The Two (Computational) Faces of AI

David Davenport

Abstract. There is no doubt that AI research has made significant progress, both in helping us understand how the human mind works and in constructing ever more sophisticated machines. But, for all this, its conceptual foundations remain remarkably unclear and even unsound. In this paper, I take a fresh look, first at the context in which agents must function and so how they must act, and second, at how it is possible for agents to communicate, store and recognise (sensory) messages. This analysis allows a principled distinction to be drawn between the symbolic and connectionist paradigms, showing them to be genuine design alternatives. Further consideration of the connectionist approach seems to offer a number of interesting clues as to how the human brain—apparently of the connectionist ilk—might actually work its incredible magic.

1 Introduction

Artificial Intelligence (AI) is both a scientific endeavour that attempts to understand human cognition and an engineering discipline that tries to construct machines with human-like capabilities. Unfortunately, the lack of a solid conceptual foundation makes AI, "an engineering discipline built on an unfinished science" (Ginsberg, 1995). Although the subject matter of this endeavour—the human mind—has traditionally been the realm of philosophy, philosophy's main contribution may have been to demonstrate that common technical words including "symbol", "representation" and "computation", are more difficult to define than they appear. Not surprisingly then, throughout its short history—beginning with the Dartmouth conference in 1956—AI research has seen a lot of heated debates, many resulting from misunderstandings due to differing goals, backgrounds and terminologies. In the

David Davenport
Computer Engineering Dept., Bilkent University, Ankara 06800 – Turkey
e-mail: david@bilkent.edu.tr

V.C. Müller (Ed.): Philosophy and Theory of Artificial Intelligence, SAPERE 5, pp. 43–58.
springerlink.com

following paragraphs, I present my own attempt to understand and bring some semblance of order to the topic. I approach the problem as an engineering task and begin by analysing the difference between the classical symbolic & connectionist paradigms. Consideration of the functional requirements for cognition, including the environmental/evolutionary contexts in which agents find themselves, offers a basis for designs which are shown to be necessarily computational in nature. This also offers a principled way in which to clarify the relation between the symbolic and connectionist paradigms and helps to set the scene for understanding how agents, and in particular we humans, might acquire meaning, consciousness, feelings, etc.

2 Explaining Cognition

In the beginning there was symbolic AI, and it was good. Indeed, John Haugeland (1985) later christened it GOFAI (Good Old-Fashioned AI), to contrast it with the upstart connectionist approach that gained support in the 1980s following the publication of Rummelhart & McClelland's (1986) book on PDP (Parallel Distributed Processing). Connectionist networks (also known as Artificial Neural Networks, ANNs) were seen by many as a means to overcome the problems being faced by the symbolic paradigm.

The classical symbolic approach sees cognition as computation, exemplified by the digital (von Neumann) computer and, perhaps to a lesser extent, the Turing machine. It is usually viewed as rule-governed manipulation of formal symbols. It appears to be inherently sequential, with a centralised control mechanism. It is generally considered to be logical and transparent, that is, its inner workings can be expressed/understood in meaningful terms. The symbolic approach has proved markedly successful, for example, so called expert systems are able to perform complex tasks, such as medical diagnosis, planning and configuration at the level of, and sometimes even better than, human experts. On the other hand, they are difficult to program, brittle (a single error often causing complete failure), not inherently able to learn, and lack a biologically plausible mapping. This results in great difficulties when it comes to building systems that can, for example, navigate around rooms or interact using spoken language—both skills that children as young as five acquire with apparent ease.

The Connectionist (ANN, PDP) approach, then, appeared to offer solutions to precisely these difficulties. As a network of "neural-like" processing units, it is naturally parallel and, with no clear centralised control, interruptable. It can learn from examples without the need for explicit programming (though the results are often opaque; difficult/impossible for humans to interpret). It is also tolerant of errors and, most importantly, has a (reasonably) obvious mapping to the human brain (which is assumed to be a network of neuron cells at the relevant level of description). To contrast it with the symbolic approach, some connectionists referred to it as "sub-symbolic" processing, a reference to the idea that concepts in ANN's were seen as represented in distributed form and that processing often worked only on parts of that representation/symbol (Chalmers, 1992), whereas in the symbolic form the entire (atomic) representation was the object of processing. This was just one way in which proponents of the ANN approach tried to differentiate it from the classical paradigm. They fully expected connectionist networks to be able to explain all of cognition. However, while it proved reasonably successful at simple low-level learning tasks, it struggled to demonstrate similar success at the higher levels where the symbolic approach had long held sway. The proper solution was therefore unclear: was one of the paradigms correct—in which case, which one—, or was it the case that a hybrid solution was needed—the ANN providing the lower-level learning that generated the symbols to drive the symbolic level—, or were they actually genuine alternatives—both able to support full-blown cognition—, or was there some other fundamentally different alternative? Not surprisingly, discussions took on the tone of religious debate, each group believing their position was the only true answer, and trying time and again to prove it.

