965 ELIZA program (Weizenbaum, 1965), which reputedly fooled several people into believing that they were indeed talking to a human psychologist.

Newell, Shaw and Simon's 'Logic Theorist' and their later 'General Problem Solver' (Newell & Simon, 1963), were among the first attempts to go beyond the purely behaviourist vision and see what the internal mechanisms of thinking might look like in practice. Winograd's 1972 SHRDLU (Winograd, 1973) program, applied these ideas to demonstrate natural language understanding in the context of a 'toy blocks world'. A pictorial representation of the state of SHRDLU's world was displayed on a monitor, which was updated in response to English sentences typed on the keyboard. Commands such as 'Put the pyramid on the blue box' would be acted upon if "understood", or would produce such responses as 'Sorry, I don't know which pyramid you mean' or 'The pyramid is already on the blue box'. This was indeed an impressive demonstration and increased expectations that truly intelligent thinking machines were here, or at least just-around-the-corner.

But this was not the case. Not only did such programs prove impossible to scale up, but a philosophical argument due to Searle showed that however sophisticated they may become, they would still never really "understand" anything at all. Searle's thought experiment, commonly known as the 'Chinese Room', (Searle, 1980) consists of a room inside of which is a non-Chinese speaking man, say Searle himself. It also contains paper, a pencil and a rule book. Native Chinese speakers converse with the room by passing messages written in Chinese through a slot in the wall. Searle-in-the-room examines the messages and, using the rule book, prepares a written response which he passes back through the same slot. The rule book simply says things like, "If the message is squiggle squiggle then output squoggle squoggle". Now, even assuming that we could actually produce a rule book sophisticated enough to enable Searle to fool the natives into believing that he understood Chinese, it is (presumably) quite clear that Searle himself still does not understand any Chinese. But Searle has simply put himself in the place of a computer which is assumed able to pass the Turing test in Chinese, so if Searle doesn't understand Chinese, then neither would the computer!

This argument hit deep at the heart of one of AI's most basic tenets, the physical symbol system (PSS) hypothesis which states that, "A physical-symbol system has the necessary and sufficient means for general intelligence" (Newell & Simon, 1979). A physical symbol system is just an implementation of a symbol system, of something which manipulates formal symbols purely on the basis of their form (syntax), not their meaning. A computer is just such a symbol system, but if, as Searle argues, it doesn't understand anything, then presumably the PSS hypothesis must be wrong.

There are a number of frequently heard replies to Searle's argument. Among them:

> The 'Systems Reply': OK, so the man in the room doesn't understand Chinese, but surely the system as a whole including the room, the man, the rule book and the paper and pencil, does. Searle's response to this is no. Imagine that the man memorises the rule book and hence dispense with the room, book etc. Now the man is the system, but he still doesn't understand Chinese.
>
> The 'Brain Simulator Reply': If we wrote a program that simulated the actual sequence of neural firings at the synapses of the brain of a native Chinese speaker, then surely it could be said to understand Chinese. Once again Searle's response is a resounding no: simulation is not duplication; ".. you could not digest pizza by running a program that simulates... digestion" (Searle, 1990). According to Searle, what distinguishes the brain, are its causal properties and its ability to produce intentional states. And these, he claims, are characteristically biological and hence not reproducible in silicon, or any other material.
>
> The 'Robot Reply': Suppose the computer were placed inside a robot and that it received some of its symbols from a television camera mounted on its head and some of its output symbols were fed to motors that caused it to act, to move around, pick things up, etc. Surely the robot could be said to understand. But again Searle responds negatively; how could it since, as with the room, Searle could take the place of the computer and would obviously have no idea what the symbols he was shuffling around, actually meant. Furthermore, since all of the robot's actions derive from its electrical wiring, while Searle-in-the-robot's actions result directly from his following the program, there is no intentionality in either of them.

Searle seemingly has all avenues of escape covered. Computers simply cannot possess understanding, hence they cannot be truly intelligent, therefore no computer, however cleverly programmed, would actually be a mind; A.I. is impossible, dead!

