Evolutionary Psychology
human-nature.com/ep – 2005. 3: 24-48

Original Article

The elementary nature of purposive behavior: Evolving minimal neural structures that display intrinsic intentionality

John S. Watson[1], Psychology Department, University of California, 3210 Tolman Hall #1650, Berkeley, CA 94720-1650, USA. Email: jwatson@socrates.berkeley.edu.

Abstract: A study of the evolution of agency in artificial life was designed to access the potential emergence of purposiveness and intentionality as these attributes of behavior have been defined in psychology and philosophy. The study involved Darwinian evolution of mobile neural nets (autonomous agents) in terms of their adaptive weight patterning and structure (number of sensory, hidden, and memory units) that controlled movement. An agent was embedded in a 10 x 10 toroidal matrix along with "containers" that held benefit or harm if entered. Sensory exposure to content of a container was only briefly available at a distance so that adaptive response to a nearby container required use of relevant memory. The best 20% of each generation of agents, based on net benefit consumed during limited lifetime, were selected to parent the following generation. Purposiveness emerged for all selected agents by 300 generations. By 4000 generations, 90% passed a test of purposive intentionality based on Piaget's criteria for Stage IV object permanence in human infants. An additional test of these agents confirmed that the behavior of 67% of them was consistent with the philosophical criterion of intention being "about" the container's contents. Given that the evolved neural structure of more than half of the successful agents had only 1 hidden and 1 memory node, it is argued that, contrary to common assumption, purposive and intentional aspects of adaptive behavior require an evolution of minimal complexity of supportive neural structure.

Keywords: agency, agent, artificial life, connectionism, emergent capacity, evolutionary psychology, intentionality, neural net, purposive behavior, goal-directed behavior.

Introduction

The concepts of purposiveness and intentionality have long been used in explanations of behavior. Aristotle's conception of final cause can be viewed in part as an attempt to give causal credit to the motivating force of anticipated future events

(Barnes, 1984, 1995). Descartes and other philosophers restricted the range of relevance of these and other mental concepts to explanations of human behavior. Animals were viewed as having insufficient mental and spiritual endowment to freely contemplate the consequences of their behavior. But humans, if sane and sufficiently mature, were credited and held accountable for behavior enacted intentionally.

Within the past century, there was a serious attempt in psychology to eliminate these mental concepts as legitimate causes of human as well as animal behavior. The radical behaviorism espoused by John B. Watson (1930/1957) and carried forward by B. F. Skinner (1938, 1953) denied any validity to explanations of behavior that were based on anticipated future events. They argued that the present stimulus situation and the history of past stimulus-response associations were the only legitimate sources of causes in a valid science of behavior. But this retrospective stance on explanation in behavioral science met a resistance that would eventually prevail. McDougal (1929) and Tolman (1932/1967) were notable voices of this resistance. They formalized their opposition in a framework McDougal termed 'purposive behaviorism.' They claimed allegiance to methodological aspects of Watson's behaviorism, but held fast to the value of explanations that included a prospective stance whereby an animal of sufficient cognitive capacity might be influenced by the mental representation of its goal.

In recent years, cognitive scientists have brought renewed focus to the nature and function of intentions in the control of behavior. Philosophers of science have argued over the necessary and sufficient conditions from which intentions might arise (Bennett, 1976; Searle, 1992; Dennett, 1971, 1996). Neuroscientists have begun investigating the brain localization and function of representations of intended motor acts in monkeys and humans (Jennerod, 1985). Developmental psychologists have studied the onset and elaboration of purposive behavior in human infants (Rovee and Rovee, 1969; Rovee-Collier, Morrongiello, Aron, and Kupersmidt, 1978; Watson, 1966, 1979) as well as the onset of human perception of purposiveness and intentions in others (Gergely, Nadasdy, Csibra, and Biro, 1995; Kelemen, 1999; Meltzoff, 1995).

Meanwhile, within jurisprudence, the causal conception of intentional behavior has remained essentially without serious challenge in western industrialized nations. The basic idea is that one is responsible for acts that one has intentionally performed. Intentional acts are presumed voluntary such that it is understood that if the individual were in the same environment, the act would not have occurred in the absence of the individual's intention to perform it. This counterfactual frame will be used as a basis of identifying instances of intentionality as distinct from the closely related concept of purposiveness.

In light of the historical attention given to purposiveness and intentionality, it is worth asking what is required in the way of cognitive structure to support these aspects of adaptive behavior. Historically, as noted above, there seems to have been an implicit assumption that intentional behavior requires relatively complex cognitive structure. With that assumption, it follows that in the evolution of behavior systems,

purposive/intentional behavior was a later derivative of earlier behavioral adaptation. The early adaptations presumably involved only simple sensory-motor structures. Yet, the level of structural complexity that might be required by the higher level adaptations is not at all clear (Sloman, 1999).

The objectives of this paper are to 1) review existing criteria for identifying instances of purposive and intentional behavior, 2) design an environment for artificial life in which purposiveness and intentionality would be an advantageous adaptation, 3) introduce a minimally structured neural-net agent that can be structurally modified by Darwinian mutation, and 4) observe if and when the targeted categories of behavior emerge and the structural complexity evolved for their support. I will conclude by considering implications of the results for some ongoing discussions within philosophy and psychology.

Criteria of Purposive Behavior

When should behavior be called purposive? As noted above, Watson (see Watson and MacDougall (sic), 1928), like Skinner (1938, 1953) somewhat later, said the answer is never. However, McDougall (1929), Tolman (1925, 1932/1967), Heider (1958), and many subsequent cognitive psychologists assume one can not avoid attributing purpose or intention to behavior if it meets certain criteria - at least not if one hopes to sensibly describe it or make predictions about subsequent behavior. Three criteria that are commonly noted are these:

> 1) *Equifinality***:** Purposive behavior results in a so called "equifinality" of behavior whereby a common end state is observed across varying situations because the behaving individual changes the observable pattern of efficacious behavior in a manner that adjusts to the changing circumstances. The ends remain the same while the means differ. Thus equfinality implies a behavior system that tends toward a particular end state. As depicted in Fig.1, Heider (1958) renders this concept in terms of agency in his naïve theory of action. As shown in Fig. 2, it can also be rendered in classic behavior theory. Heider draws a further distinction between equifinality that can arise in "impersonal causation" such as the predictable ultimate resting position of a pendulum or of a marble thrown in a bowl. In such cases, equifinality is accounted for by reference to the physical system as a whole (e.g. an "attractor" state in modern systems theory). By contrast, equifinality that arises from "personal causation" is accounted for locally by reference to the present disposition of the behaving individual.
>
> **Figure 1:** In Heider's naive theory of action, equifinality refers to the common end state (E1) achieved across varying situations (S) by

virtue of the agent (A) applying different means-ends action (m) in the different situations as guided by the particular dispositional state (D1) (goal set or intent) of the agent.

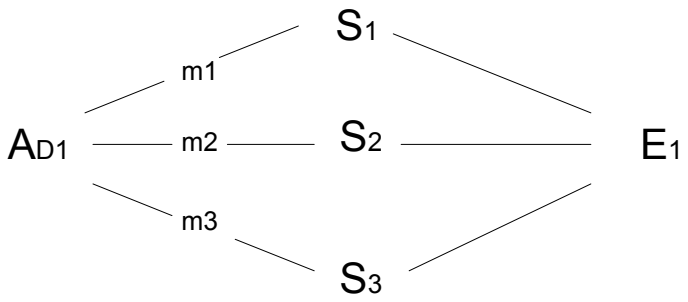
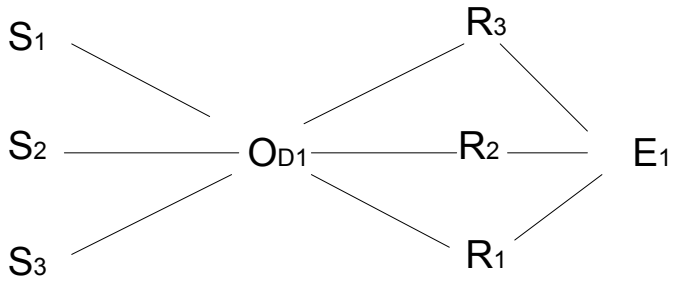


Figure 2: Equifinality can be rendered in classic (S-O-R) behavioral terms as the common end state (E1) observed across varying stimulus conditions (S) that elicit different responses (R) from the an organism in a particular dispositional state (O_{D1}).