One of the most important criticisms of the connectionist approach came from Fodor and Pylyshyn (1988), hereafter F&P. They pointed out that neural networks (of the time) were purely feedforward and lacked any sort of ability to handle sequences of inputs. This led connectionists to develop recurrent ANNs, whereby some of the output (or hidden layer) neurons feed back to form part of the input vector, so providing "context" for the next set of sensory inputs. This is equivalent to adding feedback loops to combinatorial logic to obtain sequential machines, and so provides the necessary capability, but (I suggest) does not provide a good conceptual foundation upon which to build. Another point raised by F&P was that representations must be (a) in or out of structure, and (b) active or inactive. They argued that the distributed representations favoured by most connectionists failed to have the necessary structural characteristics (since they were simple vectors), but even if they did (as, for example, in the case of a parse tree for a sentence) then there was no way for the network to make use of such information. Again, there now appear to be mechanisms (e.g. temporal) by which such structure might be extracted, so this argument against connectionism also appears to have been met. One interesting point not made by F&P, is how the symbolic approach fares with regard to these representational criteria. Clearly it manages structural concerns (using syntax and concatenation–and some semi-magical mechanism to process it), but what determines whether a representation is currently active or not? Either it is conjoined with a truth token or, in modern digital computers, it is its particular (memory) location that indicates its status–both, in effect, giving a special, perhaps unwarranted, place to the notion of truth.

It was Searle (1980) who pointed out another major problem with the classical symbolic approach: the fact that its symbols didn't actually mean anything! His now infamous Chinese Room thought experiment suggested that manipulating meaningless "squiggles" was never going to result in meaning; that syntax was not sufficient for semantics. The ensuing debate has a long and tortuous history. It began with the idea that symbols can gain meaning from their relations to other symbols. This won't work—it is like looking up a word in a foreign language dictionary—the explanation, in terms of yet more foreign words, doesn't help (even if you look them up, all you get is more foreign words). Ultimately you have to know what some of the words mean, i.e. some of them must be "grounded" (Harnad, 1993). Grounding, in essence, requires causally connecting symbols to the world, such that when an agent sees, for example, a cat, the corresponding "cat" symbol is "tokened". This suggested that the connectionists were on to something after all. Unfortunately, there are a number of difficulties with this too, including questions about how such symbols arise and how they can be in error. The best answer we have so far is that a symbol/representation/state has meaning for the agent if there is a possibility that it can use it to successfully predict and act in the world (Bickhard and Terveen, 1995; Davenport, 1999).

Of course, the most obvious difference between people and computers is our apparent subjectivity, our awareness, and the feelings we have about the world. Dreyfus(1972; 1992) argued that human intelligence and expertise rely on unconscious instincts and intuition, which cannot be captured by formal rules. If, by this, he is referring to our inability to verbalise these rules and our difficulty in realising how we solved a problem, then he is correct. However, this does not preclude physical entities, other than people, from doing likewise.

Partly in response to the difficulties perceived as inherent to the symbolic and connectionist paradigms, a number of proposals for more radical alternatives started to appear. These included dynamical systems (van Gelder, 1995), embodied cognition (Gibson, 1979), radical embodied cognition (Chemero, 2009), embedded cognition, situated cognition (Robbins and Aydede, 2009), extended cognition/mind, interactivism (Bickhard and Terveen, 1995), enactivism, etc. In essence, these divide the (design) space into those that see cognition as based on representations versus those who favour an eliminativist (non-representational) approach (computational/non-computational); those who see the body as an essential component of cognition versus those who see it as incidental (a mere input/output device), and those who see the environment as crucial versus those who see it as incidental. As is often the case, such dichotomies seem contrived, specifically designed to focus on certain aspects of the problem that have, perhaps, been neglected in the past. Ultimately, the truth must lie somewhere between the extremes and include aspects of each viewpoint. In the following sections I will present my own (simple-minded engineer's) attempt to understand and clarify things.

