Journal of Experimental & Theoretical Artificial Intelligence

Vol. 00, No. 00, Month-Month 200x, 1–24

Symbolism and enactivism: An experimental test of conflicting approaches to artificial intelligence

S. F. WORGAN and R. I. DAMPER*

Information: Signals, Images, Systems (ISIS) Research Group, School of Electronics and Computer Science,

University of Southampton, Southampton SO17 1BJ, UK.

(*Received 00 Month 200x; In final form 00 Month 200x*)

One does not have to go very far into the subject of artificial intelligence (AI) before deep philosophical issues start to surface. The Turing test, the frame problem, the Chinese room argument, symbol grounding, the role of representation: these are just a few of the topics that have generated intense discussion and debate over the years. One particular controversy surrounds symbolism: whether or not it is necessary to have explicit symbol representation and, if so, the mechanism(s) by which these symbols came into existence. How are such questions to be answered? Generally, they are considered to lie in the realm of philosophy and not to be easily amenable to experimental test. However, when even limited experimental test is feasible, it can offer valuable insight. In this paper, we present an empirical test designed to explore some important foundational issues in AI. An artificial agent inhabits a digital world (a cellular automaton) in which its cognitive abilities vary in three dimensions (size of symbol memory, percentage of symbols that are innate, planning depth), allowing us to position it in a space that reflects degree of commitment to key philosophical standpoints. One plane of this space corresponds to pure symbol attachment, another plane corresponds to pure symbol grounding, and the origin of coordinates corresponds to pure enactivism. We find that an enactivist (purely reactive) agent architecture can perform as well as one employing planning in this scenario, if properly designed. Planning has strengths when task/environment complexity make design difficult but weaknesses if an inappropriate world model is acquired (e.g. as a result of mismatch between model and task/environment complexity). However, the main claim of the paper is that empirical exploration of the kind presented here could usefully form the initial phase of the design of many practical AI systems, and forms a valuable alternative to simply declaring *a priori* adherence to a particular philosophical position.

Keywords: symbol grounding, symbol attachment, enactivism, planning, cellular automata, philosophy of AI.

*Corresponding author, email rid@ecs.soton.ac.uk

Journal of Experimental & Theoretical Artificial Intelligence
ISSN 0952-813X print/ ISSN 1362-3079 online
http://www.tandf.co.uk/journals
DOI: 10.1080/09528130xxxxxxxxxxxx

1 Introduction

Throughout the 50–60 year history of artificial intelligence (AI), the majority view of its adherents has held that cognition can be replicated on a universal Turing machine (Turing, 1936) or equivalent; all that remains is to find the correct 'algorithm of thought'. And as we move towards this goal, so we will uncover a host of benefits, not only the ability to automate increasingly complex and useful tasks previously restricted to the exercise of human 'intelligence', but also gaining insights into the nature of the mind. Although there have already been some notable successes, profound disagreement remains among researchers over how to reach AI's ultimate goal. Indeed, some consider the entire endeavour to be ultimately futile (cf. Searle 1980).

For the continued health of AI as a discipline, further practical successes are mandatory. But much as the AI practitioner or technician might wish to sidestep or ignore philosophical concerns, this is not really an option. Philosophical issues are never far below the surface in AI (Boden, 1990; Copeland, 1993) so that every practical approach adopted implies commitment to *some* standpoint. Hence, practitioners would benefit from an effective way to untangle the philosophical debate, and reach a clearer view of the way forward. Currently, the debate centres essentially around thought experiments, but these have their obvious difficulties (Häggqvist, 1996; Gendler, 2000; Peijnenburg and Atkinson, 2003; Souder, 2003; Damper, 2006). We believe that thought experiment alone is inadequate to reach definitive conclusions and, difficult as it might be, empirical data bearing on the controversy need to be collected. At present, the number of attempts to test empirically philosophical issues in AI is vanishingly small. In this regard, one might mention the experiments of Kaernbach (2005) aimed at discovering what experimental subjects in a 'Chinese room' do and do not come to understand about their formally-specified task, but there are precious few other examples. The present paper aims to add to this tiny literature. We readily concede that this approach has its own problems; to keep the empirical work focused and tractable, simplifications and constraints need to be introduced, and these may have unavoidable bearing on the philosophical points at issue. It is, therefore, necessary to keep this in mind in interpreting the outcomes.

Specifically, we focus on one particular and important division of opinion in AI; namely the tension

between symbolism and enactivism. We then attempt an experimental test of the success of an artificial agent to maintain a self-sustaining density in a simple digital world as its cognitive abilities are manipulated to reflect commitment to the tenets of three different philosophical standpoints, namely:

(i) symbolism based on 'symbol grounding';

(ii) symbolism based on 'symbol attachment';

(iii) non-symbolic 'enactivism'.

In the following section, we outline the basic ideas of these three different philosophies. Thereafter, in Section 3 the general design methodology employed in the empirical comparison of these different philosophies is described. Section 4 details our specific experimental design and the results obtained. Section 5 concludes.

2 Symbolism and enactivism

At present, the AI community is divided such that we can identify a number of philosophical approaches. Many researchers begin from the assumption that cognition arises through the rule-based manipulation and combination of symbolic tokens (e.g. Newell, 1973; Minsky, 1974; Fodor, 1975; Newell and Simon, 1976; Newell, 1980, 1990; Pylyshyn, 1984; Dietrich, 1990). In one school of thought, the atomic symbols are formed through the perception of and interaction with the surrounding environment so as to give meaning to the symbols and their combinations. But this view faces a formidable problem, famously articulated by Harnad (1990) as: 'How can the semantic interpretation of a formal symbol system be made *intrinsic* to the system, rather than just parasitic on the meanings in our heads?' (p. 335). Harnad characterises this problem as *symbol grounding*. For AI systems to achieve meaningful 'intelligence', this question needs to be addressed. In light of its importance, various solution have been proposed—see Belpaeme *et al.* (2007) for some recent work.

