A No-Strings Representation Theory for Adaptive Researchers

Chris Thornton
Cognitive and Computing Sciences
University of Sussex
Brighton
BN1 9QH
UK

Email: Chris.Thornton@cogs.susx.ac.uk
WWW: http://www.cogs.susx.ac.uk
Tel: (44)1273 678856

January 16, 1997

Abstract

Opinion is still divided over the role that internal world models can play in autonomous behaviour. Researchers who dispute the necessity of such models often have a restricted view of how they are constituted and may associate the whole enterprise of modelling with the dubious practices of GOFAI. However, this paper pursues Roitblat's approach [1] in developing a more general and less assumption-laden interpretation of what 'representationalism' means. It presents a no-strings theory of representation which shows why we should expect autonomous agents to use internal models and what these models will look like.

1 Introduction

Do autonomous agents really need internal world models? The adaptive behaviour community is still divided on the issue. Some follow Brooks' hard-line position arguing that explicit representations and models of the world are unnecessary and 'get in the way' [2,3]. Others feel that such models have a role to play but fear that their use inevitably leads back to GOFAI computationalism [4].

This paper presents a different approach to the debate. It provides a 'no-strings' theory of representation based on a simple efficiency argument. It shows why we should expect that autonomous agents will use internal representations and it shows what these representations will probably look like. However, the theory does not use any computationalist or connectionist ideas. Nor does it refer to GAs, embodiment, stochastic resonance or any other currently popular (or unpopular) approach. It is thus *paradigm-neutral.*

2 Neo-representationalism

The theory, which I call **neo-representationalism**, is based on an efficiency argument involving the concept of behavioural triggers. (A related argument is presented in [5].)

Imagine that we have an agent with a range of behaviours, one of which is called B. Let us call the set of internal phenomena which initiate this behaviour *B's internal trigger.* Assuming that B is a normal behaviour, i.e., not spontaneous or random, there must be some environmental phenomenon which triggers it. Let us call this phenomenon *B's external trigger.* We can now state the central idea of the theory.

- If external triggers are organised in structures, then the agent will save resources by replicating these structures internally.

To see why this is the case, consider the following example. Imagine that the agent has an 'attacking' behaviour and that this has as its external trigger the environmental phenomenon 'small mammal'; (i.e., the agent attacks when confronted with a small mammal.) Imagine also that the agent has a 'freezing' behaviour and that this has as its external trigger the environmental phenomenon 'large mammal'. (The agent freezes when it is confronted with a large mammal.) Finally, imagine that the agent has a 'fleeing' behaviour and that this has as its external trigger the environmental phenomenon 'small mammal with large mammal'. (The agent flees when confronted with a possible family group since in this case the large mammal may behave very aggressively.)

The external triggers here are organised in a *structure.* The situation which constitutes the 'fleeing' trigger is made up from (a) the situation which constitutes the 'freezing' trigger and (b) the situation which constitutes the 'attacking' trigger. Thus in implementing the internal trigger for the fleeing behaviour the agent must *somehow* replicate the internal trigger for freezing and the internal trigger for attacking.[1] It may *actually* replicate these triggers or it may simply re-use the originals. If the latter, then internal resources are saved.

In most contexts, internal resources are at a premium. So we can safely assume that agents will tend to re-use internal triggers where the presence of external structure makes this possible. But this assumption has interesting

[1] Of course, it is not the agent which does the implementing but rather the process which creates the agent, e.g., design, learning or evolution.