3 Engineering Intelligence

The engineering process traditionally recognises requirements, design and implementation phases, followed by test, distribution and maintenance. The requirements phase is concerned with function; identifying the problem to be solved. The design phase then represents an abstract solution to the problem, a plan which is then implemented (in the implementation phase) producing a concrete mechanism that matches the design. The mechanism (product) can then be tested for correct functioning and, if all is well, distributed to customers. Finally, the maintenance phase usually comprises incremental bug-fixes and updates that improve the system. I will now look at the process of designing a cognitive AI with reference to each of these engineering phases with the exception of the test, distribution & maintenance phases, which we need not concern ourselves over since these are clearly and very competently handled by the environment & evolution (which is unmerciful in weeding out those less successful designs).

3.1 The Requirements Phase

In order to design a product, be it an AI or a toaster (or an AI toaster), it is first necessary to determine what it needs to be able to do and what context it is to operate in. For the toaster it is relatively simple, it needs to heat up bread in such a way that the surfaces burn slightly and it must do this in a context where the bread comes in slices of certain dimensions and where there is 220 volt electricity available (not to mention an environment with certain gravity, and an appropriate atmosphere containing sufficient oxygen, etc., etc.). For an agent, natural or artificial, its primary goal will (usually) be survival in a very complex and changing environment. For biological creatures, this means maintaining an ability to move around in order to find nourishment and to sustain themselves, while avoiding any physical damage. Among other things, this requires maintaining an appropriate body temperature, blood pressure/flow, etc., being able to locate suitable food, and avoiding predators.

Some of these tasks are relatively simple, but some are extremely tricky due to the inherent vagaries of Nature. As McTear (1988) eloquently put it, "the unpredictability of the world makes intelligence necessary; the predictability makes it possible". Agents must try to take advantage of any regularities they can uncover in order to select the course of action best suited to their goals. The fact that they are small (but presumably physical) parts of the physical world, implies they are likely to have only limited knowledge and so be subject to error. Agents must somehow detect the situation they find themselves in, try to predict the outcomes of any possible actions, and then select the action that appears the most beneficial. Ideally, they will need to remember the consequences of their actions so they can learn and perhaps choose a more preferable option, should they encounter a similar situation in the future.

In an abstract sense these are all control problems, the complexity of which vary widely. Consider, for example:

- maintaining body temperature / blood pressure, ...
- tracking prey/predator even when occluded, walking/climbing, ...
- conversing in English, doing math, socialising, creating/telling fiction, ...

The (design and) mechanism for accomplishing each of these (control) tasks would be different. For instance, the first (simplest) sorts of task require only a simple fixed feedback system. The relevant decision-making data is generally available, the possible actions are few and known, and so almost no learning is needed. More sophisticated tasks may require feed-forward predictive controllers, and may be expected to work with less reliable information (incomplete & noisy sensory data), and to demonstrate very complex patterns of behaviour that might change based on experience. The most complex tasks, those so far unique to humans, require what might be called knowledge-based controllers. Such a device would be characterised by its ability to handle an extremely wide range of situations and to learn, so that it may not perform the same even under identical circumstances. It is this level of performance that is the focus of Newell & Simon's (1976) "Physical Symbol System Hypothesis" (PSSH), which claims that "A physical symbol system has the necessary and sufficient means for [human-level] intelligent action". A physical symbol system is, roughly, a physical implementation of a symbol system; that is, of a set of symbols and a set of (inference/rewrite) rules that specify how the symbols can be manipulated. Newell & Simon couch their definitions in very general terms, such that it might be taken to include both symbolic and connectionist-like systems. They also assume that the symbols involved "designate" things in the world. The evidence they offer for the truth of the PSSH is (a) the obvious successes of symbolic AI and, (b) psychological experiments that show that human problem solving involves symbol manipulation processes, which can also be modelled by AI symbol systems. Fodor's (1975) Language Of Thought theory provides further support, as does the simple fact that humans can simulate universal Turing machines. Notice that a PSS-level mechanism could perform the simpler tasks, but there is no way that the simpler mechanisms could perform the PSS-level tasks. Having thus sketched the requirements for a cognitive agent and the context in which it must function, it is now time to move on to the design phase.