Many agree with the symbolic perspective but contend that certain innate concepts need to be present at birth. For example, internal representations of space and time are required to make sense of the surrounding

environment. These innate concepts are 'attached' to the world through experience; they are not formed. Accordingly, Sloman and Chappell (2004) identify this school of thought as *symbol attachment*. Despite the necessity for innate concepts in virtually all (or perhaps literally all) practical AI systems, in that the system designer has to implant some specification or driver of desired behaviour, symbol attachment faces the charge of ignoring the challenge of 'parasitic meaning'. Biological systems derive innate information from the genome, which the long process of evolution has shaped (or grounded) through the interaction between organism and environment, but this solution is not available to designed, artificial systems.

Symbolic AI has historically faced a host of problems when situated in real world environments. This has led some to parody the symbolic approach as 'good old-fashioned AI' or GOFAI (Haugeland, 1985). As a result, certain researchers—notably Brooks (1990, 1991a, 1999)—seek to discard the symbolic notions of the mind entirely. This is typically termed 'embodied AI'. Workers in this paradigm build systems that act upon the world through cycles of perception and action based on some adaptive connectionist or statistical machine learning principles. A question remains as to whether these connectionist models truly discard symbolic representations or simply obscure them. Aside from this, embodied AI still faces a number of challenges when modelling higher cognitive functions such as language and abstract reasoning.

Having very briefly reviewed the major schism of symbolic versus non-symbolic approaches to AI, we now move to a slightly more detailed treatment of the three philosophical standpoints identified earlier.

2.1 *Symbol grounding*

Symbol grounding holds that meaning can be found in a symbolic system, so long as the symbols are derived from the classification of the system's experiences. These experiences come from the environment, through sensor readings, and so have meaning intrinsic to the system, rather than just being 'parasitic'. It is then possible to take these symbols, grounded in the environment, and manipulate them, thereby deriving new meanings (Cangelosi *et al.*, 2000). Meaning then flows from initial experiences up to higher-order symbolic manipulation.

Symbol manipulation is usually carried out through the planning paradigm. The main features of

planning can be characterised by a distinction between the perception, the action, and the reasoning functions of the agent. Significantly, the agent's reasoning involves the manipulation of an explicit world model. Symbol grounding attempts to answer the question: How do we give this world model meaning?

2.2 *Symbol attachment*

The symbol grounding approach is underpinned by the assumption that meaning could only be derived from experience. To some this heralded the rebirth of concept empiricism as put forward by Locke (1690/1979), who held that at birth the mind was a 'tabula rasa'. According to Aaron Sloman (personal communication), concept empiricism 'was refuted long ago by the philosopher Immanuel Kant (1781)', who argued that one cannot have experiences unless one already has concepts. If the content of mind does not have any initial meaning, further meanings cannot be derived. Therefore, our experiences cannot be the sole source of knowledge. Some basic concepts cannot come directly from the environment and so symbol grounding, in its strictest sense, must be flawed.

Symbol attachment brings Kant's refutation of concept empiricism into the field of AI. Like symbol grounding, it considers the structure of concepts to be essential to the nature of meaning, but these meanings do not have to come from the experience of the environment. Rather, sensory experience can be used to reduce residual indeterminacy. However, the meanings themselves are not only derived from the environment but also from an initial innate source. As the agent experiences the world, this innate knowledge becomes grounded. An instinctive, abstract concept becomes relevant through interaction with the environment.

Some researchers (MacDorman, 1997; Nenov, 1991; Klinspor *et al.*, 1996), in attempting to implement a symbol-grounding-based system, have unintentionally illustrated this problem with strict symbol grounding. When taking the planning approach, it is always the case that certain concepts (usually space, time and cause and effect) need to be 'designed in' by the programmer. Symbol attachment wholeheartedly accepts the need for this designed element but still leaves us with a number of problems.

If we are to accept the existence of genetically-encoded concepts, the only possible source of innate

knowledge, how is this to be represented in an AI system and where does it leave our concept of meaning? We could simply return to the meaningless design of various concepts for our agent, citing an abstract genetic 'learning' as our justification. However, we need to consider that this genetic encoding was a product of the environment. So rather than obtaining meaning by direct sensory experience, grounding is provided through evolutionary selection pressures. The animal as a whole, genome and all, is grounded in the environment. By comparison, simply designing a set of concepts into an artificial system leaves us open to Harnard's old objection; specifically the 'meaning' in the system is simply parasitic, being more relevant to the designer then the system. This is the artificial equivalent of psychology's *bete noir*, the homunculus, described by Lloyd (1989, p. 205) as 'the primary method of passing the buck with respect to genuine explanation'.

2.3 *Enactivism*

Others seek to avoid the question of meaning by eliminating any symbolic planning within the system, e.g. Brooks (1990, 1991a, 1999). In particular, Brooks (1991b) favoured a subsumption architecture in which different layers of control behave reactively to the environment. These reactive layers of control are arranged on top of each other, with the actions of the higher layers taking priority over the lower layers, whereas the information from the lower layers feeds upwards to the higher layers. By comparison, a planning system uses a complex world model, in combination with a detailed learning algorithm, to produce intelligent behaviour. Planning typically produces complex systems, since any failure at the task is generally taken to reflect shortcomings of the model, which is then made more complex to cope.

2.4 *Space of Cognitive Abilities*

A key aspect of this work is that we devise a space—actually three-dimensional—in which we can position the agent according to its 'adjustable' cognitive abilities, and study its success at some relevant task. This space is depicted in Figure 1. The three dimensions are:

(i) The size of the agent's symbol memory. (In this work, this ranges from 0 to 100 symbols.)

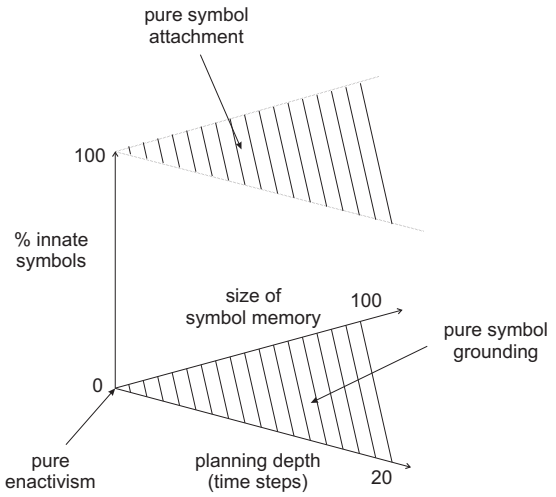

Figure 1. In this work, the agent's cognitive abilities are defined by its placement in a 3-D space, having a direct mapping to the philosophical standpoints of symbol grounding, symbol attachment and enactivism.