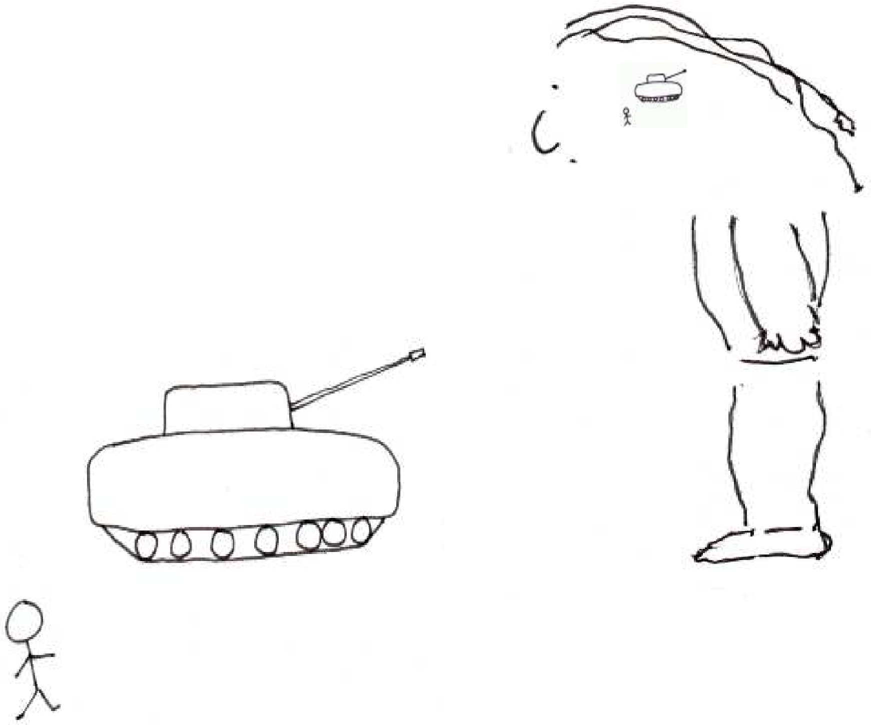

Figure 1: Ambush scenario.

consequences. In pursuing trigger re-use, agents must match up internal triggers with corresponding external triggers. This necessarily *replicates* the relevant external structure internally; i.e., it produces a structure of internal triggers which mirrors the structure of external triggers. Moreover, it enables higher-level internal triggers to exploit lower-level internal triggers as stand-ins for the relevant external phenomena. The trigger re-use strategy thus leads to (a) the production of internal structures which replicate external structures and to (b) the exploitation by some nodes in such structures (by other nodes) for *representational* purposes.[2]

The general idea is given a human slant in Figure 1 . Here we imagine that

[2] Some authors insist that a symbol which is used by an agent for representational purposes cannot be a part of the agent, i.e., that the agent cannot use a part of itself as if it were an external object [6]. However, this view seems rather suspect. Humans, for example, regularly treat parts of themselves as pseudo-external objects, e.g., when they 'use' their hands to carefully position their feet, or to comb their hair.

the agent is a human and that the external trigger for fleeing is 'ambush', the external trigger for freezing is 'tank' while the external trigger for attacking is 'infantry'. In other words, we imagine that we have a human agent which tends to attack isolated infantry, to remain frozen when confronted with a tank, and to flee the scene when confronted with an ambush comprising both tanks and infantry.

By the argument given, the agent will save resources by creating an internal model of the external situation. However this is not a model of the usual variety, i.e., some sort of caricature of the original which is lodged inside the head of the agent, as suggested in Figure 1 . Rather it is a model which replicates the salient *structural* properties of the external situation in a system of trigger re-use, as shown in Figure 2.