3.2 The Design Phase

Design is obviously constrained by requirements, but also by the properties of materials available for the implementation. To take an everyday example, if the requirements are to shelter human beings from the extremes of temperature, wind and rain (on this planet), then we might do this by building houses made of wood, of brick, of concrete, or of steel and glass (though probably not silk or banana skins). Even suitable materials obviously have different characteristics, such that we could build skyscrapers with two hundred floors from steel, but presumably not from wood. Availability is another concern; there may well be situations where steel is not an option and wood actually provides the best/only choice. What materials are available, their characteristics, and our ability to work with them, can thus significantly shape the range of design solutions. When it comes to designing a cognitive AI, however, it may not even be clear what materials are suitable. Clearly, biology works, but what of semiconductors or perhaps something else? What properties are relevant? To answer this question will require a slight digression.

I claim (and will try to argue) that cognition is essentially control, organised around a prediction/modelling mechanism, and that the design of a prediction/modelling mechanism can be expressed/described as a computation. This is a broad (and perhaps controversial) view of computation, but one I feel is justified. We humans naturally construct mental models of our environment (not to mention models of fictional worlds or situations). We then use these models to respond to questions about the actual (or imaginary) world. In the most sophisticated cases, such models are used to run "simulations" of the world, so as to predict possible future states and to see how those may change should the agent act in different ways. Armed with the outcomes of these simulations an agent can then select the action it sees as the most beneficial. Now the question is how such models (or simulations) can be constructed and run. Modeling of the real world must mean that states of the model can somehow be mapped to states of the world and that the sequence in which the states evolve also follows the same trajectories (notice that time in the model does not have to be the same as in the real world, only the sequence matters). What determines the sequence of states? Clearly, the causal pathways. When implementing such a modeling mechanism we have to rely on causation. We can either try to find an existing system with the appropriate dynamics, construct one anew, or, more commonly nowadays, we can turn to our universal machine, the (digital von Neumann) computer, that can be programmed to provide any desired causal behaviour. Notice that the only concern is the causal evolution of the system. None of the material properties matter, unless they impact the causal flow. Thus biology, semiconductors, or beer cans, are all equally suitable materials for constructing such devices. A program/algorithm/computation, then, is simply "an abstract specification for a causal mechanism" (Chalmers, 1995; Davenport, 1999) that will implement the model/computation. Learning involves changing the causal pathways so as to produce different behaviours.

Designs for the simplest sorts of tasks (e.g. maintaining body temperature or blood pressure) can now be cast in this light. Take, for instance, van Gelder's (1995) example of Watt's centrifugal steam engine speed governor (which he claimed had no representations and so was not computational and thus necessitated a dynamical systems approach). Such a governor needs to select one of only two actions (increase or decrease the steam going into the engine), directly predictable from the current engine speed. Any mechanism that provides such control is acceptable. The mechanical linkages of Watt's centrifugal governor do exactly that, simply and reliably. The device could, of course, be replaced with appropriate sensors, actuators and a control system, electronic or biological, though these may prove much less reliable. Notice, also, that there may be numerous designs for the inner workings of the control system (feedback, feedforward, bang-bang, etc.), but that, providing the requirements are met, they are all candidate solutions. Note too, that they are all causal mechanisms and so have a computational description; i.e. they are computational with reference to our broad understanding of computation.

When it comes to designing agents capable of displaying human-level behaviour, we most certainly need a more sophisticated mechanism—a PSS, equivalent to a general-purpose von Neumann machine. Given that an agent can have no a priori knowledge of the world's regularities[1], it would seem that the best it could do would be to store what it senses and detect when a similar situation occurs again. Of course, it also needs to maintain a record of the sequence of situations, including any actions it may have taken. Given "situation, action, new situation" data, it should have the information it needs to make "intelligent" actions in the future, but how? In order to shed light on this and to help resolve a number of important issues, in particular the fundamental difference between the symbolic and connectionist approaches, it is necessary to go right back to basics.

4 Back to Basics...

How can we communicate and store messages? Imagine we have a man on a far away hill, with a flag in his hand. To begin with, he can hold the flag in only two positions, either down in front of him or up above his head; initially it is down in front of him. An agent observing this distant man suddenly notices that the flag has been raised. What is he to make of this? There seem to be two possibilities. First, the pair could have established a convention beforehand, such that, for instance, the man raises his flag only if an enemy is approaching. Thus, on seeing the flag go up the observer quickly prepares, pulling up the draw-bridge and manning the battlements. The second possibility is that no convention has been established, so when the flag is raised the observer can only guess at its purport. Let's say he assumes the worst, that the raised flag means imminent attack and so he takes the precaution of pulling up the draw-bridge and manning the battlements. Thankfully he is mistaken and it was not an enemy approaching. Unfortunately, it is some rather important guests who are taken aback by the unfriendly reception. The next time the man raises the flag the observer recalls the embarrassment and so quickly begins preparations to welcome honoured guests. But again he is mistaken.

