Reactive Embodied Agents: The road to super-intelligent machines?

Amitabha Mukerjee
Computer Science and Engineering
I.I.T. Kanpur
e-mail: amit@iitk.ernet.in

Abstract

The term "agent" refers to machines or systems that can operate in limited environments with differing degrees of autonomy. An agent is "embodied" if it has a physical body as in a robot or micro-actuator. It is "reactive" if its action is decided based on the current situation, and not on the memory of past events. Reactive models require relatively smaller memories and are compact compared to global models of intelligence and have provided a useful design paradigm for practical embodied agents. However such models suffer from some shortcomings in that its "memoryless" operation may make it prone to cycles where the same action is repeated over and over. Some of these problems may be avoided by clearly defining the objectives of the agent and providing it with mechanisms for autonomously learning these behaviours in various situations. We present some results for such learning in a soccer-robot scenario, and discuss some of the long-term ramifications of such technological leaps for developing countries such as India.

1 Introduction

This paper describes an area of work involving the basic building blocks for intelligent systems called "agents". We present theoretical and implementational results for the "reactive" model for intelligent systems, which is similar to many control systems used in biology in that the resultant action depends only on the current stimulus and current objectives, and not on past actions [1, 6, 3]. This approach contrasts with "deliberative" models of intelligence which use knowledge-based models of the environment to plan and reason about the world. Biologically, many tasks requiring quick response (as in a batsman hitting the ball or tongue motions in speech) are implemented reactively, and in artificial intelligence, this model has been of significance in building real systems, as in soccer-playing robots.

In particular, we present a theoretical study of a phenomenon that is exhibited similarly in memoryless systems in machines, animals, and humans – patterns of repeated behavior or cycles. From the Artificial Intelligence point of view, we consider a simplified memoryless, reflexive mode of behavior where the current action is completely determined by the present stimulus - i.e. the set of descriptors for the present environment. We conduct a formal analysis of such systems and indicate some weaknesses of purely memoryless models.

An agent acting over a period of time may execute a number of actions, called a "sequence". Particularly significant is the case where the consequences of one or more actions contain the stimulus for some previous action , resulting in cycles that can be potentially endless. The key result [3] is the finding that in memoryless systems, the likelihood of developing such cycles increases as the tasks attempted increase in power. We also investigate some behavior modification mechanisms that are capable of reliably removing such cycles, but show how these may result in degraded performance. Finally, we discuss mechanisms for combining reactive and deliberative models as a mechanism for reducing the incidence of such cycles.

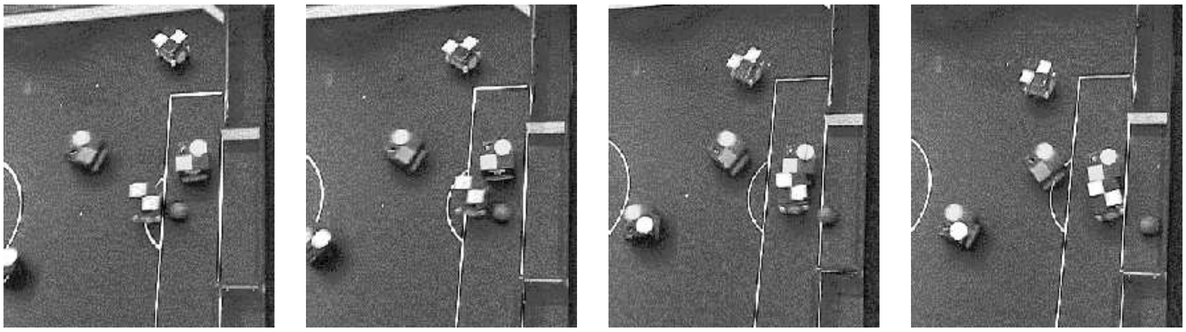


Figure 1: *Micro-Robot Agents play Soccer.* A goal is scored by two opponents against three defenders. Two colour blobs on each robot identifies the robot and its team, as well as its position (via roof-mounted camera.) Learned reactive behaviours would work well for such agents.