Well, maybe not quite. Searle's argument is actually directed against purely formal symbol manipulating systems, systems which have only syntax and no semantics. His initial response to the robot reply was to claim victory on just these grounds, since a robot is quite clearly more than just a symbol manipulator. Moreover, several authors have questioned the validity of Searle's main response to the robot reply. Boden (Boden, 1990), for example, points out that intentional states, such as understanding, are properties accredited not to brains, but to people. It is precisely in order to explain how such capacities arise, that simpler processes which lack them are postulated. Harnad too, takes up this issue (Harnad, 1993), pointing out that while Searle can put himself in place of

the computer and say he does not understand, he cannot logically do the same with the transducers responsible for the robots vision, hearing, etc. If he were to attempt to internalise these transducers too, as he did with the rule book, he would have to use his own eyes and ears, but then he would become an integral part of the system and the system he is already has considerable "understanding". It is thus unclear whether Searle-in-the-robot could legitimately claim not to understand Chinese. Indeed, some authors (e.g. Hauser, 1993) have even questioned whether Searle-in-the-room could properly claim not to understand Chinese, particularly after memorising the rule-book. As an indication of why this may not be quite so straightforward a pronouncement, consider the case of a tourist who manages to memorise a Chinese phrase-book and can thus utter an appropriate sentence in (almost) any given circumstance. Assuming the phrase-book was extensive enough, it would appear as if the tourist was actually conversing in, and to all extents and purposes understanding, Chinese. Moreover, if asked what they were doing, the tourist may quite well reply that he/she was talking in Chinese!

The difficulty which Searle's Chinese Room argument does highlight, is that computers are specifically designed to manipulate symbols without regard to what they mean, i.e. what they actually relate to in the "real" world, not simply other symbols in the machine. But true understanding most surely requires meaning. The situation resembles the problem of obtaining the meaning of a word from a (foreign language) dictionary. Each word is explained in terms of other (foreign) words, which are explained in terms of still others and so on, such that unless an external agent provides an explanation of the most primitive words, no meaning can be extracted from the dictionary. So, despite the fact that the symbols being manipulated may be capable of a systematic interpretation, there is still no understanding since this interpretation is not available to the machine. To achieve this, symbols must somehow be "grounded" in reality, and this, of course, is precisely the function that transducers perform. Thus, far from being mere extras to be added onto an intelligent computing device, transducers, both receptors and effectors, must be considered an integral part of the system, for without them there would be no understanding, no intelligence.

As noted previously, the Turing test has often been criticised precisely on the grounds that it ignores the human faculties now seen to be a vital ingredient of intelligence. It was in recognition of this, that Harnad proposed the 'Total Turing Test' (Harnad, 1991), which requires a machine to be indistinguishable from human beings, not only in their "symbolic" capacities, but also in their "robotic" capacities. Robotic here does not necessarily mean having arms and legs, but rather an ability to recognise and differentiate objects in the real world. Note that the TTT is not actually stronger than the TT since, if the above argument is correct, linguistic abilities must be founded on robotic ones. Instead it provides a clear indication that real understanding is based firmly on robotic capabilities involving transducers. Unfortunately, exactly how such a machine might be implemented and how it might function, is still something of a mystery. Simply

appending transducers to a computer is obviously not enough (otherwise SHRDLU might easily have been an A.I.). There must, presumably, be some principled method of integrating them. Harnad suggests a hybrid solution, employing a neural network between the transducers and the symbol system; the network enabling the robot to learn the invariant categories in its analogue input and then assign them a ground-level symbol. Before examining this proposal, it is perhaps worth pausing to ask whether a connectionist system (e.g. a neural network) could not solve the problem on its own, without the need for any symbol system.

Of Connections

Connectionism is an attempt to model cognition, bottom up, by copying the functional structure of the brain. This structure is presumed to consist, in essence, of a multitude of simple, independent, yet highly interconnected, processing units. Such units are called neurons and, for this reason, connectionism has become synonymous with the term artificial neural networks. Neurons have a large number of inputs, but only a single output. The output of one neuron is connected to, and hence forms the input of, many other neurons. Networks are usually considered to have no feedback loops and hence tend to form some sort of hierarchical structure. The processing which a neuron performs is generally considered to be very elementary. A neuron's state, its output, is usually modeled as a non-linear function of its inputs. A typical scheme would involve applying a threshold function to the weighted sum of the inputs. Changing the interconnection weights thus affects which specific sets of neurons respond to which input patterns. Learning then, consists in finding an appropriate set of weights, such that the desired mapping is achieved. There are a number of algorithms which allow for the automatic learning of such sets of weights, directly from example inputs.