[1] It might be argued that evolution has naturally selected for certain architectural characteristics (both physical bodies and mental structures) which, in effect, embed some a priori knowledge of the world.

This time it turns out to be the local tax collector who is overawed by the reception, but thinks it may be a bribe. The observer might continue this guessing game or he may decide to experiment, sending out invitations to various parties to see which ones invoke the man to raise his flag. Through such interactions the observer can build up a better picture of just what the signal means, and so hope to avoid any untoward situations in the future.

There are two ways in which this simple signalling system could be extended to communicate more messages. The first and most obvious way would be to allow the man to hold the flag in more positions; for example he might be able to hold it out to the left and to the right, in addition to above his head and down in front of him. This would allow him to communicate three messages. This could be further extended, in theory to allow him to signal an infinite number of messages. Of course, he would have practical difficulties doing this, as would the observer in attempting to decide which message was being sent. The distinction between a limited number of discrete messages and an infinity of messages, is the difference between the discrete and (continuous) analog forms of communication[2]. Note that the messages are mutually exclusive, so that only one message can be communicated at a time. The other way to extend the number of messages would be to have more men, each with their own flag and each of whom could communicate a message independent of (and simultaneously with) the others, so (given only two flag positions), two men could communicate two messages, three men three messages, etc. Alternatively, if the men are together considered to be communicating a single message, then two men might communicate three mutually-exclusive messages, three men seven messages, four men 15 messages, and so on[3]. One final variation would be to communicate parts of the message at different times. This corresponds to so-called serial versus parallel communication. In both cases, additional consideration may need to be given to synchronisation, but we will leave this aside for now.

In addition to communicating messages, an agent must be able to store the messages it receives and later recognise them if they occur again. Consider a set of (sensory) inputs to the agent–corresponding to the men with flags in the previous paragraphs. There appear to be two fundamental ways in which messages might be stored. The first way to remember an input pattern would be to create another set of men with flags (one man for each input), and have them simply copy the state of the corresponding input. This could be repeated for each instant, which would obviously require a lot of men and flags, unless they were reduced by storing only previously unseen patterns.

[2] The term digital is sometimes, perhaps incorrectly also applied to discrete encodings. The term analog also has another meaning, often conflated with this one, wherein it refers to a value, usually encoded in a specific material property, that varies in direct proportion to another value, e.g. the height of mercury in a thermometer varies in accord with the ambient temperature.

[3] In each case, one combination of states (e.g. no flags raised) must be used as a background, “no message”, signal.

For an agent to recognise when an input pattern reappeared would require it to compare the present input pattern with all the previously seen and stored patterns. It might attempt to do this in parallel, in which case there must be some mechanism to consolidate the ultimate "winner", or it may do it by sequentially comparing them, perhaps placing a copy of the winner into a certain winner's location. While these are possibilities, they seem messy and unintuitive, a situation that would be compounded when it became necessary to store and extract sequence information and to handle inexact matches.

Contrast this with the other fundamental way in which the storage and recognition might be managed. This time, create only a single man and flag for each instant, but establish links (wires/pathways) from him to each of the inputs that are presently signalling (but not to the inactive ones). Now, when a previously seen input pattern reappears, the links connect directly to the corresponding man who can quite easily use his flag to signal that all the previously seen inputs are once again present. In the case of partial matches, should several men share a particular input then they become alternatives. Hence, if one has more of its input pattern active, then it can suppress the other less likely combination in a winner-take-all fashion. Extending this to store and detect sequences is also relatively straightforward. Furthermore, the newly created men—that link to the (sensory) inputs—can form the input pattern for yet another higher level of men and flags, and so on for as many levels as required.

5 Is Cognition Symbolic or Connectionist?

Earlier we asked what the relation between the symbolic and connectionist paradigms was: was only one approach correct and if so which one? Was a hybrid solution required? Were they actually alternative approaches, or were neither correct and so some other solution needed to be sought? The subsequent discussion has hopefully clarified this. The copy and link storage-recognition methods provide a clear and principled way to distinguish the paradigms. The classical symbolic paradigm is based on the "copy" mode; whenever a representation is needed in an expression, a new copy (token/instance) of it is generated (in the same way that, for example, the letter 'a' is copied over and over again, as it is used throughout this paper). In contrast, the connectionist paradigm is clearly based on the "link" mode of storage; whenever a representation is needed in an expression, the expression is linked to a single existing version of the representation. This distinction is equivalent to parameter passing by-value (copy) vs. by-reference (link) in computer programming languages.