(ii) The percentage of symbols in this memory that are innate (i.e. not acquired through experience).

(iii) The extent to which the agent attempts to predict the future behaviour of the environment, quantified as planning depth (in time steps).

The space is devised to portray points and regions that reflect degrees of commitment to the three opposing philosophical standpoints of symbol grounding, symbol attachment, and enactivism. As depicted in Fig. 1, the pure symbol grounding approach to AI maps to the plane at which the percentage of innate symbols is zero, the pure symbol attachment approach maps to the plane at which the percentage of innate symbols is 100%, and pure enactivism maps to the origin of coordinates, at which there are no symbolic abilities at all. The reader is warned that the design of this space is such that its three dimensions do *not* correspond exactly to the three philosophical standpoints of concern, and should not be confused.

Sloman and Chappell (2004) detail the biological plausibility of the grounding-to-attachment continuum (the z-axis of Fig. 1), by identifying a spectrum of animals ranging from the altricial to the precocial. A precocial bird will be able to function within a few days of birth. It has been born with all of the innate knowledge it needs to survive in its environment. By comparison, an altricial bird is dependent on its parents for survival; it needs time to develop and acquire its concepts of the world through experience.

3 Design methodology

The agent—with cognitive abilities defined by its coordinates in the space of Fig. 1—is placed in an environment in which it interacts with a two-dimensional cellular automaton. In this system, α the alive cell threshold, and, β the dead cell threshold, can be adjusted within the range 0–8 allowing for a variety of cellular automata to be used. For example, by setting $\alpha = 3$ and $\beta = 2$, Conway's 'game of life' is implemented. The agent is placed at a random location within the environment; its overall task is to maintain a proportion of cells, specified by the designer, in the entire environment, in the alive state. There are two sub-tasks: either to increase the alive-cell density from an initial to a higher value, or to decrease the alive-cell density from an initial to a lower value.

3.1 *Agent abilities*

To perform its task of maintaining a target proportion of alive cells, the agent has a number of abilities: perceptual, action and cognitive. The perceptual and action abilities are *fixed*. The cognitive abilities are *variable*, and correspond to parameter settings that position the agent at some appropriate point in the 3-D space of Figure 1, so enabling us to compare experimentally the various philosophical approaches to AI embodied in this figure.

The agent moves through and alters the environment by its actions. It can move one cell in any direction (i.e. 8-neighbourhood). When a cell is within a radius of two cells from the agent (i.e. 24-neighbourhood), it is able to 'flip' the state of the cell. Potentially, it can flip all 24 cells in a single time step, but only every four time steps—this being the number of steps required to arrive in a totally new local environment.

There are basically three perceptual abilities: the agent can look ahead in a 180° arc, split into 20 equal sections, and can sense the proportion of alive cells within each section. The radius of this arc is set at 20 cells. It can sense the proportion of alive cells in the environment as a whole, enabling it to compare this number to the target proportion. The agent is also able to perceive the density of the surrounding 24 cells from its current location, and at the last location at which it performed an action, enabling it to assess the consequences of its actions.

3.2 *Pure symbol grounding agent*

A symbolic agent is one that manipulates a (symbolic) model of its environment, enabling it to explore hypothetical scenarios and theorise about how the world will behave in the future. It needs to experiment with possible approaches and learn the consequences of its actions, making it capable of planning its actions instead of simply reacting. In the case of pure symbol grounding, the agent needs to construct its world model through experience of the environment; for the symbols to have meaning, they need to be grounded in the world.

Taking the abilities detailed in Section 3.1 above, we add planning and learning. At each time step, the agent records the current perceived state of the world using its vision arc, and measures the density change from the previous time step. Thus, it acquires a set of symbols, one for each section of its vision arc, recording how densities in the environment have changed over time. These symbols are arranged in the agent's memory (world model) to provide it with a record of how neighbouring densities alter the density under consideration.

This learning process is defined by Algorithm *1*. Once the world model has been constructed over a minimum of two time steps, it can be used to explore possibilities about the world. Before deciding upon a course of action, the agent will predict the densities of the area within its vision arc for a number of time steps into the future (i.e. the planning depth).

The planning process is defined by Algorithm *2*, by which the agent predicts how the density of the surrounding environment will change over time. To enable the symbolic information to be generalised to novel situations, the recorded density changes are rounded to three decimal places (i.e. numerical precision of $p = 3$). Accordingly, similar local environments will be treated in a similar way. The agent acts on this information to alter its environment.

Having made a prediction, the agent selects a heading and an action as defined by Algorithms *3* and *4*. It does this by considering how previous heading decisions and actions have affected the world. The selected heading/action will be those that cause the greatest predicted density change towards the desired goal. If it has not encountered the surrounding 24 cell density before, it will select a random action and observe

Algorithm *1* Learning process for the symbol grounding agent.

loop
  S[ ] ← density values
  P[] ← placed cell pattern
  for all i such that $0 \leq i \leq$ S[ ].length **do**
    CS[ ] ← S[i](t) + S[i]($t-1$)
  end for
  for all j such that $0 \leq j \leq$ CS[ ].length **do**
    densityL ← CS[$j-1$]
    densityR ← CS[$j+1$]
    if densityL ≥ CS[j] and densityR ≥ CS[j] **then**
      I[ ] ← CS[j] + P[j]
    end if
    if densityL ≤ CS[j] and densityR ≤ CS[j] **then**
      D[ ] ← CS[j] + P[j]
    end if
    if densityL ≥ CS[j] and densityR ≤ CS[j] **then**
      IL[ ] ← CS[j] + P[j]
    end if
    if densityL ≤ CS[j] and densityR ≥ CS[j] **then**
      IR[ ] ← CS[j] + P[j]
    end if
  end for
  EnvInc ← $\overline{\text{I}}$
  EnvDec ← $\overline{\text{D}}$
  LefInc ← $\overline{\text{IL}}$
  RigInc ← $\overline{\text{IR}}$
end loop

Algorithm *2* Planning process for the symbol grounding agent.