2) *Persistence***:** Purposive behavior has the characteristic of what Tolman (1925) called "persistence until" (see also McDougall, 1929). Under most circumstances the behavior will persist until the end state or goal is obtained (but see Bennett, 1976).

3) *Rationality***:** Dennett (1971, 1987) has emphasized the principal of rationality in his analysis of purposive intentions and the "intentional stance." Gergely and his colleagues (Gergely & Csibra, 1997; Csibra,

Gergely, Biro, Koos, & Brockbanck, 1999) have recently employed this rationality principle in their study of when infants attribute purpose and intention to objects (see also McDougall, 1929; Piaget, 1936/1963; Tolman, 1932/67;). The rationality principle implies that, in general, purposive action will involve selecting the most efficient, least risky, and most speedy behavioral options for obtaining the goal.

Criteria of Intentional Behavior

There is much semantic overlap between the concepts of purposive and intentional behavior (e.g. as when defining intentional as doing something "on purpose"). Yet I think there is a useful distinction in what each term usually emphasizes. Purposive emphasizes the goal or end state of the behavior while intentional emphasizes the initiating state of the actor. Indeed, intentional often carries a counterfactual assumption as in a thesaurus reference to "premeditated" and "voluntary" as synonyms for intentional (Stein and Flexner, 1984). The counterfactual is the connotation that, given identical circumstances, the behavior would not have occurred if the individual had not held the same intention. So while equifinality is central to claiming purposive behavior, an additional criterion is implicated if that behavior is to be viewed as intentional/voluntary. It seems that the counter-factual assumption of what might be called 'equi-origin' is central to claiming intentional behavior. As illustrated in Fig. 3 and Fig 4, intentional implies that the same or equivalent environment would have elicited different behavior and thus a different outcome had the individual's intention (as dispositional state) been different.

Figure 3: Equi-origin refers to an assumption that in exactly the same situation (S1), the agent (A) might have enacted a different means-end action (m) and thus determined a different end state (E) if the agent's dispositional state (D) (e.g. intent) had been different.

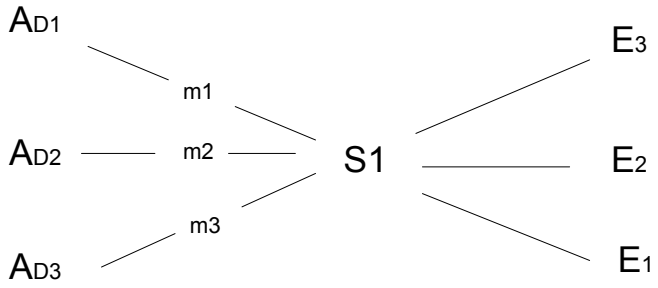
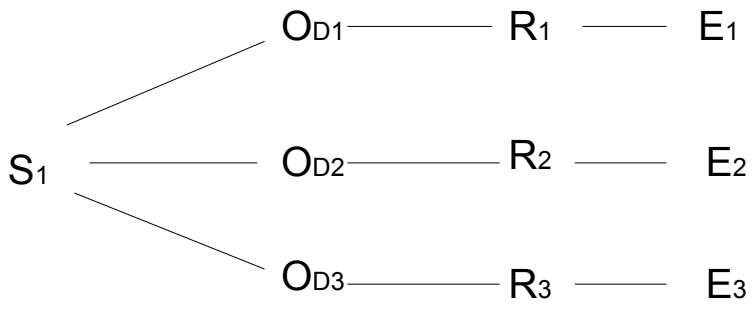


Figure 4: In classic behavioral terms, equi-origin can be rendered as the assumption that the exact same stimulus (S) might have elicited a different response (R) that would result in a different end state (E) if the organism's dispositional state (O_D) had been different.



In recent philosophical discussions, an important refinement has evolved in the concept of intentionality. I will refer to this refinement to make a distinction between what I will call weak and strong intentionality.

*Weak Intentionality***.** The weak form is much like the everyday use of the term. It implies purposive behavior that more or less meets the three criteria stated above: equifinality, persistence, and rationality. But in consideration of the equi-origin criterion, there is the added implication that the actor might have been differently disposed and thus acted differently in the given situation. So a claim of weak intentionality should provide evidence that the behavior has the following characteristics:

1) *Purposive* (meeting above criteria).

2) *Equi-origin*. The attribution of weak intentionality needs some evidence that in an equivalent situation the actor, in a different cognitive state, would behave differently.

Strong Intentionality. The strong form of intentionality is a claim that the behavior of a system is "about something," that it is not just the observable act but evidence of having a mental representation of something in mind such as a belief, a memory, a desire, or intent (Dennett, 1987; 1995, 1996). I believe that this special philosophical distinction can be nicely illustrated in the test situation that Piaget (1936/1963) devised for what he termed stage IV of object permanence. It is at this stage (about 8 months of age, but see Wishart and Bower, 1984) that human infants begin to show behavior directed toward a place where they have been shown an

attractive object become occluded by a cover (or barrier or container). Prior to this stage, they behave after watching an object be covered (or put in a container) as if the object no longer existed (at least so in terms of manual pursuit). When testing an infant for his/her capacity to search for a hidden object, one needs to demonstrate that successful behavior is not a simple interest in the occluder which, once removed, allows exposure to the object. Thus, the occluder should not be approached if it is not presently occluding the target object. Showing that the infant will only pick up the cloth (or cup) following its use to cover the target object but not otherwise is taken to imply that the act of removing the occluder was guided by an intention of obtaining the target object (and in that sense the object maintained its existence as represented in memory). The strong form of intention is implied by this act which is toward the occluder but is *about* the goal object (e.g. dependent on some level of representation of the goal object in memory).

Note that the implication of this test is to support the counter factual assumption that if the object had not been seen to be obscured by the occluder then the occluder would not have been approached. In other words, there is evidence that the same situation, the occluder is only approached when the subject has a memory of its use to cover the goal object. The attractiveness of the occluder requires the concurrent state of memory that it occludes the goal. The approach to the occluder is governed by the attractiveness of the goal. The approach is to the occluder but "about" the goal. So a claim of strong intentionality should provide evidence that the behavior has the following characteristics:

1) *Purposive:* meeting above criteria.
2) *Equi-origin:* meeting above criterion.
3) *Aboutness:* The attribution of strong intentionality, in the case of purposive action, needs some evidence that the dispositional contrast in the equi-origin behavior involves variation in representation of the agent's goal.

Intrinsic versus Derived Intentionality

When contemplating the existence of intentionality, some philosophers of cognitive science (e.g., Bowden, 1988; Dennett, 1996; Harnad, 1994; Keeley, 1994; Searle, 1980, 1992) have explored the issue of whether the intentionality is real and whether it is intrinsic to the behaving system or is derived from another source. Searle, in his now classic Chinese Room argument (Searle, 1980), argues that computer simulations of cognition do not possess intrinsic understanding and, without it, they can only be claimed to instantiate ersatz cognitive states. He argues that following a sequence of rule based decisions does not, in itself, provide understanding of what is being done. Thus, the fact that machines can be constructed to express purposive/intentional behavior (e.g. action of a thermostat or a self-guided missile, or a chess playing computer) does not mean that the intentionality is theirs.

Purposive machines can be viewed as simply expressing the derived purpose or intent of their designers or, in the case of computers, their programmers.

So then, it would seem that if one wanted to justify attributing intrinsic purposiveness and/or intentionality to a behaving system, one would need evidence that the behavior met the criteria of purposiveness and/or equi-origin plus some evidence that these features arose from the system's intrinsic nature and were not just derived (imported ready made) from some external source. However, meeting this criterion of intrinsic will require knowing something about the origin and any subsequent transformation of the behaving system in question. While this might seem a daunting requirement for the behavior of biological objects, advances in computer programming provide the possibility of having this information about behaving systems in experiments with "artificial life" (AL) (Dawkins, 1987; Levy, 1992). There has been philosophical discussion on the merits of viewing AL as an instance of real life versus, as with Searle's rejection of real computer cognition, only a model or simulation of life (Harnad, 1994; Olson, 1997).

