Essay 9

Reconstructing AI

Conor Doherty

Educational Research Centre,
St. Patrick's College, Dublin.

Abstract

Symbolic AI is argued to be epistemologically and ontologically necessary but insufficient for constructing robust AI. Two principles, embodiment and situatedness, are elaborated which any global theory of AI must incorporate. These principles require autonomous robotics to form a basis for AI. Learning is the key to the development of more autonomous robots. Artificial neural networks are evaluated for their ability to learn to integrate robust sensory categorisation with motor control. The future relationship of artificial neural networks to symbolic AI is speculated on.

9.1 Thinking Machines

How do you swat a fly with your hand? Is intelligence required for visually guided grasping of a moving object? Given the complexity and fluidity of your environment, how do you learn the appropriate representations for performing a particular task in a particular situation? Are explanations of *mental* phenomena such as language understanding ultimately contingent on explanations of *bodily* phenomena such as muscle control?

Philosophical attempts to cut through the Gordian knot connecting mind to body have been supplemented by two relatively modern methodologies. Following a natural science program, psychology, neuroscience, linguistics, anthropology, and ethology gather experimental data on humans and animals and construct accommodating theories. A novel computational methodology investigates objects that are considered imitations of living systems and their behaviour. Computers can simulate, at different levels of abstraction, the transformation of information performed by nervous systems. From the 1950s onwards, the success of the von Neumann digital computer promoted the information processing metaphor of mind as sequential software compiled onto nervous system hardware.

The metaphor implied that a computer could itself be artificially intelligent (AI). In a seminal paper, Turing (1950) suggested two, not necessarily exclusive, future directions for the development of AI:

> We may hope that machines will eventually compete with men in all purely intellectual fields. But which are the best ones to start with? Even this is a difficult decision. Many people think that a very abstract activity like playing chess would be the best. It can also be maintained that it is best to provide the machine with the best sense organs that money can buy, and then teach it to understand and speak English. The process would follow the normal teaching of a child. Things would be pointed out and named etc.
>
> Turing (1950, p. 460)

The "chess" or symbolic approach has dominated attempts to theorise about and build intelligent systems. While the symbolic AI enterprise is heterogeneous, its essential premise is that processes similar to introspected conscious reasoning underlie all intelligent behaviour.

9.2 Symbolic AI

Practitioners of symbolic AI maintain that intelligent behaviour involves the appropriate manipulation of discrete mental representations of the world (e.g., see Stillings *et al.* 1987). Typically these representations are viewed as linguistic symbols which stand for the real-world entities that they represent, and their manipulation is regarded as computation akin to formal logic (e.g., Fodor's *Language of Thought,* 1976). Traditional formal logics apply rules of inference to statements formed by applying Boolean quantifiers and connectives to symbols for individuals or classes. Classes (equating to mental concepts) are defined by rules for classifying symbols. Newell & Simon's (1976) articulation of the relationship of such logics to intelligence under the *Physical Symbol Systems Hypothesis* (PSSH) is adopted as exemplary. Physical symbol systems are collections of symbols (discretely identifiable patterns in a machine) with an associated syntax and formal rules of manipulation. Such a system, as "a machine that produces through time an evolving collection of symbol structures" (Newell & Simon 1976, p. 116), is hypothesised to have the necessary and *sufficient* means for general intelligent action. It is assumed that the symbols in such a system unproblematically denote (refer to) things in the world. System behaviour is described in terms of knowledge, goals, and actions. Evaluation of the PSSH is via verbal interrogation of a candidate system, the so called Turing Test. Ex hypothesi, perception, learning, memory, reasoning, language and action involve *only* the construction and manipulation of appropriate discrete symbolic representations of the external world.