PW ← perceived world
for j such that $0 \leq j \leq$ prediction depth **do**
  for i such that $0 \leq i \leq$ P.density **do**
    GP ←generalise(P[i])
    if GP[i] ≡ one of {I[ ], D[ ], IL[ ], IR[ ]} **then**
      PW($t+1$) ← PW+ matched {I[ ], D[ ], IL[ ], IR[ ]}
      PW($t+1$) ← PW($t+1$)+ matched {EnvInc, EnvDec, LefInc, RigInc}
      PW ← PW($t+1$)
      $t \leftarrow t+1$
    end if
  end for
end for

how this affects the local density after four time steps. Similarly, if the agent has not encountered the observed density arc before (i.e. 180° arc of 20-cell radius), it will pick a random heading and observe how this, combined with the selected action, has affected the global density. As a result, it can learn through

Algorithm 3 Heading selection process for the symbol grounding agent.

```
if PW = one of I[], D[], IL[], IR[] then
  H ← H from I[], D[], IL[], IR[]
  if H > 1 then
    H ← H from, sort(I[], D[], IL[], IR[], greatest desired density change)
  end if
else
  H ← random
  Execute H
end if
```

Algorithm 4 Action selection process for the symbol grounding agent.

```
LW ← surrounding 24 cells
if LW = one of I[], D[], IL[], IR[] then
  P ← P from I[], D[], IL[], IR[]
  if P > 1 then
    P ← P from, sort(I[], D[], IL[], IR[], greatest desired density change)
  end if
else
  P ← random
  Execute P
end if
```

its experience and construction of a generalised world model.

The world model is defined by a number of density-change symbols, sometimes linked with action patterns or headings, which the agent can store in its memory. This number is set by the researcher and can range from 0 to 100. When the symbol limit is reached, the oldest symbol is deleted to allow each new symbol to be added.

3.3 *Pure symbol attachment agent*

An agent making use of symbol attachment can use innate concepts given to it at birth. To do this, an agent will either have a hand-coded world model or it will be trained in one environment and then re-born in another. The concepts acquired through experience in one life will then become innate in the next. As the agent experiences the new environment, these innate concepts become attached to the environment as they are experienced anew, or they remain ungrounded if they are never experienced. The pure symbol attachment agent and the pure symbol grounding agent are distinguished only by the former's possession

of an established set of concepts at birth; all other abilities are equal.

3.4 *Pure enactivist agent*

The enactivist agent is implemented by linking its perception reactively to its actions. Its perception of the proportion of alive cells is linked to its choice to flip neighbouring cells. If it perceives that there are more alive cells than required, it will flip the neighbouring cells to the dead state. Conversely, if there are not enough alive cells, the agent will flip the neighbouring cells to the alive state.

The vision arc is then used to determine the agent's movement. If the agent has to reduce the number of alive cells, it will move towards the highest density of cells. Alternatively, if it has to increase the number of alive cells, it will move towards the lowest density of cells. If it cannot make a heading decision, it will pick a random direction, as no more sensible course of action is available.

3.5 *Parameter adjustment*

A few relevant parameters can be used to define a position in the space spanning these three approaches (Fig. 1). We have already specified how re-birth can create an attachment agent, but there is a continuum between attachment and grounding. The proportion of innate symbols determines the agent's location on this continuum—the z-axis of Fig 1. If there is no room in the memory for additional symbols to be formed through experience, then the agent is entirely precocial; it makes plans based on its innate concepts. However, if there is available memory, then it is capable of learning more about the world through experience.

Additionally, an increase in the planning depth allows us to move away from a pure enactivist agent. The more an agent can plan ahead, the less it is simply reacting to the current environment. The 'environment as its own model' (Brooks, 1991b) is increasingly replaced by a plan defined by the agent's internal model.

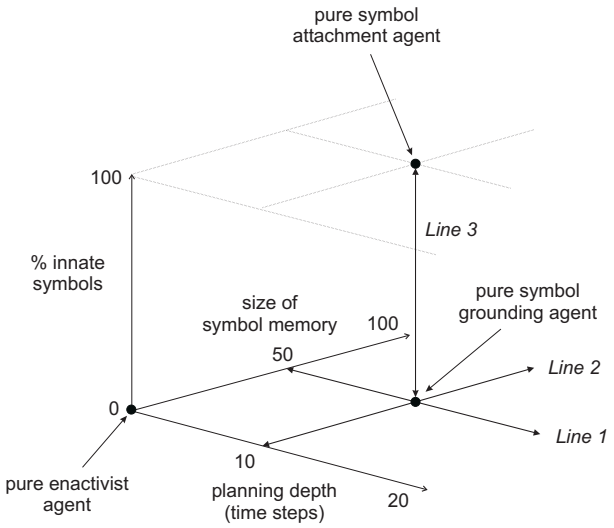

Figure 2. Positioning of pure symbol grounding agent, pure symbol attachment agent and pure attachment agent in the 3D-space of cognitive architectures. The thick solid lines denote the parameter variations employed to explore the space.

4 Experimental design and results

Clearly, the space of Fig. 1 is very large and it is not feasible to explore it fully. All we can hope is to look at a reasonable number of points (corresponding to specific agent cognitive architectures/abilities) within it. The experimental design aims at giving reasonable coverage, allowing us to draw valid and useful conclusions.

Since Fig. 1 has three axes, we have evaluated the impact of each: planning depth (x-axis), size of symbol memory (y-axis) and percentage of innate symbols in this memory (z-axis). When assessing any one of these, the other dimensions are held constant. The default parameters for each philosophical standpoint are depicted in Figure 2, along with thick lines to denote the parameter variations around (two of) the defaults.