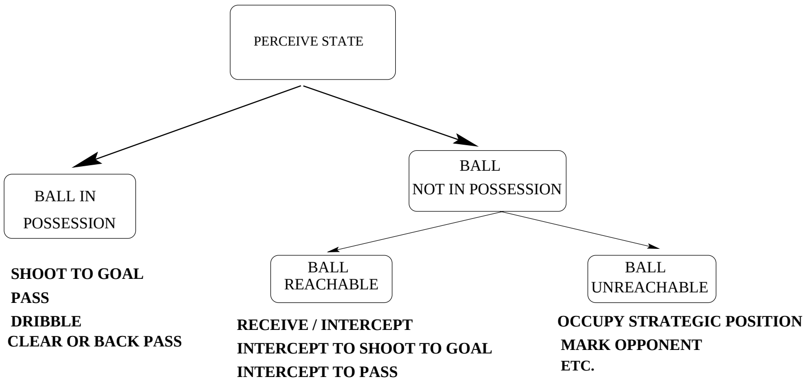


Figure 2: *A decision tree for meta-level reasoning in robot soccer.* Different behaviours are executed under differing circumstances - e.g. defence, or penalty-kick. For efficient learning, each behaviour is learned separately.

In many situations such as robot soccer (Figure 1) the response time is so short that the agent must develop a reactive model for its actions based on a very limited set of considerations [7]. This essentially decomposes the complex task with a huge state space into subtasks with fewer constraints. [1] The behaviour executed will differ depending on its objectives at the time - offence, defence, penalty kick, etc - which would be decided by a meta-level classifier. This results in a decomposition of the State Space into different regions for Offence, Defence, or Goal Striking, for example. A part of a meta-level classifier is shown as a decision tree in Figure 2.

Learning in a reactive model may be achieved using several mechanisms. We are currently testing the two following approaches which are detailed in Section 3:

1. Potential Fields : Compute different effects as a continuous field. Add to obtain overall Action Field. Tune fields using human experts.

2. Reinforcement Learning: Encode mapping from state to action as a sum of tuned Radial Basis Functions. Minimal supervision needed, but complex state-space search.

2 Cyclicity in Memoryless systems

Since reactive systems employ no memory of past events, they are often susceptible to cycles where a chain of actions is executed in an unending cycle. A reactive behaviour depends on its immediate stimulus s, and executes an action a resulting in some consequence c; we can therefore model a behaviour as the 3-tuple $< s, a, c >$.

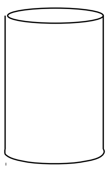
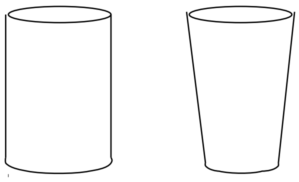
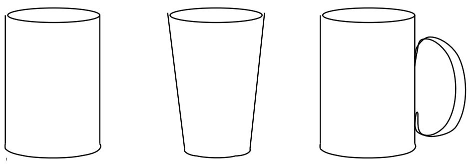
Stimuli for behaviors of increasing power		
(A)	(B)	(C)
		
$graspable(x) \wedge can(x)$	$graspable(x) \wedge (can(x) \vee cup(x))$	$graspable(x)$

Figure 3: *Power.* Behavior B can pick up cups as well as cans (has a more general stimulus). C is even more general. The notion of power is used to capture this: C is more powerful than B or A, provided all three behaviors have the same consequence.

[1]Even with a coarse position resolution of one robot-length and angle orientation of 10°, a game with two 3-robot teams would have a state space of $(20x15x20)^7$, which is too huge for handling today, but may become reachable within a few years.

To compare different behavior systems, we define a few relative measures.
Definition

- **Power :** A behavior ($\beta :=< s, a, c >$) is more powerful than ($\beta' :=< s', a', c' >) iff (s' \Rightarrow s) \bigwedge (c \Rightarrow c') \bigwedge (\beta \neq \beta')$.

In other words, the more powerful behavior can be triggered at least as frequently as the less powerful one and results in at least as strong a consequence. In Figure 3, behavior C is most powerful since it has the most general stimulus, though A, B and C all have similar consequences (the object x is picked up).

This argument can be generalized for the entire behavior space. A behavior space B is more powerful than the behavior space B' if B' can be obtained from B by replacing some behavior(s) $\beta \in B$ by less powerful behavior(s) β'.

- **Usefulness :** We define a *task* as a transition from one state of world to another, achieved by a temporal chain of behaviors. A behavior space B spans the task space τ if and only if $\forall (t \in \tau) (\exists (C \lhd B) fulfills(C, t))$. The *greatest potential task space* $\tau_G(B)$ is the largest task space that is spanned by the behavior space B. The *usefulness* of a behavior space is defined as the ratio $\frac{|\tau G(B)|}{|B|}$.