The field of artificial life was later joined by researchers who use neural nets to explore issues of evolution and development of behavioral processes (Ackley and Littman, 1991; Miglino, Lund, and Nolfi, 1995; Nolfi, Elman, and Parisi, 1994; Parisi, Cecconi, and Nolfi, 1990). Studying the adaptation of neural net structures is attractive on at least two grounds. On the one hand, it holds the promise of bottom-up explanations for the origins of adaptive behavioral functions. The functions emerge from the effects of experience on the neural structure of the agent. In addition, the manner in which experience is brought to bear on adaptive change in neural net structure (e.g. change in pattern of connection weights) is attractively natural. It is a direct effect of the experience produced as the net's behavior alters its contact with the environment. Once a "life" begins, the net's behavior determines its experience in its environment and that experience will be the source of any adaptive change it undergoes within a lifetime and/or across evolutionary generations. When a capacity emerges in this fashion, there is a basis to claim that the capacity has its origins in the process and is not simply the product of a programmer's constructive inventiveness. There is a basis to claim the capacity is intrinsic to the evolved system and not derived from similar capacity possessed by the programmer.

Recent work with simple artificial organisms, controlled solely by neural net structure, has introduced evidence that supports the claim for intrinsic possession of at least purposiveness within artificial life (Parisi et al, 1990; Nolfi et al, 1994). Nolfi and Parisi (1997) refer to these structures as artificial life neural nets. We will use the now more common term "agent" (as in "autonomous agent") when referring to these structures. Nolfi et al (1994) report that in just a hundred generations of weight pattern evolution, the evolved agents became very efficient pursuers of benefit. In this and other studies, agents display their evolved purposiveness by the fact that regardless of how they are initially placed in an environment of randomly distributed benefit and harm, they efficiently move to the benefit and avoid the harm. They would seem to meet the criteria of equfinality, persistence, and rationality. The

emergent nature of the purposive behavior would seem to fulfill the criterion of its being intrinsic to the evolved weight pattern that controls the agent's behavior. Given this evidence for purposiveness emerging in the evolution of agents endowed with a relatively simple structure (their agents were given 4 hidden and 2 recurrent motor nodes), the question arises as whether some measure of intentionality would also emerge within artificial life and, if so, at what level of structural complexity.

A Study of Emergence of Intentionality

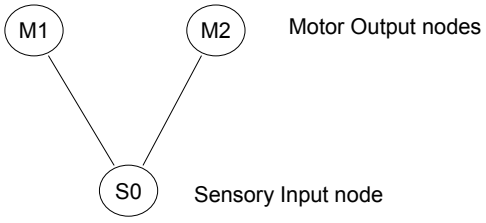
The present investigation of intentionality used procedures similar to those used by Parisi, Nolfi and their colleagues. The basic features of the computer program are presented below.

The Study

Environment. The environment in which an agent "lived" was a 10 by 10 toroidal matrix, analogous to being on a checkerboard surface of a sphere and thus having no boundaries. At the start of each generation, an agent was randomly placed and randomly oriented on the matrix. The remaining 99 cells of the matrix were either empty or held one of thirty randomly placed container objects. These containers remained stationary and the agent was free to move about at the rate of one step per moment of life as depicted in Fig. 7. Whenever the agent entered a cell that held a container, the container disappeared and the agent was recorded as consuming its state of being a benefit or harm (i.e. similar to the classic computer game Pac-Man).

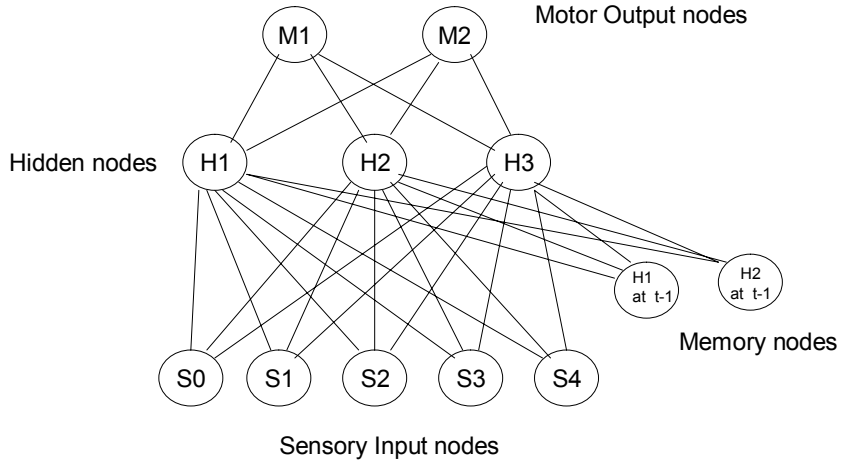
Initial Agent Structure. The beginning structure of the agents is illustrated in Figure 5. An agent had one "sensory" input node that was sensitive to the sensory value of the position the agent occupied on the matrix. It also had two "motor" nodes that determined whether the agent stepped forward one cell (when both output nodes were activated), turned 45 deg to the left (when one output node was activated), turned 45 deg to the right (when the other node was activated), or remained in its present position and orientation (when neither was activated) (as per Nolfi et al, 1994).

Figure 5: Structure of all agents at start of evolution.



Evolvable Range of Agent Structure. Although the initial structure was limited to one input node and two output nodes as depicted in Fig. 5, the potential for greater complexity was provided for in the computer program. Sensory input nodes could be added by mutation. Nodes for a hidden layer could be added and these hidden layer nodes could be connected to associated recursive nodes (i.e. memory nodes as per Elman, 1990) by mutation. Once added, the sensory, hidden, or memory nodes could also be eliminated by mutation. Fig. 6 is an example of an evolved agent with a neural net structure involving 5 input nodes, 3 hidden nodes, 2 memory nodes, and the original fixed 2 motor output nodes. Sensory node addition was such that added nodes were sensitive to progressively greater distances on the matrix in a direct line in front of the agent. The first added sensory node (S1 in Fig. 6) was sensitive to the matrix position one step in front of the agent, the second added node was sensitive to the position two steps in front of the agent, and so on for any additional nodes. When mutation eliminated a node (sensory, hidden, or memory) it eliminated the most recently added node of its type. A memory node could be eliminated either directly or as a result of its associated hidden layer node being eliminated.

Figure 6: Example of agent that has evolved 4 additional sensory nodes, 3 hidden nodes and 2 associated memory nodes.



It should be clear that the hidden and memory nodes are not morphological mutations of the initial sensory and motor nodes. However, this fact does not reduce

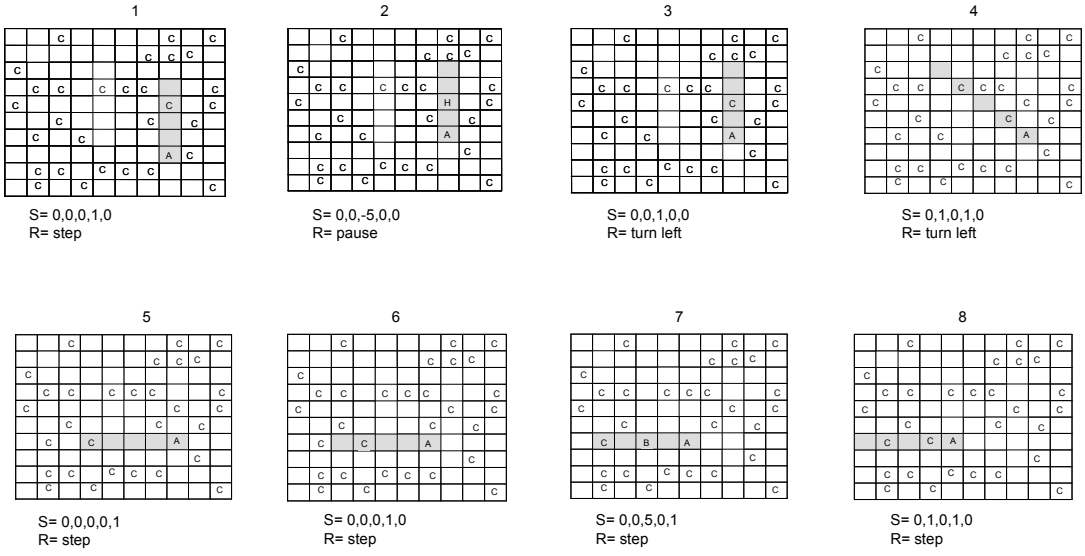
the study to a trivial exercise regarding its objectives. Recall that the focal questions concern the mechanistic complexity required to support purposive and intentional behavior and whether such complexity could be formed by mutation in a Darwinian evolutionary context. In one sense, all neural units are the same in that they share the property of being capable of being activated and of influencing the activation of other nodes. Their differentiation of function does not derive from their individual structure so much as their placement in the cluster of nodes and the resultant interconnectedness afforded by that placement. For example, functional memory does not appear simply as a direct effect of a memory node being inserted into an agent's neural net structure by mutation. Functional capacity arises by virtue of the number of nodes, their structural placement, and the evolved weight pattern on the interconnections between nodes that ultimately governs behavior (activations of motor nodes) in relation to sensory experience (activation of sensory nodes).