For the symbol attachment agent, it is necessary to have a mechanism for supplying innate symbols. This was done by placing a symbol grounding agent in an environment and running a simulation until its symbolic memory was full. It was then re-born in a new environment where it was assessed. The initial environment had the same settings as the environment in which it was reborn.

Three cellular automata were used as environments, as depicted in Table 1. In all three environments, the agent had to adjust the alive-cell density according to the initial parameters shown in Table 2, for the two sub-tasks mentioned previously. Each cellular automaton was run for 500 time steps, which was found

Table 1. Specification of the three cellular automata used in this work, denoted Complex, Stable and Exploding.

α	β	Name	Type
{2,3}	{2}	Conway's Life	Complex
{5}	{3,4,6}	Long Life	Stable
{3,4}	{3,4}	34 Life	Exploding

Table 2. The agent is set one of two sub-tasks, either to increase or to decrease the density of alive cells.

Sub-task	Initial density (%)	Desired density (%)
Increase	20	35
Decrease	20	5

to be sufficient for the agent to have a noticeable effect on its environment. The size of the environment was 35×35 cells. This is a reasonable compromise, being large enough to present the agent with a challenge yet still allowing it to have an effective influence on it.

4.1 *Comparison of the pure agent architectures*

These are compared with respect to deviation between actual and desired density of alive cells as a function of time. Fig. 3 shows the performance of the enactivist and symbol grounding architectures across the 500 time steps. Each point depicted is for an average over 5 repetitions from a different random starting point of the 2 sub-tasks and 3 environments.

As expected in view of its fixed, reactive architecture, the performance of the enactivist approach remains unchanged apart from (quite large) statistical variation caused by the uncertain (as far as the agent is concerned) changes in environment. Initially, the symbol grounding approach is more or less indistinguishable from the enactivist approach, since few or no symbols have been acquired. Over time, the symbol grounding approach undergoes a learning phase in which it is acquiring symbols and its performance is relatively poor. At time step 50, its symbolic memory is full and old symbols may be discarded for new ones. On the basis of a t-test of means, symbol grounding is poorer than enactivism over the period of 50 to 225 time steps (marginally significant $P = 0.07$). Thereafter, as the symbol memory starts to stabilise, performance improves and becomes competitive with the enactivist architecture. (The pattern of performance was largely unchanged for time steps great than 500.)

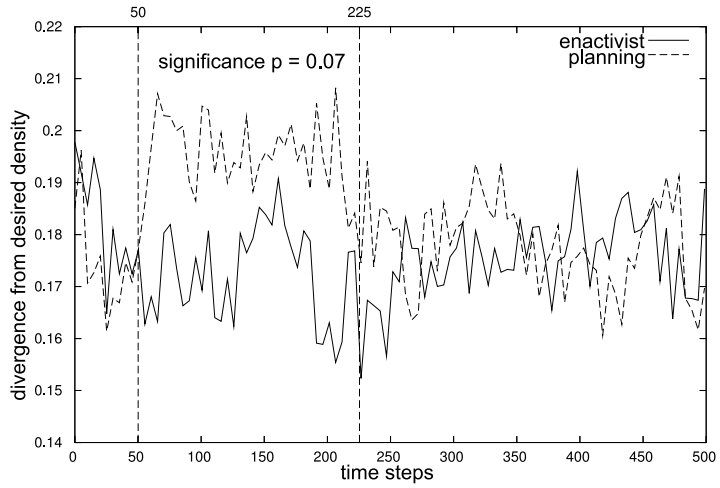

Figure 3. Performance of enactivist architecture and symbol grounding architecture over time.

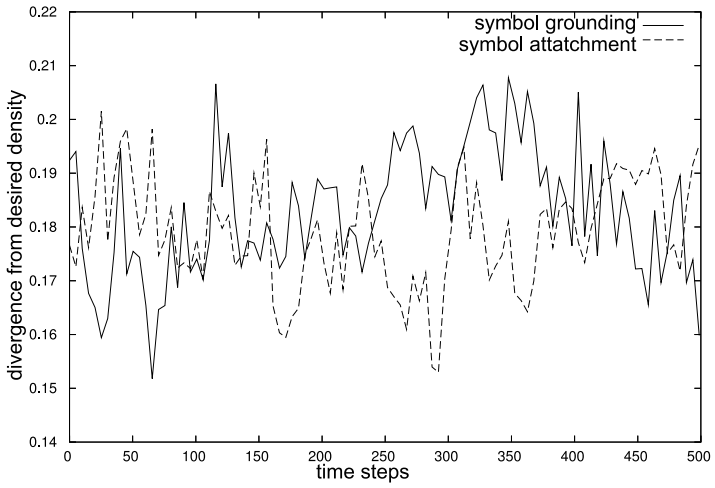

Figure 4. Performance of symbol attachment architecture and symbol grounding architecture versus time.

Fig. 4 shows the performance of the symbol attachment and symbol grounding architectures across the 500 time steps. Again, each point depicted is for an average over 5 repetitions of the 2 sub-tasks and 3 environments. At face value, and given the effect of the uncertain (as far as the agent is concerned) environmental changes, it is hard to say that there is much systematic difference between these two. We do, of course, expect greater variability in the performance of the symbol grounding architecture simply because it is a learning system; its world model is subject to constant change. Although symbol attachment might speed the learning process, the innate concepts need to be relevant to the actions and performance of the agent.

In summary, there is little to choose between the two pure symbolic architectures (Fig. 4). The attachment architecture performs generally better in the short term (Fig. 3), since it does not have to learn, and learning can be an error-prone process. However, after successful learning, a symbolic architecture can do just as well. A well-designed purely reactive system uses the world as its own model and can behave appropriately as a result. It may have a prediction depth of zero, but by the same token, there will be zero error in its world model (since it doesn't have one). In this work, the environment stops while the agent builds its world model; this does not happen in the real world! Previous work, e.g. Shakey the robot (Nilsson, 1984), has shown that the time taken to construct a world model is a crucial problem when a real-world robot is constructed. While the agent is building its representation, the world itself carries on, making any plan built for a static environment potentially irrelevant. This gives an edge to the enactivist agent when compared to the symbol grounding agent.