- **Modularity :** A behavior space is more modular if different modules in that space are more independent, in the sense of minimal interference between its modules. One measure of interference in a behavior space is the incidence of cyclic behavior. We therefore define the modularity of behavior space B as an inverse of the likelihood that cyclic conflicts will arise in B.

Behavior Modification Theorem.

Given two behavior spaces B and B' such that B is more powerful than B' (i.e. B' is obtained from B by replacing some behaviors β of B by the less powerful ones β') then:
(a) The greatest potential task space of behavior space B' is smaller than that of B, i.e.
$|\tau_G(B')| < |\tau_G(B)|$
(b) Usefulness of B is larger than that of B' i.e.
$\frac{|\tau G(B)|}{|B|} > frac | \tau_G(B') || B' |$.
(c)Likelihood of a cycle in B is at least as large as that for B'. (see [3] for more details).

The basic result then is that reactive systems may work in a limited domain, but attempts to broaden its range or output may result in temporal cycles. This has been actually observed in a number of robots (first reported in 1990, but frequently seen in any situation involving autonomous robots - e.g. soccer robots that knock on the boundary repeatedly). Thus, there is a need for a meta-level arbitrator or overseer, and how to design these is one of the issues on which current debate in Artificial Intelligence rages.
Modularity Theorem. Given cycle free behavior spaces B and B' (B' is obtained from B by replacing some behaviors of B by less powerful ones) and a module λ not belonging to B and B' is added to both of them, then the space B$\cup\{\lambda\}$ is at the most as modular as

the space $B' \cup \{\lambda\}$.

This result, which follows from the above, shows that for memoryless models of intelligence, the more powerful one tries to make a system, the more difficult it is to avoid destructive cycles. Also, when expanding the behavior repertoire, it is likely that new behaviors will need to be developed along with the old ones, many times at the cost of modularity.

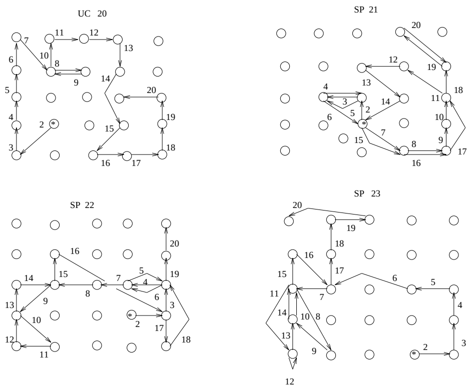


Figure 4: *Cycles in Memory-deficient monkeys.* The monkeys are looking for food under each trapdoor, starting with the one marked by a star. The test group, with a lesion in the sulcus principalis (SP), revisits some nodes several times, indicating some type of cycle in their behavior graph. Normal monkeys (UC = unoperated control) avoid such cycles. (Figure reproduced from [4].)

These results are of importance to the debate within the AI community on the role of reactive systems. However, in a broader sense, these also provide an important insight into the role of memory in general architectures of intelligence, which becomes clearer when we compare these results with those from biological systems under conditions of memory deficit, as with the lesions in the parts of a brain or its pathways (Figure 4).

Next, we consider the question of developing individual behaviours.

3 Learning Individual Behaviours

Any behaviour is a map from a set of stimuli into some actions. If this map is provided by the programmer, it can incorporate only those situations foreseen by the programmer, whereas if the behaviour can be learned, it can over time, handle a far varied set of stimuli than most prior programming. Here we shall consider two models of behaviour learning.

3.1 Supervised Learning using Groundtruth

In the first mode, a function is hypothesized for combining all the effects into a resultant action. Such a function should typically be continuous, and may be obtained as the deriva-