Evolutionary selection process. As in the cited studies by Nolfi and Parisi and their colleagues, evolution in the present study was initiated by producing a random set of connection weights for each of 100 agents. These 100 agents were run individually for their life cycle. Life lasted 200 time units (Nolfi et al used 5000 time units). The best 20 of this first generation (determined by comparing the number of benefits they consumed minus the number of harms consumed) were subjected to mutation. Mutation involved randomly selecting a few connection lines (5% on average) and randomly increasing or decreasing their connection strength (by a magnitude ranging from -1 to +1). Thus, the pattern of connection weights could be slightly altered and this might change some aspect of the agents' responsiveness.[2] Similarly, the modification of neural net structure in terms of sensory, hidden, and memory units was altered by a mutation rate of .5 and a magnitude ranging from -2 to +2 (values chosen in respect of units, unlike weights, being added or subtracted as integers). In this manner, 5 offspring were produced from each of the 20 best agents of the first generation. The resultant 100 agents comprised the second generation. The procedure was repeated for 4000 generations. The basic question is this: Can random variation design a structure and associated pattern of connection weights that will control behavior in a manner that meets our criteria of purposiveness and some level of intentionality - when the only constraint on that design process is a Darwinian selection on the basis of net benefit obtained in a life time?

Containers of benefit or harm. In order to establish an evolutionary pressure for strong intentionality, the environmental objects were given a surface value that did not disclose whether they held benefit or harm. That is, an input node received sensory input associated with the presence of a container on the matrix position to which the node was sensitive. The sensory value of the container was 1 regardless of whether consuming the container would increase or decrease the net's competitive score. An empty matrix position had a sensory value of 0. In order to provide a potential basis for adaptive discrimination of a container of harm versus a container of benefit, the relevant information was provided momentarily at a distance. In the present study, when a container was 2 steps in front of the agent, it exposed a sensory

value (+5) indicating it contained benefit or a sensory value (-5) indicating it contained harm. This exposure was for 1 moment of life. The container then resumed display of its surface value of 1. Thus the decision to enter a container or not could now be related to its contents given that (and this was the evolutionary challenge) the agent controlled that decision by a memory related to the prior exposure of whether it held benefit or harm (in terms of human memory research, this could be viewed as working memory). If the agent entered a container only when previously informed that it marked a position containing benefit, and not in the absence of such information, then a case might be made for at least a primitive form of intentionality.

Figure 7: Eight successive moments of life of agent (A) embedded in matrix with 30 containers (C). Shaded cells show the five cells to which agent's 5 sensory nodes are sensitive. Sensory input (S) and agent's reaction (R) are shown for each moment of life. Contents of harm (H) and benefit (B) are exposed on 2nd and 7th moments respectively. Note that criterion of equi-origin is met by contrast in reactions on 4th and 8th moments (turning versus stepping respectively) although agent has received the same sensory input (0,1,0,1,0).



At the beginning of a life cycle, half of the 30 containers held benefit and half held harm. When a container was exposed to the agent's input node that sensed the cell two steps in front of it (assuming the agent had evolved such a node), then the container's contents were exposed. This exposure could happen as a result of the agent stepping toward a container (from three cells away as depicted in Fig. 7) or by turning in a situation where it happened that a container was two cells away and now was brought into sensory contact with the input node sensitive to that distance (S2 in Fig. 6).

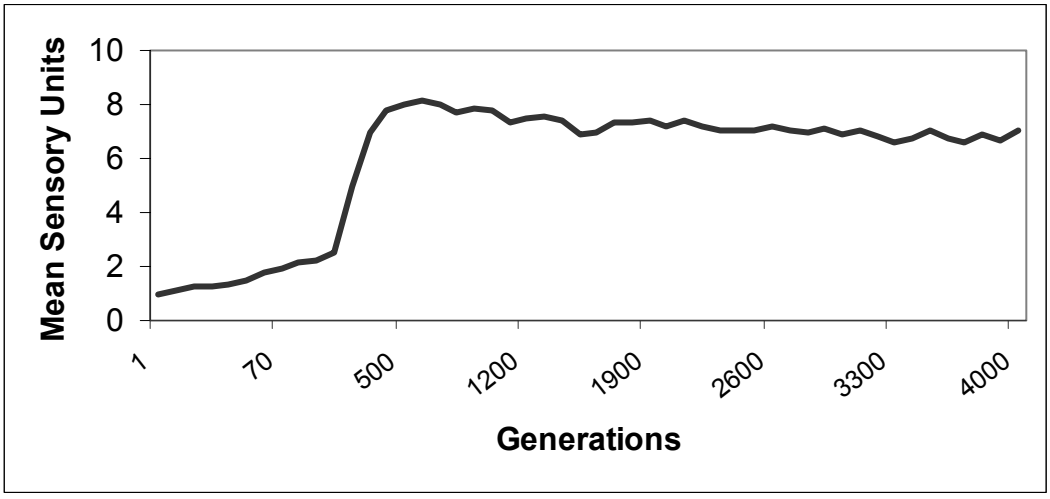
At the beginning of an agent's life, the positions of the 30 containers and of the agent were randomly specified as was the agent's initial facing direction. Thus the unique starting conditions for distribution of containers were the combinations of 100 things taken 30 at a time. For each of these, there were 70 positions times 8 orientations that specified the options for the starting placement of the agent. In sum, the unique starting conditions number in the range of 10^{100}. Thus it is very unlikely that any two agents begin life in exactly the same relationship to their environments. If one adds consideration of the variation in contents of a container being benefit or harm, the estimate of likelihood is even less. In order to control for any potential effects of the starting conditions, the evolution was repeated 25 times providing a total of 500 agents for evaluation.

Results

Evolution of Structure

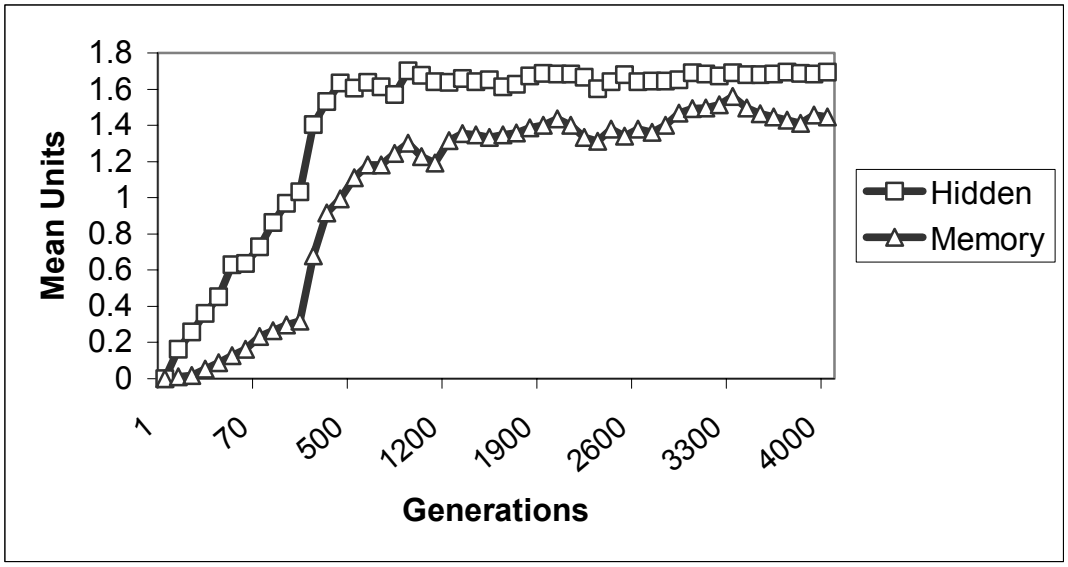
Sensory units. The agents' environment was such that any discrimination of the positions of benefit and harm would require possession of at least three sensory units. Thus, at least two would need to be added by evolution before an agent could forage effectively. How many additional sensory nodes would add to its competitive advantage was not clear on a priori grounds. As shown in Fig. 8, sensory units evolved relatively quickly to about 8 on average by 600 generations and then settled back to about 7 by 3000 generations.

Figure 8: Evolution of sensory units in agent structure.