4.2 *Variation of planning depth*

Here, we test the agent's performance as a function of planning depth by varying its position along *Line 1* in Fig. 2 from depth 0 to 20, in increments of 5. Note that when planning depth equals 0, we have a pure enactivist agent—symbol memory size is irrelevant if there is no symbolic planning.

Figure 5 shows the deviation from the desired density. Each point depicted is for a particular time step (of the 500) averaged over 5 repetitions of the 2 sub-tasks and 3 environments. It seems that there is a disadvantage to having planning capability, or at least, the agent is unable to exploit it to build an accurate world model.

Why is the model inaccurate? Is it a result of poor generalisation? To answer these question, the difference between the agent's world model and the actual world was computed, averaged over both sub-tasks, the 3 environments, 500 time steps and 5 repetitions, as a function of planning depth and level of generalisation (i.e. the degree of numerical rounding, or precision, p). The result is shown in Figure 6. At this point, we should emphasise the (subtle) difference between the y-axes of Figs. 5 and 6. The y-axis of Fig. 5 (divergence from desired density) indicates how far the agent is from achieving its goal, i.e. it is a measure

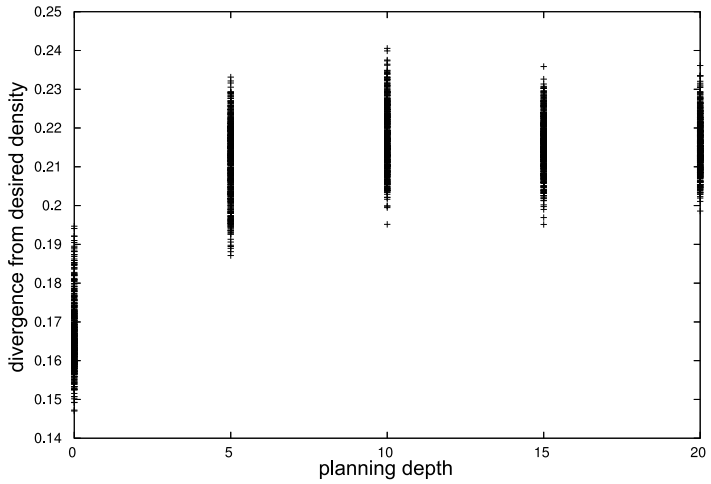

Figure 5. Average performance difference between actual and desired alive cell density as a function of planning depth, for the default value of precision parameter, $p = 3$.

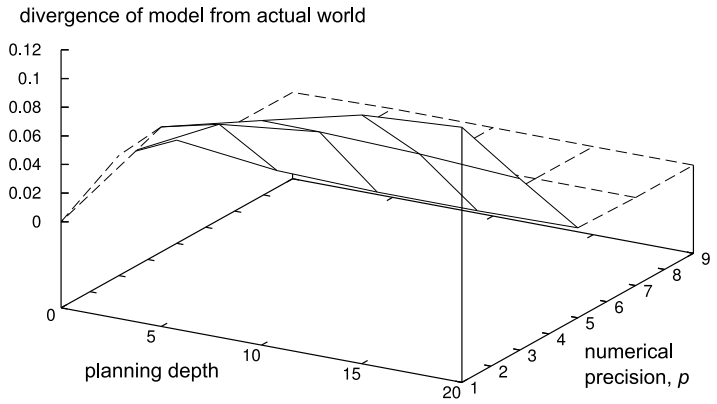

Figure 6. World model accuracy for different levels of planning depth and numerical precision, p. Note that p is an inverse measure of the agent's level of generalisation.

of 'performance'. However, the y-axis of Fig. 6 (divergence of agent's model from actual world) indicates how far the agent's internal world model is from the actual world.

The curve in Fig. 6 for the higher levels of generalisation ($p \sim 3$) is very similar in shape to that of Fig. 5, indicating that the deterioration in performance as planning depth increases is most likely due to a failure of the world model when generalisation is too severe. This error affects the agent's predictions causing it to make the wrong heading decisions and action choices; it can be countered by relaxing the degree of generalisation, i.e. increasing the numerical precision by increasing p. The divergence of the agent's model from the actual world then falls to zero, irrespective of planning depth, for $p \geq 6$. The question arising is

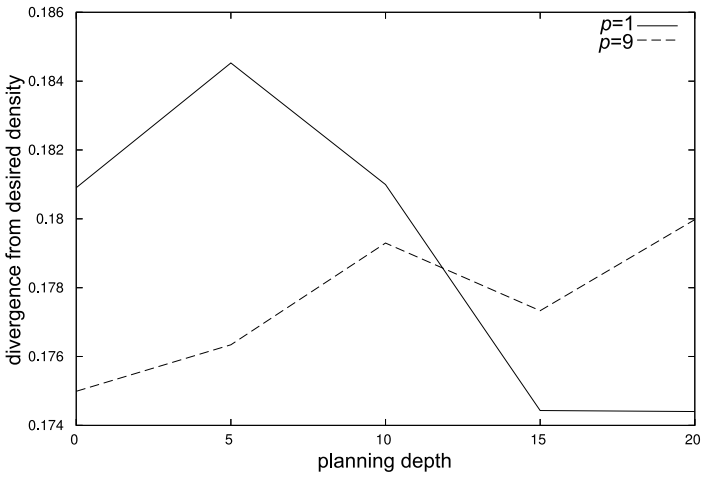

Figure 7. Difference in performance over different planning depths when the numerical precision is 1 and 9.

how this reduced generalisation capability would be reflected in agent performance.

As can been seen in Figure 7, when the precision parameter is set to $p = 1$, the performance improves as the planning depth increases. This suggests that a reasonably high level of generalisation makes the world model useful, in spite of its divergence from the actual world, as one would expect intuitively (by avoidance of over-fitting to specific cases). By comparison, when the symbols have high levels of precision, $p = 9$, the performance appears to decrease with planning depth, probably because the precise world model scenarios are used only very few times (i.e. there is over-fitting). As a result the agent is forced to choose a new random action and heading at each time step.