Symbolic AI is competence oriented, modelling specific, often very advanced, human abilities (e.g.,

chess playing, medical diagnosis). The PSSH induces a top-down engineering methodology. First, a problem domain is selected. Then, formal models of this micro-world are defined, typically based on axiomatic (frequently introspected) outputs of sensory systems. It is assumed that sensory systems can deliver these in a straight forward manner. Top-down versions of planning problems, learning problems, etc. are solved in these formal micro models. The resulting problem representations, ingeniously crafted by AI programmers, are task specific and not readily generalisable to other domains. Nevertheless, the great success of symbolic AI at modelling serial conscious reasoning in clean precise problem domains, especially inherently formal ones like chess, has sanctioned certain fundamental assumptions about constructing AIs.

The PSSH and Turing's universal theory of computation (cf. Davis & Weyuker 1983) have perpetrated a doctrine of implementation independence. Given that all computers can be simulated by a special universal Turing machine and given that software ideally compiles on different hardware platforms, formal models are widely conceived to be sufficient to determine mental computation. However it is also well recognised that Turing's formalisation of computation is disengaged from the space/time complexity constraints on real-world computation imposed by finite memory, physical instability of different classes of computers, etc. The severance of formal models from implementation under the PSSH occludes the impact of implementation on generation and selection of appropriate formal atoms. Worse still, PSSH style attemps to explicitly formalise intelligence operate within a restricted conception of natural intelligence.

9.3 Insufficiency of Symbolic AI

9.3.1 Epistemological Insufficiency

Human cognition only approximates the ideal rationality of logic (e.g., Gelman 1988). Is this evidence of some sort of weakness of the processing system or the result of the underlying human processing mechanism? While our capacity to reason is vital for intelligent behaviour, it is probably more akin to pattern recognition than formal logic. Rumelhart *et al.* (1986b) argue that mental logics (e.g., iconic and analogical reasoning) are contingent on internalisation of external notation systems appropriate for a particular inferential task. This important aspect of human reasoning, development and selection of appropriate formalism, is peripheral to much symbolic AI.

Among the most impressive and powerful results of thirty odd years of research in symbolic AI are expert systems. These are formal codifications of technical domains which serve as important tools for enhancing naive human decision-making. They function as special-purpose intelligently organised information repositories. However, the obvious limitations of expert systems as a theory of general natural intelligence foregrounds the centrality of representational *autonomy* for any genuine AI. The development of autonomous intelligence using symbolic AI persistantly receeds. Nevertheless, negative results are valuable for hardening many issues pertinent to AI. Perhaps the most sustained symbolic attempt is the 10 year program of Lenat *et al.* to construct an AI system by encoding encyclopaedic knowledge into a very large knowledge base (Lenat & Feigenbaum 1991). Despite their periodic claims to be on the verge of breakthrough, Smith (1991) argues that what the project has in fact demonstrated is that making explicit aspects of knowledge which result from bodily participation in the world is the most intractable barrier to the formalisation of human knowledge.

The engineering shortcomings of the PSSH parallel philosophical cracks in the rationalist assumptions underpinning the hypothesis. During the eighteenth century enlightenment, it became widely accepted in Europe that reason applied to the control of mechanical nature resulted in progress. The rationalist conception of Truth elaborated Platonic and Cartesian dualism by requiring that:

> ...there is a unique set of concepts and a unique set of propositions employing these concepts that adequately express the nature of the world, and that these propositions form a system and could ideally be recognised as a set of [*a priori*] necessary truths.
>
> Willams (1972, p. 73)

This may seem eminently plausable and it is the *modus operandi* of many logicians but it is also hotly disputed by subsequent romantic and modern philosophers. The most devastating epistemological critique of rationalism or transcendental objectivism is due to Hilary Putnam (1981). Model theories are formal frameworks for objectivism. For model-theories, the meaning (semantics) of a sentence is a function that assigns a truth value to the

sentence in each possible situation. The meaning of each term in a sentence is that which it *refers* to in each possible situation. Thus, summation of the truth values of a sentence's elements results in a sentence's truth value. However, for any successful theory of meaning, changing the meanings of the components must alter the meaning of the whole sentence. If this requirement is violated by even one valid sentence governed by a candidate theory of meaning, it negates the theory.