Hidden and Memory units. Agents began evolution with no hidden or memory units. It was expected that some hidden units would be necessary for successful foraging given that some memory would be needed and any memory units would be associated with a subset of hidden units. However, what was not clear was how many hidden units would be required. Fig. 9 shows the evolution of hidden units and their associated memory units. Hidden units evolved to an average of about 1.6 units by 400 generations while memory units kept increasing until about 2 to 3 thousand generations where they averaged about 1.45 units.

Figure 9: Evolution of hidden and memory units in agent structure.

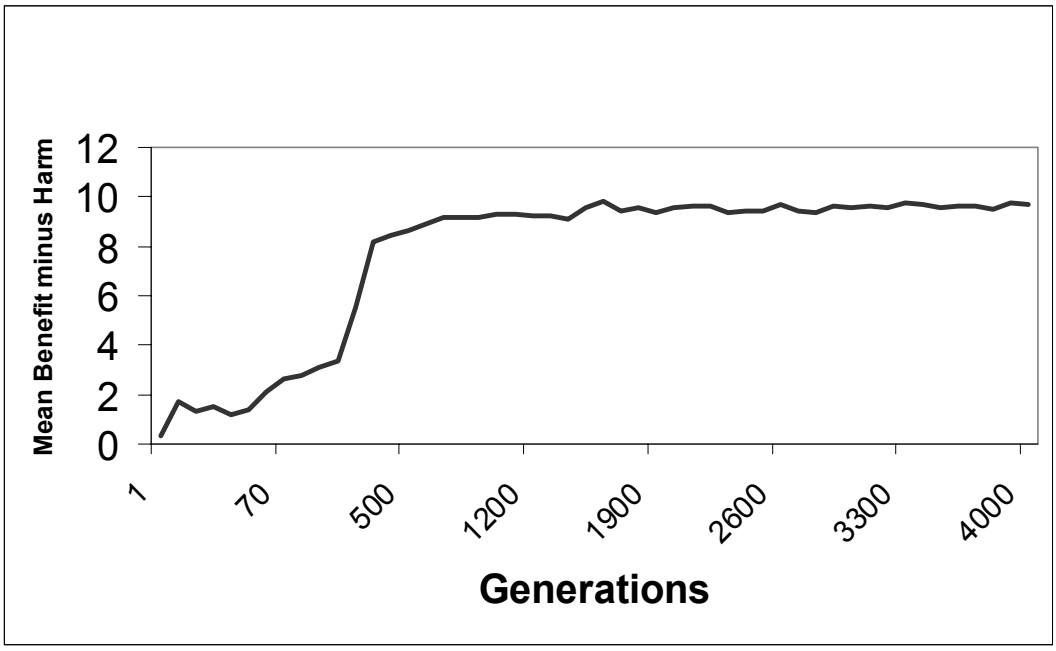


Evolution of Function

Assessment of Purposiveness. The agents rapidly evolved behavior that would appear to meet the three criteria of purposiveness. Equifinality was displayed by the fact that regardless of how the evolved agent was initially oriented and placed among a random distribution of benefits, it eventually proceeded to search and find almost all of them (93% on average) within its limited lifespan. Persistence was displayed by the continuous search, especially as the remaining benefits became fewer over time. All agents evolved a disposition to approach containers that were beyond 2 steps in front of them as they came within their evolved depth of perception. Many evolved a disposition to change directions after a few steps revealed no container within their sensory inputs. Rationality was displayed by the selective discrimination of benefit and harm, approaching the former while avoiding the latter, and the efficiency displayed in doing so. The evolution of foraging success (containers of

benefit entered minus containers of harm entered) is presented in Fig. 10.

Figure 10: Evolution of foraging success defined as benefit minus harm containers consumed.



A specific test of purposiveness. In an effort to compare the evolution of purposiveness with the additional attributes of intentionality, a simple test of purposiveness was devised based primarily on the criterion of rationality. The specific structure and weight pattern of each of the 500 evolved agents were exposed to two sets of test sequences of sensory inputs as illustrated in Fig. 11. Each sequence of the test set was composed of 9 successive input patterns. The first six inputs were used to assure a controlled background experience for these agents that varied in sensory and interior structure. The input sequence began with the stimulus array generated by a container being 9 steps in front of the agent. This was followed by the stimulus array generated by a container being 8 steps in front, and so on until the container was 3 steps in front. After this point, the sequences of the test set were distinct. In one sequence, the agent was exposed to the container's contents 2 steps away being a benefit and in another sequence to the contents being harm. To pass this test of purposiveness, the agent needed to step forward twice following exposure to the container of benefit but not to step forward twice when exposed to harm. The test was given to each of the 500 selected agents at the end of each generation. The evolution of purposiveness is presented in Fig. 14. All 500 selected agents passed

this test from the 300th generation onward.

Figure 11: Test of purposiveness where potential sensory stimulus is introduced to agent (A) representing container (C) being progressively nearer with benefit displayed at distance of two cells in Part 1 and harm displayed in Part 2 as described in text.

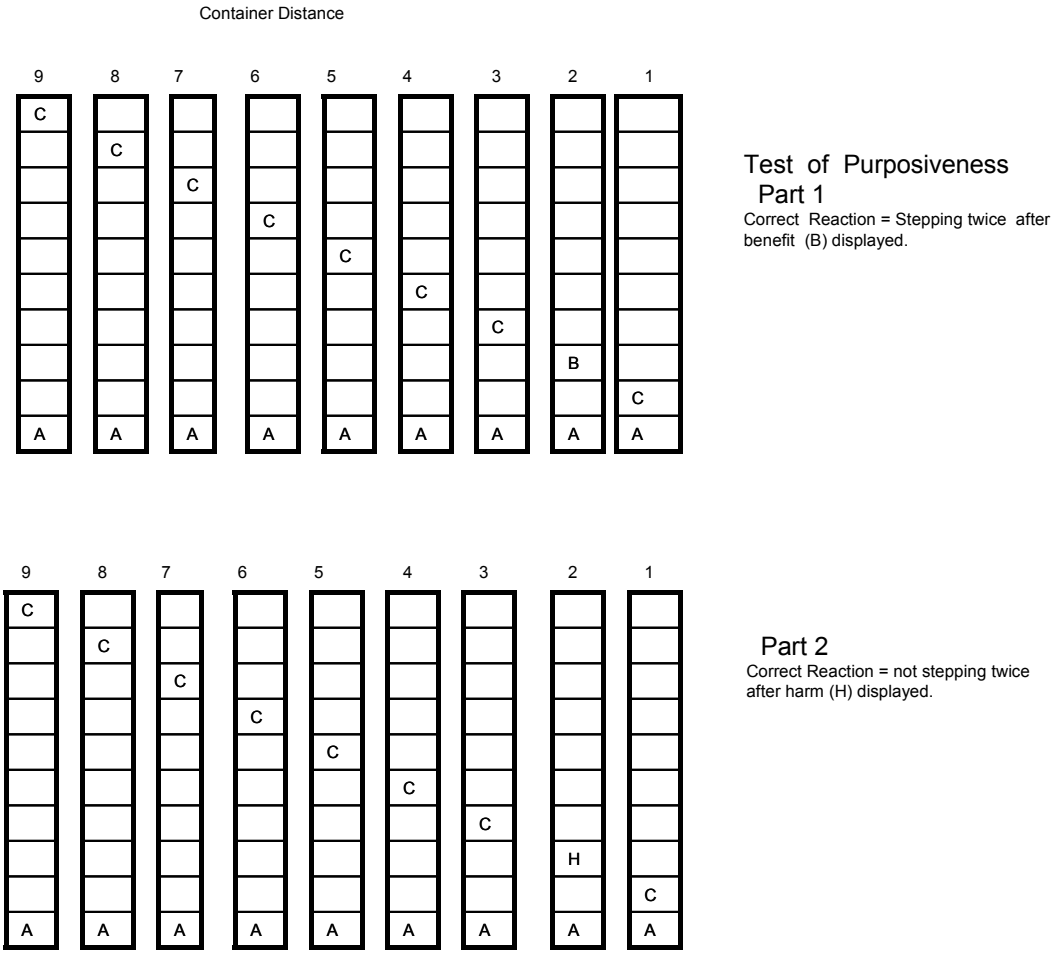


Figure 12: Test of equi-origin and weak intentionality where potential sensory stimulus is introduced to agent (A) representing container (C) being progressively nearer with benefit displayed at distance of two cells in Part 1 and harm displayed in Part 2. Same sequence but different scoring than for simple purposiveness as described in text.