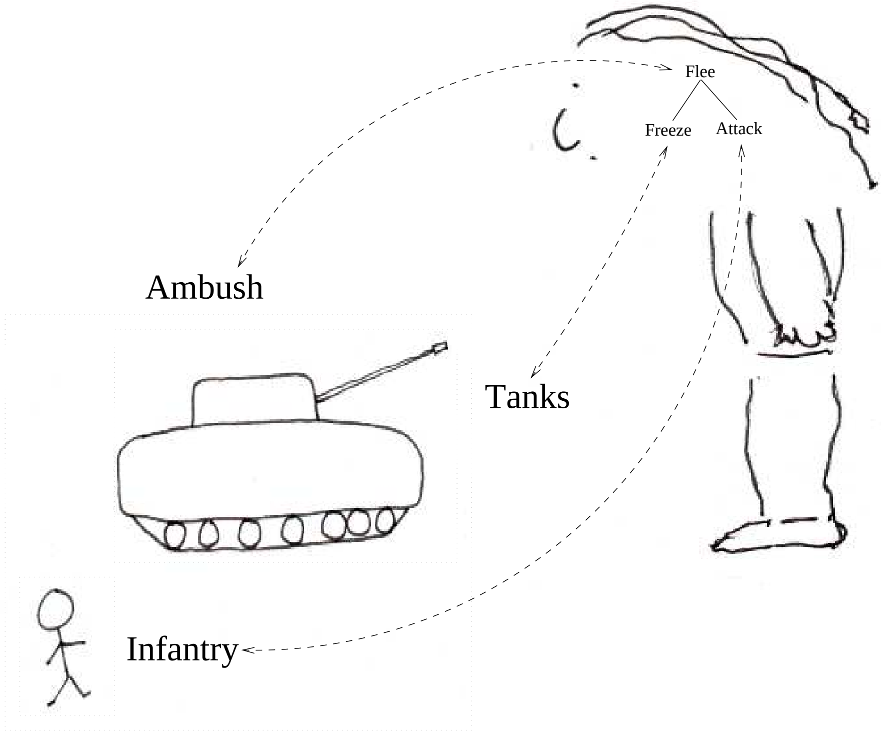

Figure 2: Structural model through trigger re-use.

3 Discussion

Many in the adaptive behaviour community may still be sceptical about the role that internal world models and representational mechanisms can/should/will play in the construction and explanation of autonomous agents. Some researchers may even feel that representationalist notions lead inevitably back into the murky waters of symbol processing and GOFAI. However, as Roitblat commented:

> There is no compelling reason to believe ... that ... representations must resemble the kind of word-like tokens that play a central role in strong symbol systems. Rather, organisms can use any number of alternative forms of representation. If experience at one time is to affect behaviour at another, then the organism must have some means of representing that experience. Some change in the organism must depend on the experience, which can influence later behaviour. Such changes are representations. [1]

The present theory adds a new twist to this. It shows that where resource constraints apply there will be pressure to arrange the 'representations' which Roitblat envisages in structural replications of external phenomena, and to enable higher level nodes in these replications to use lower-level nodes as stand-ins or symbols of external phenomena. The theory thus leads to a strong, representationalist position. However, it is essentially just an application of the old idea that representation affects processing efficiency. The implications of the theory apply to any autonomous agent engaged in the production of environmentally contingent behaviour.

Neo-representationalism has implications for both artificial and natural agents. It implies that artificial autonomous agents *should* use internal representations (of the described type) whenever internal resources are at a premium and the environment is structured. It predicts that natural autonomous agents (animals) *will* tend to use internal representations of the described type whenever (a) internal resources are limited and (b) the relevant evolutionary, learning or developmental processes allow the relevant trigger re-use strategy to be pursued. The theory makes no assumptions about what the triggers actually are, how they work or how they can best be described. In fact it makes no assumptions (beyond the pivotal one about resource efficiency) and is thus completely neutral with respect to choice of paradigm.

References

[1] Roitblat, H. (1994). Mechanism and process in animal behaviour: models of ANimals, animals as models. In D. Cliff, P. Husbands, J. Meyer and

S.W. Wilson (Eds.), *Proceedings of the Third International Conference on Simulation of Adaptive Behaviour* (SAB-94) (pp. 12-21). Brighton, UK.

[2] Brooks, R. (1991). Intelligence without representation. *Artificial Intelligence, 47* (pp. 139-159).

[3] van Gelder, T. (1992). What might cognition be if not computation?. Research Report 75, Bloomington, IN47405: Cognitive Science, Indiana University (Indiana).

[4] Prescott, T. (1994). Spatial learning and representation in animats. In D. Cliff, P. Husbands, J. Meyer and S.W. Wilson (Eds.), *Proceedings of the Third International Conference on Simulation of Adaptive Behaviour* (SAB-94) (pp. 164-173). Brighton, UK.

[5] Thornton, C. (1994). Emergent representation/green cognition. *Proceedings of DRABC-94: On the Role of Dynamics and Representation in Adaptive Behaviour and Cognition* (pp. 192-193).

[6] Harvey, I. (1992). Untimed and misrepresented. CSRP 245, Cognitive and Computing Sciences, University of Sussex.