The aim of connectionism is to demonstrate that such 'parallel distributed processing' as is afforded by artificial neural networks, is in fact a sufficient basis for cognition. If this is true then the brain, contrary to the classical stance, need not be a PSS. Fodor and Pylyshyn examined this claim in detail (Fodor & Pylyshyn, 1989). They concluded that neural nets, as presently conceived, were in fact not powerful enough to provide a basis for cognition, although they went on to suggest that such nets might offer a partial implementation-level account of the mind's PSS. F&P's argument is based on the observed productivity, systematicity and compositionality of language, and on the apparent inferential coherence of thought. In order to produce such effects, they claim, an agent's internal representation must have semantic and syntactic constituent structure. F&P observe that neural networks do not possess such a representation and, moreover, that they implement an already discredited philosophy (associationalism), and thus cannot offer a suitable basis for cognition. Given

this argument, Harnad's proposal for a hybrid architecture seems reasonable, perhaps even essential.

But is this really the case? How can the shortcomings of the individual techniques be overcome simply by combining them? Answering this requires an appreciation of the underlying difference between the symbolic and connectionist paradigms. What really distinguishes them is the nature of symbols and the manner in which they combine (Davenport, 1993a). The symbolic paradigm is based on idea that signals coming from the environment can be shown to be mutually-exclusive. Symbols are thus considered to be like "objects" which can be moved and copied at will, composite symbols being formed by an encoding process, usually concatenation (the literal gluing together of copies of the component symbols). In contrast, the connectionist paradigm is founded on the conjunction of inputs from the environment. Connectionist symbols are correspondingly static, unmovable 'signals', which combine by being 'linked' (effectively and'ed) together. F&P argued that linking lacked the ability to retain potentially vital 'relational' information, as required, for example, to distinguish 'john loves mary' from 'mary loves john'. However, in principle, the structure of the linking can be used to handle these situations, just as in classical syntax diagrams. The catch, of course, is that a special "mechanism" is required to "decompose" such structure, since the net itself cannot do it. Obviously, it would be nice to be able to account for this process using the same linking technique, however, it should be noted that the syntactic method similarly demands extraneous mechanisms which are also presently unexplained! Assuming then, that both 'paradigms' are capable of representing the necessary knowledge, they are, in a sense, equivalent, and hence may be viewed merely as alternative means of implementation. But on this account, Harnad's proposal for a hybrid system is obviously flawed, combining techniques is not going to make any difference, at best it would simply remove the need for a symbol system but that would leave the neural net to do all the work, which, according to F&P, it still couldn't manage even given a solution to the representational structure problem.

Summary

To recap, Searle's Chinese Room argument has destroyed the hope that a symbol system alone could possibly be an artificial intelligence, while F&P have demonstrated that neural nets are also unsuitable, finally Harnad's proposal to combine the techniques has also been shown to fail. Is A.I. impossible then? Well, maybe not, there remains at least one further option. F&P dismissed neural nets on two grounds, representational structure and Associationalism. The former, as noted above, could be resolved given a suitable mechanism to "extract" the network structure. Can the philosophical difficulties be overcome too? The next section presents a new theory which suggests that they can be. It also appears to offer a suitable basis for an extraction mechanism.

Enter Inscriptors

The world, according to the inscriptor theory (Davenport, 1993b), is a vast, complex, chaotic place. Yet, for all its apparent randomness, certain "states-of-affairs" do recur in both space and time. An agent which could identify such repetitions might be able to predict other "world-states". It could then use this knowledge when selecting its actions, thus increasing its chances of success/survival. For example, it could utter the same sounds which had previously persuaded its mother to relieve the discomfort caused by a soiled nappy. It could first search locations similar to where it had previously found food. It could come to recognise and avoid undesirable, even potentially dangerous, situations. The better an agent was at predicting the world, the better its chances of survival (and presumably, the more intelligent we would consider it!)

An agent's task then, is to detect and store recurrent "states- of-affairs", and to later recognise an existing situation as appearing to match one of them, so as to use this knowledge to decide upon the next action. Since it is impossible to know which states will repeat, the best an agent can do is to remember whatever it can, giving priority eventually to those states which are actually observed to repeat in practice. In essence then, an agent merely has to remember each state that it observes. The basic unit of memory which accomplishes this is the inscriptor.