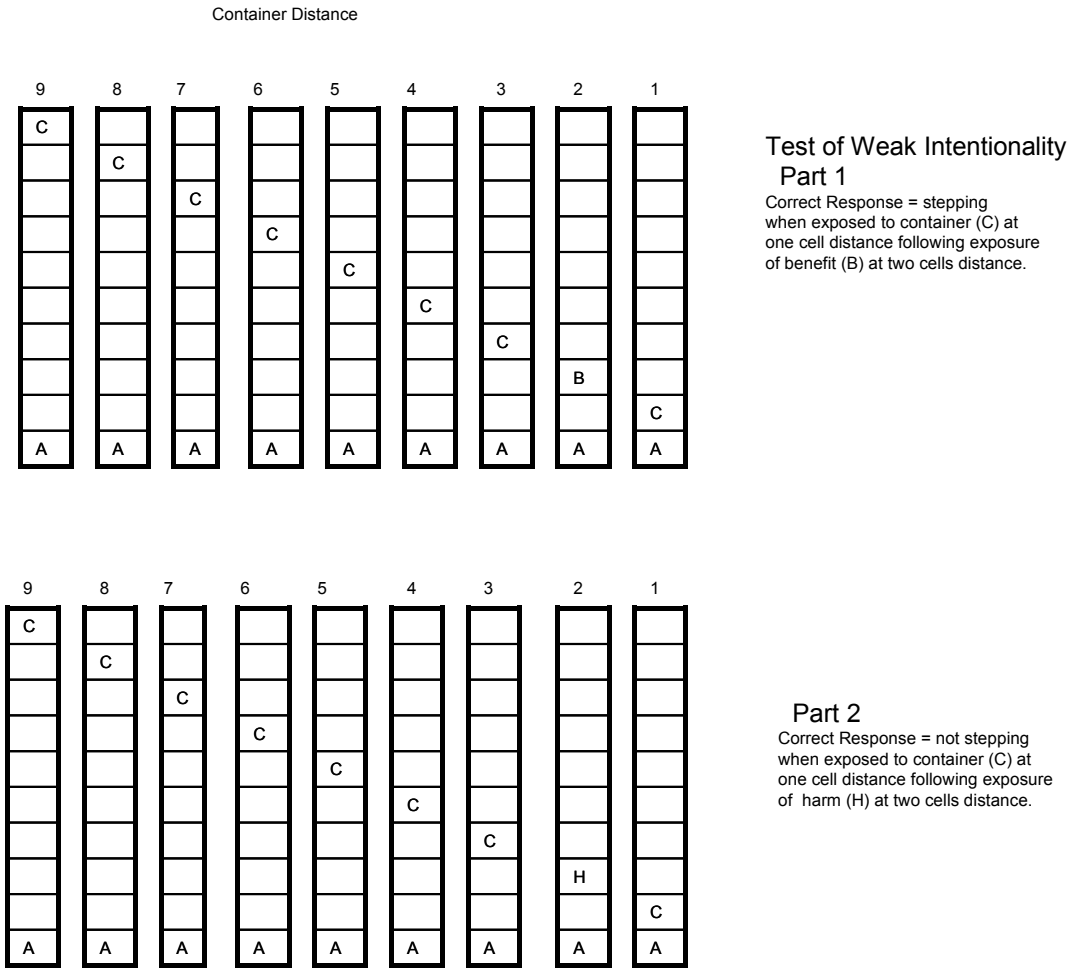


Figure 13: Test of strong intentionality where potential sensory stimulus is introduced to agent (A) representing container (C) being progressively nearer with content of container not displayed at distance of two cells. If memory of benefit has been basis of action for agents passing test of equi-origin, then they will not step when container is at distance of one cell as described in text.

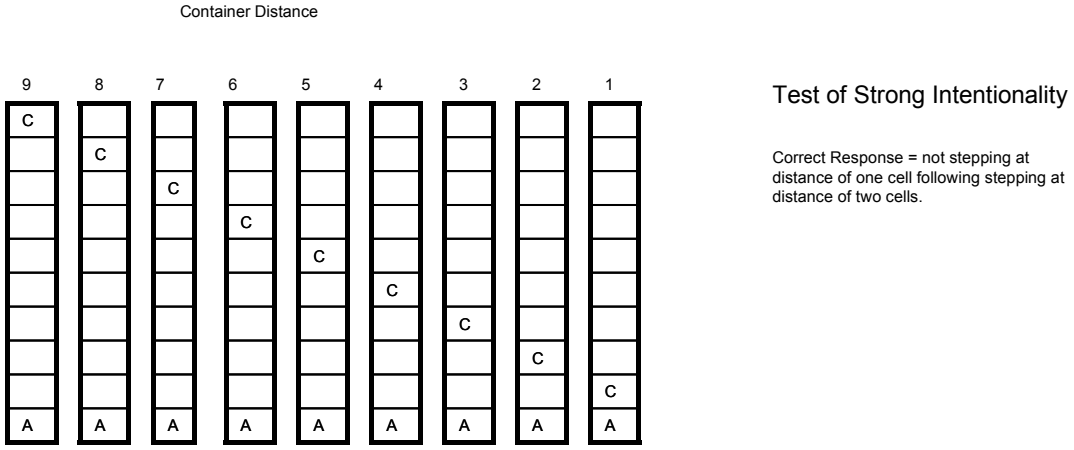
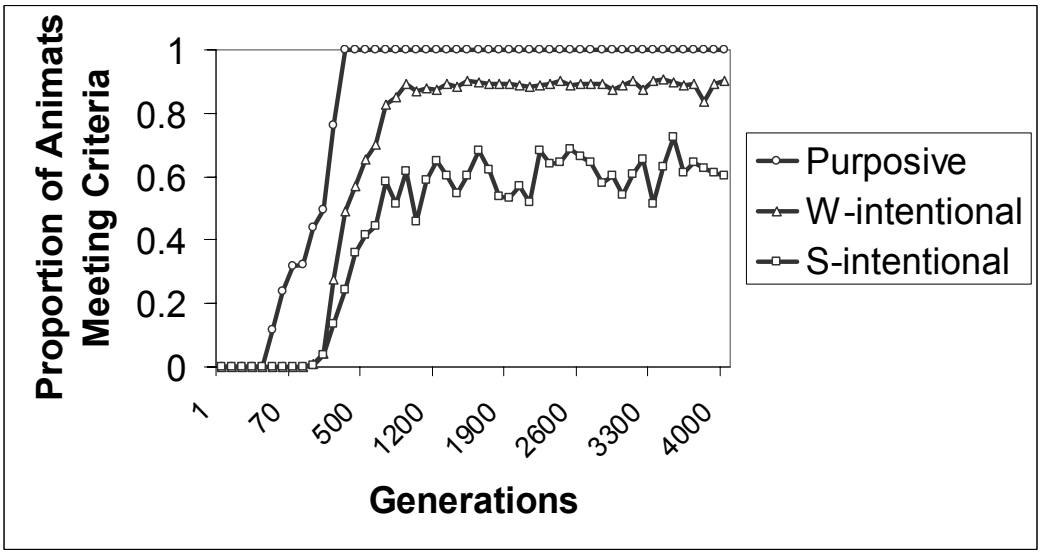


Figure 14: Proportion of the 500 agents that met criteria for purposiveness, weak (w-) intentionality, and strong (s-) intentionality.



Assessment of intentionality. While an evaluation of evolved purposiveness was rather clear from the overall foraging success displayed in Fig. 10, a similar inference about intentionality is not so easy. For example, as in the specific test for simple purposiveness, a certain amount of success at obtaining benefit and avoiding harm might be accomplished by a simple pattern of stepping forward under all

sensory input conditions except the input associated with harm and turning in that case. But such a response pattern would not meet the criterion of equi-origin for intentionality wherein the exact same sensory input requires appropriately distinctive responses on separate occasions. A special test was devised to assess the agents with regard to the equi-origin criterion.