The symbolic and connectionist paradigms thus appear to be genuine alternatives; that is, an intelligent agent could equally well be designed and implemented using either approach. To be fair, it is still not certain that the connectionist (link) scheme can actually provide the necessary functionality, but as I will try to demonstrate below, I believe it can. Indeed, given that the human brain almost certainly uses the "link" scheme, it makes sense to concentrate efforts on explicating it. In the following paragraphs I will try to show how the connectionist approach outlined above, may provide insights into some of the more perplexing puzzles regarding human consciousness experience–intuitions which the symbolic approach fails to give any real clue about.

6 Connectionist Insights

So far we have seen how an agent can store sensory input by creating causal links from the active inputs to a newly allocated man and flag. It can take the outputs from such men and use them as inputs to another tier, and so on, to produce a (loose) storage hierarchy. When new input is received it is matched against this hierarchy. In cases where only a partial match can be found, links to any unmatched inputs form "expectations" (since they were active on a previous occasion). Hence the agent has "anticipations" of the missing signals. If we assume that the men (like neurons) retain their state for at least a short amount of time after the input signals are removed, such expectations can serve to "prime" the hierarchy so that interpretation of subsequent inputs will tend towards matches that include previous solutions.

Assume, now, that "nodes" (men or neurons) can detect and store either simultaneous input signals (as before) or signals that arrive in a particular order (i.e. signal "A" precedes signal "B"). Coupling this with the idea of "state-retaining" nodes provides a means to accommodate sequence processing (an alternative to the feedback employed by recurrent neural networks). Finally, notice that state-retention can provide a decoupling of nodes higher up the hierarchy from those closer to the input layer. Together, such features may provide a way in which goal-directed behaviour might be achieved and understood–the very top level nodes remaining active and providing the expectations that guide the lower levels.

We have already seen that agents use the information they store for the prediction and selection of appropriate actions. The mechanism thus forms a model that can be run to simulate what will happen. Of course, initially models will be very simple and incomplete; each man and flag being essentially a "model" of some relationship in the environment. Over time, however, more sophisticated models will evolve and be refined as a result of interactions with the environment. The decoupling mechanism is a vital part of this since it allows models to run simulations independent of the current sensory input, enabling longer term planning and actions. Note that agents will almost certainly retain and use multiple models with varying levels of detail and completeness, so that they can respond rapidly when the need arises, but be able to “think it through” if time allows.

Given that agents are part of the environment they act in, they naturally need to model themselves too. Any reasonably sophisticated agent will thus develop a "self" model—that we ultimately label "I" or "me" after acquiring language. Our history of interactions with the world, including other agents in it, becomes associated with this model, gradually building up our personal identity.

Finally, and much more speculatively, the expectations that result from missing inputs may help explain "feelings". Should a node become active without any of its inputs—as a result of higher-level expectations, for example—then it will produce "expectations" on all of its inputs. These "prime" each of the inputs; a situation very much like the original one when the actual inputs were present. Note that similar stimulation may result during dreaming and during brain surgery when a neurosurgeon stimulates individual neurons.

6.1 Internal & External Symbols

So far only signals (representations/symbols) that are internal to the agent have been considered. External symbols, words, signs, etc., also have internal representations. How can such external symbols gain meaning? Assume we have one hierarchy (of men with flags) that store and recognise certain physical states of affairs, for example a cat or a dog, when seen by the agent. Assume also that there is another hierarchy in which the agent stores and recognises audible states of affairs, for example the spoken word "cat" or "dog". Situations will arise in which the spoken word and the actual entity are present simultaneously, and the agent will store these states of affairs in the same manner as any other (linking the relevant nodes in each hierarchy to a new node). Subsequently, on seeing a cat, the previous situation (in which the word & object occurred together) will be recalled and so–as a result of the normal mechanism that fills-in missing inputs–the expectation of the word "cat" will arise. Similarly, should the word "cat" be heard it will produce an expectation (a mental image) of a cat. In this way, then, external symbols acquire meaning for the agent.

When it comes to actually describing a situation it is necessary to "extract" structure from the network to form verbal sentences. Recall that F&P argued that this was not possible in ANNs, but if we accept that the agent abstracts the natural language's grammar in yet another heirarchy, it is not inconceivable that the basic mechanisms described above could combine it with the specific situation to generate the necessary words one-by-one.