Using this requirement as leverage, Putnam *proves* that model theories refer inconsistently to entities in the world. Given any model theory, he demonstrates that the truth value of a sentence like *a cat is on a mat* can remain constant while the referent of *cat* changes from cats to cherries and the referent of *mat* changes from mats to trees (using judicious definitions of cat and mat). But ... if the reference of the elements changes while sentence truth is preserved, then meaning evaporates. PSSH, also characterisable as the pairing of meaningless strings of symbols with meaningless model structures, thus does not qualify as a theory of meaning. The logical inadequacy of an objective representational interlanguage is supported by evidence on natural categories in language (Lakoff 1987, 1987, see below) and arguments from evolutionary biology (Maturana & Varela 1980, see below).

Consideration of the above leads me to make a strong claim: *The post-enlightenment failure to explicitly articulate and encode the totality of human knowledge results from the absence of a fixed, objective universal description of the world.* If a stable general representation of the world independent of a given task is unavailable for encoding into formal representations to be operated on by symbolic AI systems, then a critical problem for such systems is their extra-referential status outside their programmer's mind. This appears to be the basis for Searle's (1980) Chinese Room argument that strong symbolic AI is devoid of semantic content.

What nonetheless accounts for the intuitive persuasiveness of combining discrete symbols according to rules to explain our *rational* behaviour? When people are asked to give retrospective reports of their mental processes, they tend instead to provide a justification for their actions. Rationality is essentially this type of social justification (Harrè 1983). What is labelled rationality/logical reasoning, and attributed to the working of the individual mind, is a public reconstruction meant to legitimate a conclusion by showing it can be derived by procedures recognised as valid. According to this viewpoint, rules are part of the regulative framework for the social construction of rationality. Personal rationality, results from turning the social process of justification inward upon one's own thoughts. The cognitive apparatus is increasingly constrained to output according to what is publicly justifiable rather than driven by any intrinsic formal system. Rules, thus may play an important role as knowledge that enters into behavioural computations but do not constitute the computational algorithms themselves. Once it is accepted that rules operating on discrete symbols are insufficient for generating intelligent behaviour, then we may wonder just what is?

Extant successful biological algorithms take into account that action is essentially situated (i.e., contingent on actual unfolding situations). It appears that *a priori* prescriptions can not anticipate all contingencies that could arise during a given interaction of a system with the real world. Furthermore, *a posteriori* rationalisations of actions often suppress details that are critical during an action. Agre & Chapman (1987) conclude, based on analysis of the inadequacy of rule-based robot motion planning, that rule-based representations are useful for post hoc communication about intelligent behaviour rather than mechanisms underlying such behaviour. The kernel of intelligent behaviour is the ability to adapt to dynamic environments. An agent engaged in ongoing interaction with her environment continuously adjusts to the changing internal and external circumstances of that interaction in such a way as to achieve her objectives. Maturana & Varela (1980) argue that adaptive animal behaviour only requires such structural congruences between the dynamics of an intelligent animal's internal mechanism and the dynamics of the external world. There is no necessity that an observer be able to distinguish discrete internal structural configurations, or complex functions of such representations which correspond to the animal's environment (à la PSSH).

From these perspectives, the impressive performance of many symbolic AI systems has limited relevance for understanding natural intelligence. Symbolic AI emphasises algorithmic processes like search or exact reasoning while neglecting such basic natural adaptive abilities as perception and categorisation. Perception and motor control are the hard problems that any autonomous intelligent system must solve. Input/output representations constrain greatly the engineering of aspects of intelligence, like inference, that are contingent on categorisation. Rosch's (1981) investigation of natural linguistic categories indicates that many are structured probabilistically and interconnected in a manner not amenable to intensional definition. Problems con-

necting the arbitrary PSSH symbols used in internal reasoning with external physical stimuli, "symbol grounding" (Harnad 1992), and resulting AI system failure in domains even a little different from the ones they were programmed for, "brittleness" (Holland 1986), highlight the need for bottom-up design. During evolution, environmental demands channeled sensor and effector design which in turn structured adaptive control architectures. Neglect of the effect on categorisation of sensor and effector structure leads to symbolic AI's second deficiency.