An inscriptor, like a neuron, has a large number of inputs and a single output. The output of one inscriptor can feed to the inputs of other inscriptors such that collections of inscriptors form a loose hierarchy with the system inputs at the lowest level. Each inscriptor has a "learning threshold" and if the signals incident on the inputs are sufficient in number to exceed this threshold, the inscriptor is said to 'fire'. This causes it to remember the combination of inputs which made it fire, and from then on to ignore signals on those connections which did not contribute to the firing.

The outcome is that an inscriptor records some observed input combination, some "state-of-affairs"; that is, whatever (logical) input connections it ends up possessing (after firing), signals have been concurrent on those inputs. The resulting (logical) network structure is thus a consequence of direct causal interaction with the world and is therefore also, in some sense, a 'correct' representation of it. While subsequent "situations" should be stored in a similar manner, they must also be matched against the knowledge already inherent in the network, to help discover recurrent states and hence to help select desirable courses of action. To achieve this, inscriptors which have fired respond to signals on their (logical) inputs by changing their "activation level" and by generating another signal on their output. Signals from the environment thus alter the state of various inscriptor nodes, such that there is a direct causal relationship between the world and the activation state of the network. To understand how this can help

'recognise and predict' consider a typical inscriptor network such as that depicted in figure 1.

... insert figure 1 about here ...

Signals incident on the inputs to the network provide "evidence" for the inscriptor nodes to which they are connected. In other words, an input signal is evidence that the same or similar set of inputs as caused the inscriptor to fire in the first place, are being repeated (i.e. the same situation or concept is again apparent to the agent's senses). Thus, in figure 1, input C is evidence for both concepts X and Y, while input A is evidence for concept X (but not Y) and input E is evidence for concept Y (but not X). Inscriptor outputs function in a 'winner-take-all' fashion, such that "evidence" is gradually redirected away from nodes with lower activation levels in favour of those with higher activation. Since this tends to make nodes with higher activation even more active, the result 'snowballs' towards the most likely conclusion, i.e. towards the node(s) for which there is most evidence. So, if, in figure 1, both A and C inputs were incident on the network, there would be more evidence for concept X than for concept Y, and thus X would "succeed", whereas if inputs C and E were incident there would be more evidence for concept Y than X, and it would win-out. Note that this outcome is reached in the absence of a signal on input D (or B in the former case), the network being able to draw "conclusions" even if the information it is given is incomplete. An inscriptor (node) will also receive evidence (and hence become "active"), even in the absence of any inputs, if some of the nodes to which its output is connected are themselves active. Consequently, if, in figure 1, the network has settled on concept Y in response to inputs C and E, node D would actually become active since it is connected to the active Y node. Furthermore, evidence from Y is passed on up to R and S, so that Z also sees its output connections active and so becomes active itself. The network thus displays 'input completion', or equivalently, raises 'expectations' of certain inputs. Note too, that this process is entirely symmetrical, in that evidence from input F will pass (via Z) to nodes R and S causing them to become slightly active, and that this, in turn, will cause node Y to receive evidence and hence become active (even in the absence of inputs on C, D and E).

Observing the state of the network once it has settled on a solution, shows that it could be described as possessing very logical properties, e.g. if A and B then X, if B and C and D then Y (actually it has been shown that a more suitable description would be, if X then A and B, if Y then B and C and D, see Davenport, 1992). These letters may even be replaced with suitable propositions, e.g. if barks and animal then dog, if animal and meows and claws then cat. It is easy to imagine that input signals could be derived from transducers and that the network would thus properly recognise things such as cats, dogs, or whatever (such a system would obviously be extremely complex, however, it is, in principle, possible). While we have achieved a certain degree of understanding here, we have still not yet achieved conversational/language understanding

since the labels are not actually part of the system, but merely a convenience for human interlopers.

Using language requires some further complication, the agent must be able to recognise and utter any desired word, and words heard must be related to objects seen and vice versa. Figure 2 illustrates how the previous network may be extended to achieve this.

.. insert figure 2 about here ...

There are now two sets of inputs, one originating from visual senses as before, another from the hearing senses. Both function in the same manner. Some nodes are connected to both input sets, and represent situations in which a word is heard and something is seen. Assuming that the net is constructed in such a way that the word is the one used to refer to the thing seen, we now have the required "(word) symbol grounding". Seeing something results in a particular set of visual inputs which are "processed" as described above, the net settling into a state such that the corresponding 'word' node, will be active, since its output is connected to an active node. Having identified the appropriate word, it is but 'a short leap and a jump' to actually utter it. Alternatively, if a sequence of sounds are heard, these too are "processed" and result in the recognition of a particular word. Again a 'side effect' of this, is that the related nodes in the visual hierarchy are activated producing a corresponding mental image in the "mind's eye".