6.2 Connectionist Logic?

The link (connection) storage method is clearly reminiscent of the connectionist (ANN) approach, each newly created man being like a neuron with multiple incoming links (dendrites) and a single output (axon). Given that the man is created when all his inputs (say a, b, & c) are active, his output (z) might be expressed as "if a & b & c then z". This form is known as a Horn clause and is common in Prolog and rule-based expert systems. Unfortunately, when interpreted as a material implication P -> Q, the "logic" is all wrong. At the very least, since z was created only when a, b & c were active, given z we should definitely be able to say a & b & c were true. Logic, however, dictates that given Q & P -> Q, no conclusion can be made regarding P; see Davenport (1993b) for a more detailed discussion.

Inscriptors (Davenport, 1993a) seem to offer a much better formulation, viz. "if z then a & b & c". Viewed as a material implication, this allows us to conclude a & b & c from z (as desired) and, using abduction, correctly suggests that z may be true if any subset of a, b and/or c are true. It is only possible to actually conclude z if it is the only candidate, and even then it may be wrong. A similar form of reasoning was used very successfully by Peng and Reggia (1990) in medical diagnostic expert systems based on their Parsimonious Covering Theory, providing some evidence that the approach described here is viable.

It is interesting to note that the mechanism employed by each "neuron" provides logical non-monotonic reasoning. The decisions it reaches must be (logically) correct—given the agent's limited knowledge and processing resources. In other words, such agents have bounded rationality. Agents will have evolved this way, since the correct solution is, by definition, the one that most benefits the agent, and agents whose mechanism failed to make the correct choices would presumably have died out long ago.

7 To Conclude

There is no doubt that AI research has made huge strides, both in helping us to understand how the human mind works and in constructing ever better machines. Yet, for all the progress, its foundations remain shaky at best. This paper has been an attempt to build solid foundations by returning to first principles and adopting an engineering approach to the problem. The result is hopefully a much clearer and simpler picture of how agents may function. Examination of the possible ways in which agents could communicate, store and recognise messages, led to a better understanding of the processes involved, and so provided a principled distinction between the symbolic and connectionist paradigms. Since both approaches can achieve the same function it is clear that they are genuine alternatives. In other words, an intelligent agent could equally well be designed using either symbolic or connectionist techniques. To use the analogy of house building, the designer is essentially free to build above or below ground, the choice being irrelevant as regards the function–though of course other tangential considerations may tip the balance one way or the other. Likewise, then, the representations and processing in an agent may be continuous or discrete, serial or parallel, symbolic or connectionist (or any combination thereof). And, just as a house could be designed and built in a variety of materials, so too could an agent–both symbolic and connectionist "faces" being computational and so multiply-realisable.

Given this, what of the PSSH; is it wrong to claim that a symbol system is necessary and sufficient for intelligent action? Actually, no. The PSSH is concerned with the higher functional level, not with the design and implementation. All it says is that such and such abilities are needed (Nilsson, 2007). The problem has been that those requirements have often been conflated with the design/implementation, since there seemed to be no means to create a symbol system, other than the "copy" (token) one. Now we can see that the "link" option is a real alternative, perhaps we need different vocabulary for the PSSH level so as to clearly distinguish it from the design/implementation level.

And what of the newer contenders—embodied, embedded, and extended theories of cognition—that reject representations, emphasise the role of the agent's body and/or their situation within the world? Most of these research programs primarily aim at controlling bodily movements, usually by modeling the agent and its environment, and analysing the coupled system in dynamical terms. For observers this is fine, but it doesn't explain how an agent can come to know the "outside" world, which is exactly what makes cognition difficult and intelligence necessary! An alternative, simpler approach, has been to avoid having the agent build detailed internal models of the world at all, and instead have them "look-up" the information from the environment as needed; "the world as its own model." These, however, are all relatively low-level functions and, as we saw earlier, they simply cannot account for the full range of human intelligent behaviour. At the other end of the spectrum are approaches, such as situated cognition, which promise to somehow combine low-level behaviour with higher cognitive levels. Here, the focus has been on language and how we interact with other agents in a socio-cultural environment. Much of this behaviour is undoubtedly a consequence of certain incidental biological needs (for nourishment, warmth, sex, etc.), and/or limitations (of memory and processing speed—leading to extended mind like interactions), and so not a function of cognition per se. Hopefully, the analysis presented here provides an outline of how a computational mechanism (an agent), with a body, operating in a socio-cultural environment, may actually come into existence and function.