9.3.2 Ontological Insufficiency

Historically, symbolic AI ignored neuroscience as largely irrelevant to its goals because of an implementation independence doctrine. Marr (1982) elaborated the most sophisticated version of this autonomy thesis. He argued that vision can be analysed into three loosely coupled functional levels:

1. A *Computational Theory* level characterises the goal of a computation.
2. This goal is realised by an input output *representation and algorithm* level.
3. A physical *hardware implementation* level realises the algorithm that realises the goal.

Marr's analysis of operation levels is complete for stable systems not undergoing morphological transformation as a result of processing, unlike adaptive biological systems. Considering similar neurbiological evidence, Arbib (1987) concludes that:

> ... Marr's (e.g., [Marr 1982]) notion of an independent computational level of analysis as mistaken—for example, one cannot give an a priori analysis of depth perception because different animals (or different subsystems of a given animal) may make different uses of different cues that cannot be discovered until 'implementational details' (the data of neuroscience) are taken into account.
>
> Arbib (1987, p. 407)

Arbib's claim is that a biological system's implementational details contribute centrally to the way it's task is conceived and described at the functional level. Additionally, structural network levels of description provide a compact way of encompassing large classes of symbolic algorithmic descriptions of the behaviour of a system where a network with a particular learning algorithm can generate many specific algorithms to solve large classes of problems. Current neuroscientific data indicates that:

> ... software and hardware are one [and] the same in the nervous system. ... Extensive evidence indicates that the brain is not an immutable series of circuits of invariant elements; rather it is in constant structural and functional flux. The digital computer analogy is fatally misleading.
>
> Black (1988, pp. 2–3)

At the other end of the mind/body continium, the consequences for mental functioning of bodily structure have been investigated by Phenomenology. Phenomenologists claim that meaning is located at the intersection of a person's social, historical, bodily and spatial situation. Merleau-Ponty (1964) argued that whenever a person perceives or acts on the world, it is articulated in perspectives centrally related to the body, such as within or beyond reach, above or below, etc. Dreyfus's critique of symbolic AI (e.g., Dreyfus 1979) is grounded in phenomenological theories of meaning and since these in turn are derived from continental European dialectical and hermeneutic modes of inquiry, it is not surprising that Anglo-American logico-empiricists reject the claim that the frame problem results from representational context sensitivity all the way down to the sensorimotor foundations of cognitive activity.

Weaving epistemological and ontological strands together, Lakoff (1987) argues that the development of linguistic natural categories depends on a subconceptual layer of bodily experiences and imagery that are directly meaningful, with concepts drawing their meaning from their relations to the sub conceptual meanings. Such a functional semantics, in contrast to symbolic AI's model-theoretic semantics, assigns internal representations their meanings by virtue of their causal role in the mental processes of the instantiating organism. In other words, *what you know depends on what you do.*

Mechatronics concurrently integrates mechanical, electronic and informatic constraints to produce optimal structures, communication and control in machines. It provides engineering support for the benefits of tighter interrelation of functions and structures. Projected miniaturisation technologies enabling processes such as 3D molecular computing (Hansson 1991), will accelerate the collapse of the present software/hardware distinction. As molecular computers are developed, we may hope that

computation will become more brain like and even beyond. However, how can such molecular machines be released into constantly changing environments if explicit programming is inadequate?