4.3 *Variation of symbol memory capacity*

By varying the parameters along *Line 2* of Fig. 2, we investigated the role of the size of the agent's symbolic memory. As seen in Figure 8, the memory size has no discernible effect on performance.

4.4 *Variation of proportion of innate concepts*

We assessed a range of configurations, from grounding to attachment, by varying the proportion of innate concepts along *Line 3* of Fig. 2. The cognitive architecture of the agent was initially set to pure symbol grounding with its memory capacity set to 100 symbols. Further configurations were tested by increasing

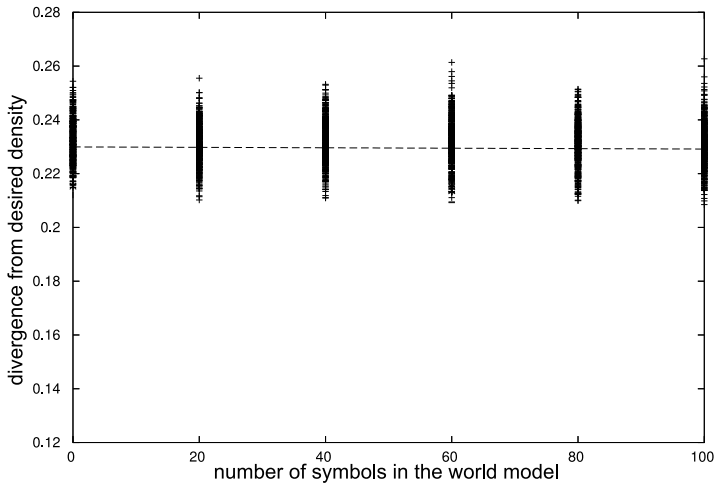

Figure 8. World model accuracy for different levels of planning depth and generalisation.

the proportion of innate concepts in the memory from 0 to 100 in increments of 20 symbols. The average deviation from the desired density was recorded for all combinations of the 2 sub-tasks and 3 environments, and for 5 repetitions. In addition, the deviation in prediction accuracy (between the world model and the cellular automaton) was recorded and averaged.

Figure 9 shows the deviation from desired cell density (i.e. performance) for these different configurations as a function of time. There is a slight decrease in performance as proportion of attached symbols increases, and performance is more or less constant over time. Figure 10 shows divergence between the agent's world model and the actual world for the same situation as depicted in Fig. 9. The results indicate a possible problem with the inflexibility of innate concepts, in that accuracy decreases with increased proportion of attached symbols and also over time, although there is no impact on the performance (Fig. 9). This occurs even when the symbol attachment agent has been trained in a similar environment to the test environment, and may be a result of changes in the initial configuration of the two environments: concepts acquired in one do not necessarily apply in the other. This further illustrates the dangers of using inflexible innate concepts even if they appear to be correct.

It is noteworthy that here the deterioration in model accuracy was not reflected in depressed performance whereas in Section 4.2 it was. The simplest interpretation of this finding is that it is the interplay of planning depth and proportion of attached symbols that is the cause of deterioration.

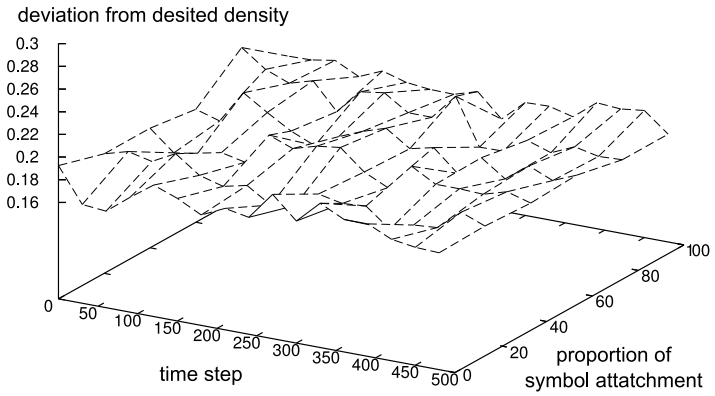

Figure 9. The world model deviation from the environment on the altricial-precocial scale.

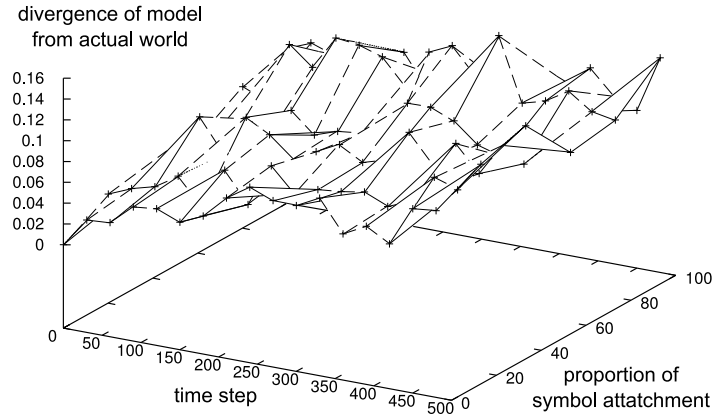

Figure 10. The world model deviation from the environment on the altricial-precocial scale.

5 Conclusion

In this paper, we have attempted to show how empirical or experimental research might offer at least some insight into foundational, philosophical questions underpinning artificial intelligence (AI). Of course, the findings here are highly specific to the situation studied, and it would be foolhardy to suggest that they had general validity for all fields of AI. Rather, our main claim is that an empirical exploration of the space of possibilities, in the way that we have done, could usefully form the initial phase of the design of many practical AI systems. It forms a valuable alternative to the AI technician simply declaring *a priori* adherence to a particular philosophical position, as tends to happen at present. This *a priori* allegiance to a school of thought anyway carries the danger of ending up being compromised by the practical necessity of making a system work.

With this important caveat, and given that we must always be careful when drawing general conclusions from designed models, what insights have we gleaned and how portable are they between applications? Clearly, both our agent and its artificial environment were highly simplified, to keep the work focused and tractable; hence, we readily concede that portability is low. However, as we scale up complexity, we believe that the sort of issues identified here will only become more acute.