Language proper must be built on top of the foundation offered by this ability to relate word to object. The processes involved are the same but this time require linkage between abstract situations and abstract sentence patterns (as opposed to abstract word and abstract object). Having selected an appropriate sentence form for the situation to be described, utter it word by word, each time selecting the most suitable word based on the pattern and the actual situation.

As an illustration that the combination of two hierarchies provides the required "understanding/grounding", consider a robot with only hearing senses and which only ever hears people speaking, (or equivalently imagine yourself - blind - surrounded by people who only speak in a foreign tongue, or perhaps just listening to foreign language radio broadcasts). The robot (you) would gradually come to recognise syllables and words and eventually sentence patterns (extended sequences of words which differ by only a word or so in certain places). It (you) would probably be able to offer a list of words which could potentially be used in specific places within a given sentence pattern. Most of these candidate words would be ones which had actually been heard in sentences which matched the pattern, although it is possible that others may be suggested by comparison with the lists of words in other sentence patterns. In this way a (partial) syntax of the language could gradually be built up, but without the slightest idea of what any sentence or word actually meant, i.e. actually referred to in the real world. The addition of, and linkage to, another hierarchy

which somehow encodes spatial and other knowledge of the world, can give "meaning" to this. The second hierarchy determines groups of features, objects, situations, etc. based on similarities which actually occur in the real world. Such groupings are 'linked' to a word-symbol in the first hierarchy, such that even though the agent may not have heard a sentence containing a particular word, it can generate/understand a sentence with that word since it is in the same grouping with (semantically) 'similar' words/concepts from which it originally created the pattern.

Harnad's explanation concerning the grounding of unseen concepts is also clearly valid in this scheme. An agent which knows (has nodes in both the visual and spoken word hierarchies for) 'horse' and 'striped', and which is told that a 'zebra is a striped horse', can successfully imagine, and thus identify, a 'zebra' when it first encounters one. Similarly with the more fanciful peekaboo unicorn, a horse with a horn which disappears the moment you attempt to look at it (and hence is unseeable in principle) and so, in a similar vein, with abstract concepts such as love, barter, freedom, etc. Lastly, notice that the second hierarchy need not result from audio senses, but could equally well come from touch or even vision (as evidenced by blind people who learn to touch read Braille and deaf people who can read/sign).

In this way then, inscriptors overcome the major difficulties besetting both the symbolic and connectionist paradigms. Being based on a connectionist, linked, architecture, they avoid the difficult (impossible) questions concerning where symbols come from, how to decide which set of symbols to employ and how to correctly manipulate them. By not employing interconnection weights inscriptors are able to evade the philosophical problems facing neural networks. Interconnection weights are the embodiment of the Associationalist philosophy which suggested that one concept was related to another with some fixed "degree". The inscriptor theory recognises that the relationship between concepts is not fixed but varies with context. By accumulating evidence only from those situations that match the specified context inscriptors effectively compute the relevancy of any concept as and when needed. Notice that while one may view the firing process as setting binary weights, these are not used in subsequent computations as they are in neural networks. Finally, by providing a clear distinction between internal (signal) symbols and external (word) symbols, and by employing two hierarchies (one for spatio-temporal information, the other for the matching external symbols) the inscriptor theory suggests a solution to the problem of constituent representational structure.

Of Intelligence and Minds

Inscriptors thus seem to be able to endow a system with "understanding", but does this make it intelligent? What is intelligence? I offer the following definition;

> Intelligence is the ability of an agent to detect, store and subsequently use for its own advantage, the regularities which exist in an ever changing world. A truly intelligent agent may also be expected to exhibit creativity whereby it fortuitously combines circumstances which are not otherwise naturally related.

With the proviso that an agent have a goal, or set of goals (intentional states) so that it does not remain a purely passive entity, inscriptors can thus seemingly provide the abilities necessary for intelligent behaviour (of course, this is cheating somewhat, since the definition, although quite reasonable, is obviously framed to suit the needs of the paper!)