To date, most AI work has concentrated on the symbolic paradigm or on connectionist networks that tended to use distributed representations, had little or no feedback mechanism to handle sequence, and required thousands of epochs to train. Consideration of the requirements level has shown that there is a more realistic design alternative for the "link" (connectionist) form. This is important since it would appear Nature has adopted this "link" scheme for use in our brains. Further work is needed to fully expound the mechanism and its implications, but, as we saw above, it does seem to offer clues regarding some of the most intractable problems AI has faced, including intentionality, feelings, and even consciousness, as well as deeper philosophical conundrums regarding the ontology of the world, our place in it, and the notion of truth (Davenport, 2009; Floridi, 2011).

At the very least, I hope this paper has presented the core concepts and arguments in a clear and understandable form, and that it affords a framework that will help others put the vast literature into some sort of perspective.

References

Bickhard, M.H., Terveen, L.: Foundational Issues in Artificial Intelligence and Cognitive Science: Impasse and Solution. Elsevier Scientific (1995)

Chalmers, D.: On implementing a computation. Minds and Machines 4, 391–402 (1995)

Chalmers, D.J.: Subsymbolic computation and the chinese room. In: Dinsmore, J. (ed.) The Symbolic and Connectionist Paradigms: Closing the Gap, ch. 2, pp. 25–48. Lawrence Erlbaum Associates (1992)

Chemero, A.: Radical Embodied Cognitive Science. MIT Press (2009)

Davenport, D.: Inscriptors: Knowledge representation for cognition. In: Gun, L., Onvural, R., Gelenbe, E. (eds.) Proceedings of the 8th International Symposium on Computer and Information Science, Istanbul (1993)

Davenport, D.: Intelligent systems: the weakest link? In: Kaynak, O., Honderd, G., Grant, E. (eds.) NATO ARW on "Intelligent Systems: Safety, Reliability and Maintainability Issues", Kusadasi, 1992. Springer, Berlin (1993)

Davenport, D.: Computationalism: The very idea. In: New Trends in Cognitive Science, Vienna (1999), `http://www.cs.bilkent.edu.tr/~david/papers/computationalism.doc`; also published on MIT's COGNET and in Conceptus Studien 14 (Fall 2000)

Davenport, D.: Revisited: A computational account of the notion of truth. In: Vallverdu, J. (ed.) ECAP 2009, Proceedings of the 7th European Conference on Philosophy and Computing, Universitat Autonoma de Barcelona (2009)

Dreyfus, H.: What Computers Can't Do. MIT Press (1972)

Dreyfus, H.L.: What Computers Still Can't Do: A Critique of Artificial Reason. MIT Press (1992)

Floridi, L.: The Philosophy of Information. Oxford University Press (2011)

Fodor, J.: The Language of Thought. Harvard University Press, Cambridge (1975)

Fodor, J.A., Pylyshyn, Z.W.: Connectionism and cognitive architecture: a critical analysis. Cognition 28(1-2), 3–71 (1988)

van Gelder, T.: What might cognition be if not computation? Journal of Philosophy 91, 345–381 (1995)

Gibson, J.: The Ecological Approach to Visual Perception. Houghton-Mifflin, Boston (1979)

Ginsberg, M.: SIGART Bulletin 6(2) (1995)

Harnad, S.: Grounding symbols in the analog world with neural nets. Think 2, 1–16 (1993)

Haugeland, J.: Artificial Intelligence: The Very Idea. MIT Press (1985)

McTear, M.: Understanding Cognitive Science. Horwood Ltd. (1988)

Newell, A., Simon, H.: Computer science as empirical inquiry: Symbols and search. Communications of the ACM 19(3), 113–126 (1976)

Nilsson, N.J.: The Physical Symbol System Hypothesis: Status and Prospects. In: Lungarella, M., Iida, F., Bongard, J.C., Pfeifer, R. (eds.) 50 Years of AI. LNCS (LNAI), vol. 4850, pp. 9–17. Springer, Heidelberg (2007)

Peng, Y., Reggia, J.: Abductive Inference Models for Diagnostic Problem Solving. Springer, New York (1990)

Robbins, P., Aydede, M. (eds.): The Cambridge Handbook of Situated Cognition. Cambridge University Press (2009)

Rumelhart, D.E., McClelland, J.L.: Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1. MIT Press (1986)

Searle, J.R.: Minds, brains, and programs. Behavioral and Brain Sciences 3(03), 417–424 (1980)