Specifically, we have found the following. In some circumstances, a purely reactive (enactivist) system may not only be sufficient, but can have concrete advantages. These stem from the avoidance of having to acquire a world model that has the right level of generalisation, tracks changes in the environment on an appropriate time scale, etc. Well-known problems such as the bias-variance, stability-plasticity and exploration-exploitation dilemmas raise their head here. Of course, the reactive system has to be properly designed in the first place, and herein lies a real problem when the application confronts a significant degree of complexity. Although symbolic planning offers a solution in this case, the complexity of the world model and the algorithm for acquiring it need to be matched to the complexity of the environment and of the task, as explored here through the symbolic memory size.

Although obviously simplified, we argue that the environment devised for this work has many positive aspects; it is dynamic and has a useful degree of complexity, providing a valid challenge for the agent. But this dynamic nature places a limit on the significance of the results as the density changes effected by the agent are often dwarfed by the density changes of the environment. By averaging the results over 3 cellular automata, 2 sub-tasks and multiple runs, we limit the effect of the environments' own interactions and expose the agent's achievements.

With the empirical framework outlined here, researchers can begin to define, test, and justify their own approaches. It might even lead to a reappraisal of apparently competing philosophies, and perhaps to their extension and/or reconciliation. This is important because otherwise we are left with verbal theories based loosely on the mind-as-machine metaphor. These verbal theories can suffer from a lack of precision and hidden assumptions. Ultimately, let the success of AI be judged by its results and not its philosophy.

Acknowledgement

The authors are indebted to Aaron Sloman who originated some of the ideas embodied in this paper, and contributed to the development of several others.

References

Belpaeme, T., Cowley, S., and MacDorman, K. F., editors (2007). Symbol Grounding: Special issue of *Interaction Studies*, **8**(1).

Boden, M. A., editor (1990). *The Philosophy of Artificial Inteliigence*. Oxford University Press, Oxford, UK.

Brooks, R. A. (1990). Elephants don't play chess. *Robotics and Autonomous Systems*, **6**(1), 3–15.

Brooks, R. A. (1991a). Intelligence without representation. *Artificial Intelligence*, **47**(1–3), 139–159.

Brooks, R. A. (1991b). The role of learning in autonomous robots. In *Proceedings of the Fourth Annual Workshop on Computational Learning Theory*, pages 5–10, San Mateo, CA. Morgan Kaufmann.

Brooks, R. A. (1999). *Cambrian Intelligence*. Bradford Books/MIT Press, Cambridge, MA.

Cangelosi, A., Greco, A., and Harnad, S. (2000). From robotic toil to symbolic theft: Grounding transfer from entry-level to higher-level categories. *Connection Science*, **12**(2), 143–162.

Copeland, B. J. (1993). *Artificial Intelligence: A Philosophical Introduction*. Blackwell, Oxford, UK.

Damper, R. I. (2006). Thought experiments can be harmful. *The Pantaneto Forum,*, **Issue 26**. `http://www.pantaneto.co.uk/`.

Dietrich, E. (1990). Computationalism. *Social Epistemology*, **4**(2), 135–154.

Fodor, J. (1975). *The Language of Thought*. Crowell, New York, NY.

Gendler, T. S. (2000). *Thought Experiment: On the Powers and Limits of Imaginary Cases*. Garland Press, New York, NY.

Häggqvist, S. (1996). *Thought Experiments in Philosophy*. Almqvist & Wiksell, Stockholm, Sweden.

Harnad, S. (1990). The symbol grounding problem. *Physica D*, **42**, 335–346.

Haugeland, J. (1985). *Artificial Intelligence: The Very Idea*. Bradford Books/MIT Press, Cambridge, MA.

Kaernbach, C. (2005). No virtual mind in the Chinese room. *Journal of Consciousness Studies*, **12**(11), 31–42.

Klinspor, V., Morik, K. J., and Rieger, A. D. (1996). Learning concepts from sensor data of a mobile robot. *Machine Learning*, **23**(2–3), 305–332.

Lloyd, D. (1989). *Simple Minds*. Bradford Books/MIT Press, Cambridge, MA.

Locke, J. (1690/1979). An essay concerning human understanding. P. H. Nidditch, editor, Oxford University Press, Oxford, UK. Original work published 1690.

MacDorman, K. F. (1997). *Symbol Grounding: Learning Categorical and Sensorimotor Predictions for Coordination in Autonomous Robots*. PhD thesis, University of Cambridge, UK.

Minsky, M. (1974). A framework for representing knowledge. Technical note AIM–306, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA.

Nenov, V. I. (1991). *Perceptually Grounded Language Acquistion: A Neural/Procedural Hybrid Model*. PhD thesis, University of California, Los Angeles.

Newell, A. (1973). Artificial intelligence and the concept of mind. In R. C. Shank and K. M. Colby, editors, *Computer Models of Thought and Language*, pages 1–60. Freeman, San Francisco, CA.

Newell, A. (1980). Physical symbol systems. *Cognitive Science*, **4**(2), 135–183.

Newell, A. (1990). *Unified Theories of Cognition*. Harvard University Press, Cambridge, MA.

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

Nilsson, N. J. (1984). Shakey the robot. Technical note 323, SRI International, Menlo Park, CA.

Peijnenburg, J. and Atkinson, D. (2003). When are thought experiments poor ones? *Journal for General Philosophy of Science*, **34**(2), 305–322.

Pylyshyn, Z. W. (1984). *Computation and Cognition: Toward a Foundation for Cognitive Science*. Bradford Books/MIT Press, Cambridge, MA.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, **3**(3), 417–457. (Including peer commentary).

Sloman, A. and Chappell, J. (2004). The altricial-precocial spectrum for robots. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*, pages 1187–1192, Edinburgh, UK.

Souder, L. (2003). What are we to think about thought experiments? *Argumentation*, **17**(2), 203–217.

Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society (Series 2)*, **42**, 230–265.