A specific test of equi-origin and weak intentionality. The test involved a slight but essential variation in scoring agents' behavior on the stimulus sequence used for the test of purposiveness. In this test, whether or not the agent turned or paused when exposed to harm at 2 steps ahead, the next exposure was of a container 1 step ahead (e.g., as could occur during evolution after a turn as in Fig. 7). Agents that stepped forward to the container, after either turning or stepping forward in response to exposure of harm, were scored as failing the test. Although the appearance of the container at 1 step ahead following a pause was an "unrealistic" event, agents that stepped forward to the container, following a pause to the harm, were scored as failing the test. In the sequence of exposure of benefit, agents who did not step forward twice in succession were scored as failing the test. Thus, for both of these sequences, the final input was of the container (a sensory value of 1) at one step in front of the agent. At this final input, evidence of successful adaptation demands two distinctly different responses to the exact same environmental stimulus display. An agent is judged to have passed the test of weak intentionality if it steps forward to the container after the prior exposure to its associated benefit, and not stepping forward after the prior exposure to its associated harm. As shown in Fig. 10, at the end of 4000 generations 90% of the agents passed this test.

Although the assessment of equi-origin in the weak test of intentionality provides the requisite evidence of at least primitive intentionality, it does not speak to the question of how the adaptive dispositional variation is formed. It is clear that the agent is disposed to step forward to the container after the prior exposure to benefit and not to do so after prior exposure to harm. However, a claim that the disposition to act is based on some representation of the container's contents, that it is an act towards the container but "about" its contents, requires additional evidence.

Evolving a responsive pattern of turning when exposed to harm in the container and stepping forward when exposed to benefit would, in itself, not require any reference to past experience (indeed, as visible in Fig. 10, all agents developed this responsive pattern in the early stages of evolution as their behavior was becoming purposive). Yet, once evolved, this pattern might serve as a basis for adapting to the challenge of responding appropriately when faced with a container 1 step away. Remembering that the prior act was stepping forward could be used to reactivate stepping and remembering that the prior act was turning away (or pausing) could be used to reactivate turning (or pausing). Thus, an alternative possibility in the present case is that the agent does not remember the contents of the container (i.e. does not possess an activation pattern in its memory nodes that is unique to the prior stimulus), but rather remembers its past act of stepping forward or not (e.g. possesses an activation pattern in its memory nodes that is unique to the requisite for causing the

output motor nodes to activate the pattern for stepping forward). In such a case, the agent's adaptive dispositional variation could not be said to be guided by a representation of the container's contents but simply reflective of the agent's memory of prior behavior.

Another alternative possibility is that the agents evolve a pattern wherein they always step forward toward a container unless they possess a memory of exposure to harm on the preceding moment of life. While this basis of dispositional variation is in reference to the container's contents, i.e. do not step following evidence of harm, it does not seem to justify a claim that when stepping toward a container of benefit, when it is one step away, the agent is performing an act that is toward the container but "about" the container's contents of benefit. At best, one might claim it is about the container not possessing harm, but that could be claimed about stepping into empty space.

Due to the agent's directional perceptual system in which it evolves sensory sensitivity to some depth in front of it, there should be some competitive advantage to guiding behavior by memory of perceived benefit versus memory of perceived harm. Were an agent to evolve a response of turning (or turning after pausing) when perceiving harm in the container two steps in front of it, there is a possibility of being faced with another container just one step away. If it turns again, it still faces that possibility again. Guidance by reference to past action could lead to dysfunctional spinning in place. Likewise, as memory fades, guidance by reference to a preceding exposure to harm could lead to risky entry into containers with unknown contents. Alternatively, guidance by reference to the remembered perception of benefit should present virtually no risk. Thus, it would seem that there would be some environmental pressure favoring the evolution of behavioral control in relation to a memory representation of the agent's purposive goal.

A test to assess strong intentionality. An additional test sequence was applied to the agents to assess how they were passing the equi-origin test. As depicted in Fig. 13, the test of strong intentionality involved the same initial 7 exposures as in the test for purposiveness. Then, when the container was two steps away, the point at which contents would normally be displayed, the agent was exposed simply to the container's surface value. The final exposure of the test sequence was once again the surface value of the container one step away. At the end of 4000 generations, 98% of agents stepped forward in response to the non-informative display when 2 steps away from the container. Now at one step from the container, if guided by memory of their past act (past step forward leading to step again and past turning (or pausing) leading to turning (or pausing) again) or by memory of prior exposure to harm (past harm leading to turning (or pausing) and absent that, stepping forward), then they should now step forward. If, on the other hand, agents were guided by memory of exposure to benefit, they should not step forward. Fig. 10 displays the results of this test. Notable is the fact that 67% of the agents who passed the equi-origin test of weak intentionality, and thus displayed their disposition to step forward when exposed to benefit, now did not step forward without that memory of benefit.

General Discussion

The data presented here support a claim that at least a primitive form of strong intentionality can rapidly emerge in a Darwinian process of self-design. After four thousand generations, the agents met the criteria of equifinality, persistence, and rationality in their foraging for benefit while avoiding harm. The vast majority also evolved a structure for and a systematic use of memory. When confronted by the opportunity to enter the environmental position of a container (and thereby obtain its contents), they went "beyond the information given" in the immediate sensory input designating a container and governed their behavior on the basis of their memory. Two thirds of these were guided by memory of the attractive contents of the container. Thus it would seem that in the philosophical sense of strong intentionality, the behavior of a large majority of these agents had the right kind of "aboutness" about it. It was about the objective of their purposive behavior and not about the behavior per se. It also seems clear that this adaptive use of an internal representation of the container's contents was not derived from some external source. Rather, it emerged as intrinsic to the Darwinian self-design of these autonomous agents. The major philosophical implication I draw from this finding is that it would seem to support Dennett more than Searle in their argument over possible evolutionary origins of real intentionality and the associated question of how intention might emerge from the interaction of lower order events that do not possess the attribute (Dennett, 1995).

The fact that such a high percentage of agents evolved the capacity to pass the tests of intentionality is a notable outcome of the present study. What is perhaps more surprising, however, is the simplicity of the agents' internal neural net structure that evolved in support of that capacity. The average internal structure involved only 1.7 hidden nodes and 1.4 associated memory nodes. Fifty-six percent of the agents that passed the test of weak intentionality and 41% that passed the test of strong intentionality possessed just 1 hidden node and an associated 1 memory node. It was predictable that agents would need to evolve at least 2 sensory nodes (for a total of three) and some level of memory function in order to pass the tests of intentionality. That was because information about contents of containers was provided at a distance of two steps away. However, it was not obvious how complex an internal (hidden and memory) structure would be required, if it were even possible to design in this evolutionary setting. The present results would seem to indicate that when behavior affects reproductive fitness of a mutable interconnected modular system, purposiveness and intentionality can be expected to emerge as elementary features of the evolving entity. It should be clear that there is no implication that the control system that evolved was anything beyond what Sloman (1999) has termed a "reactive subsystem," but that should not detract from an appreciation of the historical salience of the control achieved.

The results of the study support the claim that relatively simple agents can and will evolve some degree of the strong form of intentional behavior. Moreover, it seems reasonable to claim that the emergent capacity is intrinsic to the agents that

manifest it. Darwinian evolution fashioned the designs of neural structure and the weight pattern that governed the adaptive behavior of the successful agents. The composition of supportive structure and weight patterns were not designed by a programmer and built into the agents. They designed themselves (at least in terms of committed numbers of neural units in the sensory, hidden, and memory layers, and the specification of weights on all their interconnections) by adaptive modification resulting from random mutation. It was not obvious ahead of time that a Darwinian search through "potential design space" (Dawkins, 1987; Dennett, 1995, 1996) would lead to the emergence of either weak or strong intentionality.

There would seem to be at least two implications of this finding for psychology. The first has to do with the development of intentionality in the child. How can a simple neural net display the essence of stage IV object permanence, given that human infants develop for 6 to 8 months before they display the capacity? At least three answers seem worth considering. One possibility is that it may be for the reasons Hebb (1949) suggested for explaining the relative incompetence of human infants compared to the sensory-motor capacities of the young of other species, namely that the large size of the association area relative to the sensory and motor areas in the human brain is an initial handicap for learning. Another possibility is that it may be linked to a slow maturation of memory capacity in human infants (Case, 1985). A third possibility is that the human infant may be coping with a fundamentally more difficult issue, namely that of the permanence of numerical identity (e.g. that toy, the mom) while the neural net may only be coping with a problem of permanence of equivalence (e.g. a benefit, a harm)(see Meltzoff and Moore, 1998).

The second implication I would propose is that the apparent speed with which purposiveness and intentionality can emerge in the evolution of simple life forms lends some support to the idea that infants may be pre-organized to discriminate these attributes in other living objects. Thus, Csibra and Gergely's (1998) conception of very young infants employing the ''teleological stance" and later the "intentional stance," which are roughly equivalent to discriminating purposiveness and then intentionality in their environment, seems more plausible as perceptual capacities organized so early in life given that purposiveness and intentionality are likely to be primitive aspects of virtually any behaving organism in the infant's environment.