Turing's second suggested AI research avenue incorporated situatedness and embodiment. He recommended placing a robot, *tabula rasa,* in a real sensory environment where it could learn by experience and interaction with a teacher. Some early research in artificial neural networks (ANNs) followed such a program.

9.4 Numerical AI

ANNs are computational systems loosely inspired by information processing in nervous tissue. In 1943, McCulloch & Pitts published a landmark paper on computation in the nervous system that implied that the brain computed logical functions using networks of simple threshold logic units (McCulloch & Pitts 1943). In 1958, Rosenblatt developed the *Perceptron,* an algorithm which by changing numerical values of connections between a pattern input layer and a classification output layer learns to associate input patterns with output patterns (Rosenblat 1958). In 1969, Minsky & Papert proved the Perceptron incapable of learning non linear classification and surmised (incorrectly it subsequently transpired) that a multi-layer learning algorithm did not exist which could learn such tasks (Minsky & Papert 1969). Their analysis precipitated the ascendancy of symbolic AI and the refraction of the first ANN phase.

ANNs are massively parallel processing architectures characterised by properties such as the ability to adapt and learn, to cluster or organise data, and to generalise. While the electrochemical dynamics of real neurons are only partially understood, they can ideally be considered as switching elements that sum their inputs and output a signal if the sum exceeds a threshold. Excitatory and inhibitory input connection "strengths" are represented by positive and negative numbers. Actual neural signals, which are pulse trains, are represented by a single average real value. This numerical style of computation does not assign a discrete referential meaning to connection values unlike symbols under the PSSH.

ANNs can be regarded as generalised tensor maps that transform input vectors to output vectors (Pellionisz & Llinas 1985). If a cost function is associated with the output vectors, it can be optimised by changing the "strengths" of network connections. Such optimisation is called learning. Learning is central to the construction of robust intelligence. The revival of ANNs during the 1980s depended on new learning algorithms such as backpropagation (Rumelhart *et al.* 1986a), and cheap serial workstations for simulation. The backpropagation algorithm overturned Minsky & Papert's (1969) speculation that a multi-layer Perceptron learning algorithm could not be developed which was capable of learning non linear classification.

ANNs have demonstrated themselves capable of learning variable binding (Touretsky 1990), albeit operating inefficiently compared to symbolic AI parsers. ANNs have widened the definition of compositional structure[1] to include functional as well as spatial concatenative composition (Elman 1990; van Gelder 1990). ANN's ability to bind variables and recursively combine them counters Fodor & Pylyshyn's (1988) charge that they are fundamentally incapable of natural language processing. ANNs promise adequate representational capacity for sensori-motor mappings required by autonomous robots. Autonomous robotics must form the basis of any theory of AI which encompasses situatedness and embodiment. Since the internal structure of representations learnt by ANNs reflects the *implicit* processing semantics of a given task without necessarily making *a priori* task assumptions (e.g., learning sample distributions), adaptation to environmental demands can form a basis for non-brittle categorisation and determinate reference. ANNs as universal function approximators can potentially learn the temporally convoluted functions that map sensor inputs to motor outputs.

Nevertheless, while ANNs are among the few computations that can automatically exploit an arbitrary amount of parallelism (Bleloch & Rosenberg 1988), an experimental disadvantage is that learning is NP-complete (Judd 1990) and prone to local maxima (like all gradient descent algorithms). In biological systems, natural selection has overcome these problems by inducing relatively precise organisation of the morphology and connectivity of neurons, specified epigenetically (softwired), that channels search to relevant sections of problem spaces during learning.

Similar techniques are required for designing ANNs. Large problem spaces can be functionally decomposed into smaller spaces which are searched by network modules that are later assembled to solve the larger task. Most successful contemporary applications of ANNs also invariably perform some form of input/output feature coding to combat learning complexity. Such problem decomposition into *a priori* representational spaces must proceed carefully

[1] The ability to recursively manipulate symbol structures.

to avoid the indeterminacy of reference which we have been alerted to by top-down AI's limitations. While ANN hardware developments rapidly increase the ability of ANNs to deal directly with sensor data (e.g., Mead 1989), substantial developments in sensor and motor processing are contingent on learning algorithms more isomorphic to the functional/structural organisation of real neural computation.