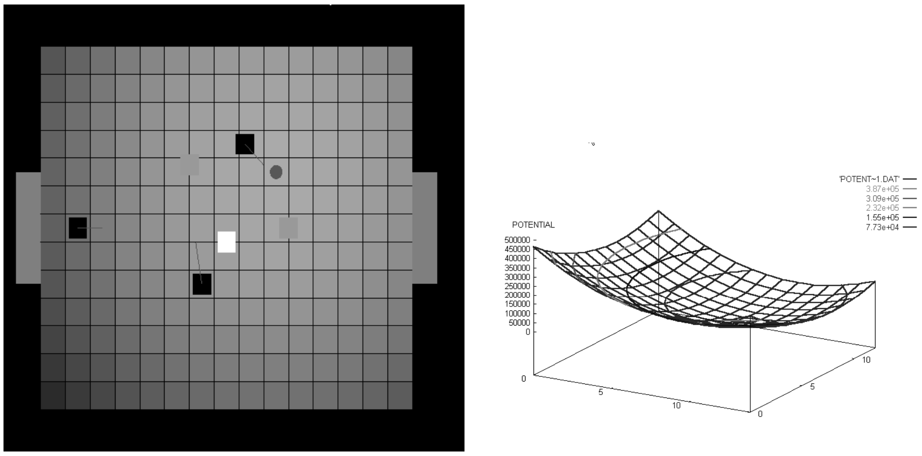


Figure 5: *Potential Field for ball attraction:* This field is lowest at the ball to attract the robot to get near the ball.

tive of a field, which is high in the undesirable zones and low otherwise. Such a model, called a potential function, may be learned based on a training set of soccer robot situations. For each setup in the training set a "groundtruth" or optimal action model is created by querying a number of users[2]. Designing such behaviours involves planning for the individual robot as well the team dynamics; thus the reactive function is defined on a state vector which consists of the position and orientation relative to goal for all six robots, and also the ball [5, 7]. Now, different sets of potentials need to be created for different team objectives (defence, pass receiving, striking, etc.). In our approach, this is done by defining the strategy as a sum of several potential fields (which are a form of multi-dimensional fuzzy functions). For example, the *Ball attraction field* (Figure 5) is higher away from the ball, so that the robot may have a tendency to go towards the ball. At the same time, the *Energy-Conservation Field* has a minima at the current robot position so that the robot does not move out too fast. Other fields include an *Occlusion Field* (Don't stay in regions where the ball can't reach), *Other Player field* (This is repulsive: Don't hit the other players), *Goal Obstruction field* (Stay between opponent and own goal) and a *Ball Obstruction field* (Stay between opponent and ball). Now, if the ball field is defined as

$$Field_{ball} = K_{ball} \times (dist)^{\eta \mathrm{ball}},$$

then learning consists of selecting the appropriate $K_{ball}, and \eta_{ball}$ that works best across the wide variety of situations. Such a tuning may be achieved using nonlinear optimization techniques such as Genetic Algorithms. (See [2] for more details). On the whole however, since the structure of the function is provided by the programmer, such models of learning are relatively weaker than general function learning systems such as Neural nets or Function approximators.

[2]The term "Groundtruth" arises in Satellite Image Processing, where it implies the correct labeling of zones in the image

3.2 Unsupervised Learning

In unsupervised modes of learning one does not depend on human experts to decide whether an action is good or bad. In **Reinforcement Learning**, the system obtains the result, and updates the expected reward $E[r_t] = R(\mathcal{S}_t \times \mathcal{A} \times \mathcal{S}_{t+1})$. We use Radial basis function (RBF) networks, which are a class of universal function approximators, to model the mapping from the state-action pair to an expected utility function. Unlike sigmoidal feed forward networks, RBF networks consists of a number of locally approximating units spanning a suitable amount of the input space. The output of these local units are combined linearly to give the final approximation.

Also, since the result of an action will be known only in a future epoch, rewards are passed back in time based on the probability of its reaching a consequent state with known utility $\mathcal{S}_{t+1}$. In such multi-epoch learning, the quantity to be optimised is

$$Return = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} \ldots = \sum \gamma^i r_{t+i}, \quad 0 < \gamma < 1$$

where γ is the discount factor for rewards in the future. Now the action value function $\mathcal{Q}$ depends on both the action $\mathcal{A}$ and the current state $\mathcal{S}$. Where the process dynamics cannot be explicitly modeled, a popular method is Q-learning, where the the action value function $\mathcal{Q}$ for some state,action pair is updated based on the probability of this action leading to a favourable outcome. Some results of Q-learning from striker mode are shown in figure 6. Other behaviours that have been developed and tested on simulated and real robots include goalkeeping, passing, and pass interception.