So an inscriptor-based agent seems capable of understanding and of intelligent behaviour. Most certainly it thus provides a reasonably good MODEL of the mind, but could it actually BE a mind? Presumably this would depend a lot on the meaning attached to the word "mind". Most people would probably balk at the conception of an artificial (man-made) mind, either on purely religious grounds or because, to date, all minds have been natural so that the idea of an artificial one seems somehow contradictory. Even allowing for this it is not certain whether such a system would count as a mind, since exactly what the necessary/requisite properties are is far from certain. Whether attributes such as consciousness, causal/intentional powers, emotions etc. are essential to, irrelevant to or a by-product of minds is unclear.

The concept of self (of self awareness, of consciousness) may, for example, merely be the label we attach to the internal model we each have of ourselves (our bodies). Such a model is a pre-requisite for any sort of intelligent action, since it is vital to be able to predict the consequences of any action before carrying it out, in order to avoid potentially disastrous results. Thus, by definition (cheating again) there is every reason to suppose that a machine could, indeed should, be conscious. In a similar vein, just as direct physical pain is part of the control mechanism which helps us develop such a model by teaching us to avoid immediate personal injury, so the more ethereal emotions/feelings such as love/hate/jealousy etc. may be learnt to help describe and control social interactions!

As regards causal and intentional properties, Searle has long argued that a 'computer' could not be a mind precisely because it has the wrong causal and intentional properties. Searle's insistence on "causal" powers being uniquely biological is understandable, all he is saying (I believe) is that much/most/almost-all of our existence, our very being, is bound to our biology. Things like hunger, thirst etc. are irrelevant to anything non-biological, hence such an agent would have no use for, and hence no real understanding of, such concepts. It may, of course, acquire a behavioural conception of them by observing human beings, but this is a far cry from "living" them. As an analogy consider the case of men and women; men have no (and never can have any) real understanding of what

it feels like to be a woman, of "womanhood", of giving birth, since these are uniquely tied to biological factors and the wrong biology at that. In an exactly similar vein Collins (Collins, 1987) has argued that a machine must "share a culture" in order to exhibit the understanding necessary for social interaction. Intentional states too, were founded largely on biological needs, the desire to survive, to reproduce, to eat, etc., although these are now often translated into social pressures such as getting money, promotion, marriage, children, etc.

On the other hand, Searle also claims that the 'type' of causal powers embodied in the electronic computer are wrong. He points out that simulation is not duplication. This is correct, but the argument in this case is wrong because, in essence, the brain's input and output is of the same "casual type" as the simulation, i.e. electrical signals. Thus, just as a mechanical simulation of a heart could pump blood while a computer simulation of it would not, so too an electronic simulation of a brain could cause muscles to move while a mechanical simulation of it could not. In other words, just as a persons real heart could be replaced with a mechanical one, so too their brain could be replaced with an electrical machine (suitably programmed)!

Concluding Remarks

The inscriptor theory appears to present a plausible alternative to both the classical symbolic and connectionist paradigms. It avoids the problems posed by Searle's Chinese room argument since it is not a symbol system per se. It could presumably form the basis of a virtual symbol system though, since it offers a means by which (external, word) symbols can be grounded, i.e. by which they can attain meaning. It also overcomes the difficulties faced by neural nets (as pointed out by F&P) by adopting a more realistic view of the world. While it retains a connectionist type, linked, architecture it is not affected by Searle's Chinese Gym argument (Searle, 1990), indeed it seems to show the fallacy of it by demonstrating the clear distinction between internal and external symbols. Of course, there are still several pieces of the puzzle left to put in place, in particular exactly how an agent can come to understand and initiate actions such as uttering words. But these are relatively minor details, for, if the inscriptor theory is right, an artificial thinking, understanding, intelligent, even conscious, emotional, mechanism, appears to be a real theoretical possibility (although in practice constructing one with human level capabilities may prove impossible and pointless!).

But then again, all these ideas may be partly or even wholly wrong. There may be much more to the mind than the superficial observations presented here, as Turing hinted, it really might be intimately bound to a spiritual world far beyond our materialistic reach, so that an artificial mind really is a misnomer.

Acknowledgments: The author wishes to thank Varol Akman, Mujdat Pakkan and Aaron Sloman for their helpful comments on the ideas in, and earlier drafts of, this paper.

References

Boden, M. (1990), "Escaping from the Chinese Room", in *The Philosophy of Artificial Intelligence*, Boden, M., Oxford University Press.