9.5 Incremental AI

Sensory and motoric systems have mutually co-evolved. Interacting with signals from sensory sheets, motor ensembles provide the nervous system with a mechanism for honing sensory feature correlation. Based on feature detection, categorisation occurs, in parallel and motor driven, of topological invariances and continuities that are crucial for detailed sensory abstractions. Edelman (1989) argues that motor activity is *the* essential mechanism underlying sensory categorisation and this in turn is essential for motor learning. During development:

> ... action is fundamental to perception, and sensory sheets and motor ensembles must operate together to yield a sufficient basis for perceptual categorisation ... The major and essential contribution of motor ensembles to perception is feature correlation, which arises out of the continuity properties of motion and the continual focusing of sensory signals by creating postural and gestural movements.
>
> Edelman (1989, p. 238)

Such a view of cellular reorganisation corresponds well to Piaget's theory of the construction of cognitive functioning. While the precise details of Piaget's *genetic epistemology* are controversial, it is widely accepted that individual language development and abstract reasoning are critically contingent on adequate early infant interaction with the world. Observations of children convinced him that during human development:

> ...concrete action precedes and makes possible the use of intellect ... As the infant begins to manipulate the objects which surround him, he gradually develops a practical 'understanding' of external reality ... Thus the acquisitions of the sensorimotor period form the foundations of the individual's mental development ...
>
> Ginsburg & Opper (1969, p. 106)

During ontogenesis, Piaget argues that development of spatial and self/body awareness is mediated by active visually co-ordinated reaching for and grasping of objects. At a phylogenetic level, relatively specific cortical circuitry (a "module" roughly equivalent to Broca's area) underlies the hierarchically organised combination of elements in the development of *both* speech and sequential manual action such as tool use (Greenfield 1992).

For experimenters, there is a simulation tradeoff between niche behavioural complexity and sensor complexity. Modelling complex behavioural patterns of simple complete animats requires simplified sensors which can interact with the data structures of a simulated world. Complex (e.g., visual) sensors require simple environments and thus restrict the behavioural repertoire or restrict modelling to behavioural sub-components. Such sub-components must be self-contained and incrementally grounded to minimise brittleness.

9.6 Conclusion

Symbolic reasoning is a cultural superstructure for enhancing adaptation. Explanation of discrete symbol manipulation probably requires explanation of the large chunks of adaptive behaviour that are not discrete-representational. The royal road to robust symbolic AI is paradoxically contingent on simulation of the neural mechanisms underlying the abilities of simpler organisms to effectively cope with the niches in which they are embedded. Robust autonomous intelligence requires a system to create its own task-driven representations grounded in the environment. Roboticists such as Brooks (1991a) argue that most intelligent control behaviour does not need, and indeed is hampered by symbolic formalisms. The MIT group builds robots bottom-up, enabling debugged sensor-effector competencies to be *subsumed* into successively more complex behaviours. Such a method replicates evolution's own incremental design algorithm.

Emulation of higher level cognitive functions such as reasoning and language evidently requires some form of symbolic representation. In order for future robots to autonomously perform complex tasks, (e.g., deep sea repair, household hoovering), a high level of symbolic reasoning will be nessesary. Whether modular ANNs capable of low-level sen-

sorimotor competences can be interfaced with pre-programmed symbolic AI programs to produce hybrid systems (e.g. Dyer 1991) or whether robust AI symbolic manipulation will ultimately require perceptual internalisation of initially external symbol systems as in humans remains to be seen.

Acknowledgements

I wish to thank Ronan Reilly for his perceptive comments and encouragement. I came across many of the ideas in the paper during the ESPRIT NERVES project.