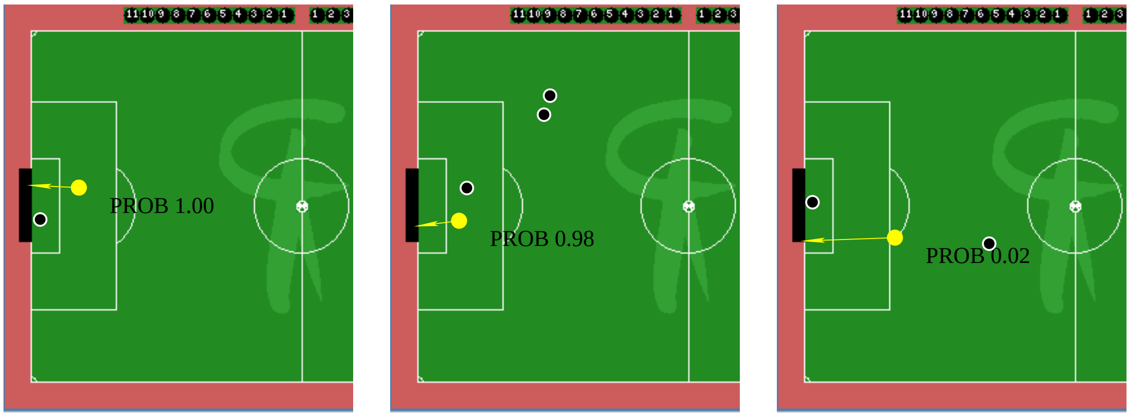


Figure 6: *Learning to shoot to goal.* The yellow robot is trying to score a goal in the presence of ideal opponents. Given a large set of results from actual play, the system models the most likely action and its success probability using a set of Radial Basis Functions. These rules are equivalent to fuzzy functions, resulting in a human-readable model.

4 Super intelligence and consequences for developing nations

The worldwide effort towards constructing intelligent agents, of which these results are but an insignificant part, has deep consequences, especially for countries like India. Machines already defeat the worlds best players at chess, and missiles can hone in onto targets based on sight. With exponential growth in computer speed, software performance, and electronics, it is but a matter of time before we are able to construct systems that would be able to defeat a team of humans in sports such as table tennis or even soccer. Such systems would not stop here however, but would keep on increasing in capability. It is unlikely that all global citizens will equally share in the benefits such technology may provide (and in its dangers); clearly developing nations like India have much to worry about from such explosive opportunities.

A nineteenth century scientist shown a robot soccer match involving teams of autonomous robots would probably call it sheer witchcraft, and much of India's populace today may also view it rather similarly. As has happened repeatedly in the history of man, such technological advances have unfortunately have had its deleterious uses. From the trenches we have but a limited view of the horizon ahead, and while the hazy picture contains much positive potential, there are also worrisome issues meriting considerable thought and discussion.

References

[1] Rodney A. Brooks. A layered intelligent control system for a mobile robot. In *Robotics Research,* ed O. D. Faugeras & Georges Giralt, pages 365–372. MIT press, 1986.

[2] Amitabha Mukerjee, Tamhant Jain, and Sreangsu Acharyya. Optimizing reactive soccer strategies by polling human experts. In *Proc. Intl. Symposium on Autonomous Robots and Agents (ISARA'2000)*, Singapore, May 23-26, 2000. [http://www.cs.iitk.ac.in/users/amit/pub/isara_potf.ps.gz]

[3] Amitabha Mukerjee and Amol D. Mali. Reactive robots are like amnesics : A comparative study in memoryless behavior. *IEEE Trans. Systems Man and Cybernetics*, May 1999.

[4] R. E. Passingham. Memory of monkeys (*macaca mulatta*) with lesions in pre-frontal cortex. *Behavioral Neuroscience*, 99(1):3–21, 1985.

[5] Peter Stone and Manuela Veloso. Multi agent systems: Survey from a machine learning perspective. Technical Report CMU-CS-97-193, 29 pages., CMU-RI, 1997.

[6] Jun Tani. Self-organization of symbolic processes through interaction with the physical world. In *IJCAI-95*, pages 112–118, 1995.

[7] Ashley Tews and Gordon Wyeth. Using centralised control and potential fields for multi-robot cooperation in robotic soccer. In *Proc. Pacific Rim Intl. Workshop on Multi-Agents (PRIMA),* Singapore, Nov 1998, pages 176–190.