Churchland, P.S. & Sejnowski, T.J. (1989), "Neural representation and Neural Computation", in *Neural Connections and Mental Computation*, MIT.

Collins, H.M. (1987), "Expert Systems, Artificial Intelligence and the behavioural co-ordinates of Skill", in Brian P. Bloomfield, ed., *The Question of Artificial Intelligence*, Croom Helm Ltd., Kent.

Davenport, D. (1992), "Intelligent Systems: the weakest link?", in Kaynak, O., Honderd, G. & Grant, E., eds., *Intelligent Systems: Safety, Reliability and Maintainability Issues*, Springer Verlag, 1993

Davenport, D. (1993a), "Cognitive Architecture and Symbols of the Mind", Bilkent University Technical Report CIS9307

Davenport, D. (1993b), "Inscriptors: Knowledge Representation for Cognition", in *Proceedings of the International Symposium on Computer and Information Science-8*, Istanbul, 1-3 Nov. 1993 (also available as Bilkent University Technical Report CIS93xx)

Fodor, J. and Pylyshyn, Z. (1989), "Connectionism and cognitive architectures: A critical analysis" , in Pinker S. and Mehler J., eds., *Connections and Symbols* (A special issue of the journal Cognition), Bradford Books, MIT Press 1989, 1990

Harnad, S. (1991), "Other Bodies, Other Minds: A machine incarnation of an old philosophical problem", in *Minds and Machines 1*: pp.43-54.

Harnad, S. (1993), "Grounding Symbols in the Analog World with Neural Nets", in *Think* (Special issue on Machine Learning)

Hauser, L. (1993), Reaping the Whirlwind: Reply to Harnad's "Other Bodies, Other Minds", *Minds and Machines* 3, pp.219-237.

Martin, C.D. (1993), "The Myth of the Awesome Thinking Machine", *Communications of the ACM*, Vol.36 No.4, pp.120-133, April.

Newell, A. and Simon, H.A. (1963), 'GPS - A Program that Simulates Human Thought' in Feigenbaum, E.A. and Feldman, J.A., eds, *Computers and Thought*, pp.279-96, New York: McGraw- Hill.

Newell, A. and Simon, H.A. (1979), "Computer Science as Empirical Enquiry", , reprinted in *The Philosophy of Artificial Intelligence*, Boden, M., Oxford University Press, 1990. (ideas originally outlined in Newell, A. (1979) 'Physical Symbol Systems', Lecture at the La Jolla Conference on Cognitive Science. later published in Cognitive Science 4 (1980):135-83)

Searle, J. (1980) "Minds, Brains and Programs", reprinted in Boden, M., ed., *The Philosophy of Artificial Intelligence*, Oxford University Press, 1990.

Searle, J. (1990), "Is the Brain's Mind a Computer Program?", *Scientific American*, January.

Turing, A.M. (1950), "Computing Machinery and Intelligence", reprinted in Boden, M, ed., *The Philosophy of Artificial Intelligence*, Oxford University Press, 1990.

Weizenbaum, J. (1965), 'ELIZA - A Computer Program for the study of Natural Language Communication Between Man and Machine', *Communications of the ACM,* Vol. 9, pp.36-45

Winograd, T. (1973), 'A Procedural Model of Language Understanding', in Schank, R.C. and Colby, K.M., eds., *Computer Models of Thought and Language*, p152-86, San Francisco: W.H. Freeman.

figure 1

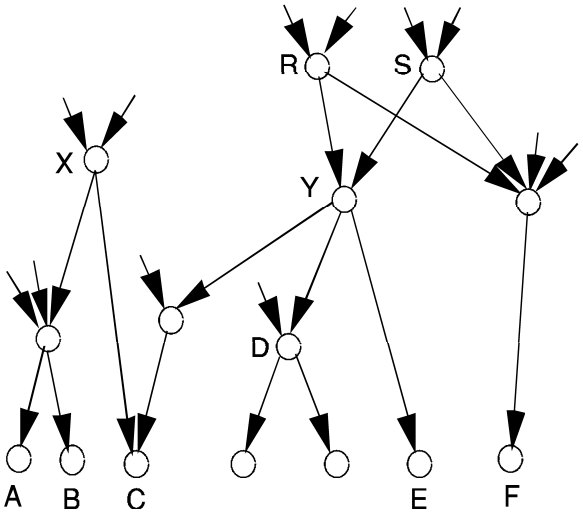

figure 2