Notes

1. The author is indebted to Earl Wagner for his writing of the programs used in this study and to the University of California's non-competitive research grants that financed the computers employed.

2. The output nodes and their motor effects were virtually the same as used previously by Nolfi, Parisi and their colleagues wherein the net could step forward, turn right or left 45 deg, (Nolfi et al, 1994 used 90 deg.) or stay still as the result of

activation transmitted through the net on each time unit of a life. All hidden and output units were also connected to a bias unit that had a constant activation of 1 which is not shown in figs. 5 and 6. Activation of hidden units and output units was by the sigmoid function $\mathbf{a = 1 / (1 + e^{-Sw})}$ where Sw is the sum of weight x activation of all input connections to the unit. The output units were squashed to one of two states: on if activation exceeded .5 or off if activation was .5 or less. As noted in the text, the motor action rules were the same as used by Nolfi and Parisi, namely step forward if both output units are on, turn 45 deg. left if one unit is on, turn 45 deg right if the other is on, and pause if both are off. It is perhaps worth noting that the results do not depend on this particular motor rule assignment nor do they depend on the use of positive numbers for sensory designation of benefit and negative numbers for designation of harm.

Received 29 April, 2004, Revision received 1 November, 2004, Accepted 7 December, 2004.

References

Ackley, D. and Littman, M. (1991). Interactions between learning and evolution. In Langton, C. G., Taylor, C., Farmer, J. D. and Rasmussen, S. (Eds.), *Artificial Life II*. (pp. 487-502). Redwood City, CA: Addison Wesley.

Barnes, J. (1984). *The Complete Works of Aristotle: The Revised Oxford Translation*. Princeton, N.J.: The Princeton University Press.

Barnes, J. (1995). *The Cambridge Companion to Aristotle*. New York: Cambridge University Press.

Bennett, J. (1976). *Linguistic Behaviour*. Cambridge: Cambridge University Press.

Boden, M. A. (1988). *Computer Models of Mind*. New York: Cambridge University Press.

Case, R. (1985). *Intellectual Development: Birth to Adulthood*. Orlando, FL.: Academic Press.

Csibra, G., and Gergely, G. (1998). The teleological origins of mentalistic action explanations: A developmental hypothesis. *Developmental Science*, *1:* 255-259.

Csibra, G., Gergely, G., Bíró, S., Koós, O. and Brockbank, M. (1999). Goal attribution without agency cues: The perception of 'pure reason' in infancy. *Cognition*, *72*: 237-267.

Dawkins, R. (1987). *The Blind Watchmaker*. New York: Norton.

Dennett, D. C. (1987). *The Intentional Stance*. Cambridge, MA.: MIT Press.

Dennett, D. C. (1995). *Darwin's Dangerous Idea*. New York: Simon & Schuster.

Dennett, D. C. (1996). *Kinds of minds*. New York: Basic Books.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*: 179-211.

Gergely, G. and Csibra, G. (1997). Teleological reasoning in infancy: The infant's naive theory of rational action: A reply to Premack and Premack. *Cognition*,

63: 227-233.

Gergely, G., Nadasdy, Z., Csibra, G. and Biro, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, *56*: 165-193.

Harnad, S. (1994). Artificial life: Synthetic vs. virtual. In Langton, C. G. (Ed.), *Artificial Life III* (pp. 539-552). Reading, MA: Addison Wesley.

Hebb, D. O. (1949). *The Organization of Behavior*. New York: John Wiley & Sons.

Heider, F. (1958). *The Psychology of Interpersonal Relations*. New York: John Wiley & Sons.

Jennerod, M. (1985). *The Brain Machine: The Development of Neurophysiological Thought*. (Urion, D.; trans.) Boston: Harvard University Press.

Keeley, B. L. (1994). Against the global replacement: On the application of the philosophy of artificial intelligence to artificial life. In Langton, C. G. (Ed.), *Artificial Life III* (pp. 569-587). Reading, MA: Addison Wesley.

Kelemen, D. (1999). Beliefs about purpose: On the origins of teleological thought. In Corballis, M. and Lea, S. E. G. (Eds.), *The Descent of Mind: Psychological Perspectives on Hominid Evolution* (pp. 278-294). Oxford: Oxford University Press.

Levy, S. (1992). *Artificial Life*. New York: Pantheon Books.

McDougall, W. (1929). *Modern Materialism and Emergent Evolution*. London: Methuen & Co.

Meltzoff, A. N. (1995). Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology*, *31*: 838-850.

Meltzoff, A. N. and Moore, M. K. (1998). Object representation, identity, and the paradox of early permanence: Steps toward a new framework. *Infant Behavior & Development*, *21:* 201-235.

Miglino, O., Lund, H. H., and Nolfi, S. (1995). Evolving mobile robots in simulated and real environments. *Artificial Life*, *2:* 417-434.

Moore, M. K. and Meltzoff, A. N. (1999). New findings on object permanence: A developmental difference between two types of occlusion. *British Journal of Developmental psychology*, *17:* 563-584.

Nolfi, S., Elman, J. L., and Parisi, D. (1994). Learning and evolution in neural networks. *Adaptive Behavior*, *3:* 5-28.

Nolfi, S. & Parisi, D. (1997). Neural networks in an artificial life perspective. In Gerstner, W., Germond, A., Hasler, M. and Nicoud, J. (Eds.). *Artificial Neural Networks*. Berlin: Springer-Verlag.

Olson, E. T. (1997). The ontological basis of strong artificial life. *Artificial Life*, *3:* 29-39.

Parisi, D., Cecconi, F. and Nolfi, S. (1990). Econets: Neural networks that learn in an environment. *Network*, *1*: 149-168.

Piaget, J. (1936/1952). *Origins of Intelligence in Children*. New York: Norton.

Rovee, C. K. and Rovee, D. T. (1969). Conjugate reinforcement in infant exploratory behavior. *Journal of Experimental Child Psychology*, *8*: 33-39.

Rovee-Collier, C. K., Morrongiello, B. A., Aron, M., and Kupersmidt, J. (1978). Topographical response differentiation and reversal in 3-month-old infants. *Infant Behavior and Development*, *1*: 323-333.

Russell, S. J. and Norvig, P. (1995). *Artificial Intelligence: A Modern Approach*. Upper Saddle River, NJ: Prentice Hall.

Searle, J. R. (1980). Minds, brains, and programs. *The Behavioral and Brain Sciences*, *3*: 417-457.

Searle, J. R. (1992). *The rediscovery of the mind*. Cambridge, MA: MIT Press.

Skinner, B. F. (1938). *The Behavior of Organisms*. New York: Appleton-Century-Crofts.

Skinner, B. F. (1953). *Science and Human Behavior*. New York: The Macmillan Co.

Stein, J. and Flexner, S. B. (1984). *The Random House Thesaurus: College Edition*. New York: Random House.

Tolman, E. C. (1925) Behaviorism and purpose. *Journal of Philosophy* As reprinted in Tolman, E. C. (1966). *Behavior and Psychological Man* (pp. 32-37). Berkeley, CA: University of California Press.

Tolman, E. C. (1932/1967). *Purposive Behavior in Animals and Men.* New York: Appleton-Century-Crofts.

Watson, J. B. (1930/1957). *Behaviorism*. Chicago: Phoenix Books, University of Chicago Press.

Watson, J. B. and MacDougall, W. (1928). *The Battle of Behaviorism*. London: K. Paul, Trench, Trubner & Co., Ltd.

Watson, J. S. (1966). The development and generalization of `contingency awareness' in early infancy: Some hypotheses. *Merrill-Palmer Quarterly, 12*, 73-94.

Watson, J. S. (1979). Perception of contingency as a determinant of social responsiveness. In Thoman E. B. (Ed.), *The Origins of Social Responsiveness* (pp. 33-64). New York: Erlbaum.

Wishart, J. G. and Bower, T. G. R. (1984). Spatial relations and object concept: a normative study. In Lipsitt, L.P. and Rovie-Collier, C. (Eds.), *Advances in Infant Research, Vol. 3* (pp.57-123). Norwood, N.J.: ABLEX Publishing Corp.

Wooldridge, M. and Rao, A. (Eds.) (1999). *Foundations of Rational Agency*. Dordrecht, The Netherlands: Kluwer Academic Publishers.