

The integration and control of behaviour: Insights from neuroscience and AI

David W. Glasspool Advanced Computation Laboratory,
Cancer Research UK, Lincoln's Inn Fields, London,
and Institute of Cognitive Neuroscience, University College London.
david.glaspool@cancer.org.uk

Building a computationally specified theory of a human-level mind is an ambitious goal. How can the cognitive disciplines - artificial intelligence (AI) and cognitive psychology - contribute to such an undertaking?

Both psychology and AI have tended to study small areas of cognition and work with theories of single empirical phenomena. In a full scale cognitive theory two related issues must be addressed, those of integration (how are numerous cognitive theories or models organised into a coherent whole, rather than descending into behavioural chaos?) and control (how are these modules to be co-ordinated by an explicit goal?).

Within cognitive psychology rather few theories have emerged which attempt to deal with this breadth of human cognitive function. One which does is the framework proposed by Norman and Shallice (1986). In the context of the goal mentioned above this theory is interesting for a number of reasons.

Firstly, it takes the form of a layered architecture, with two or (in more recent versions) three layers of control operating simultaneously. There is a strong parallel with certain forms of layered agent architecture which have been particularly successful within AI, although the rationale for the architecture is very different in each case – on the one hand, observation of the slips and errors people make in everyday behaviour, and of the effects of neurological damage, and on the other a pragmatic need to find an efficient engineering solution to the problem of controlling an autonomous agent or robot.

Secondly, the lower layer of the Norman and Shallice framework is strikingly similar to an established AI model of action organisation (Maes 1991), although again the motivation for proposing the approach is different in each case – empirical observation versus engineering design.

Finally, the various elements of the Norman and Shallice framework are now sufficiently well specified that computational implementations have been produced of the entire framework. However this has been achieved in the case of the higher-level layer of the framework through analogy with yet another model drawn from AI.

As well as being a promising basis for a large-scale model of cognition, the Norman and Shallice framework thus presents an interesting example both of apparent theoretical convergence between AI and empirical psychology, and of the way in which theoretical work in both fields can benefit from interaction between them.

The proposed chapter explores these issues as follows:

1. The Norman and Shallice (N&S) framework compared with three-layer agent architectures.

The structure of the N&S framework is described, along with a brief review of the empirical evidence which motivates it. The framework comprises three layers of control: The lowest level is a

motor control system which is capable of relatively simple actions, but which closely couples sensory input with motor control to produce actions which are well tailored to their context. The middle layer is a Contention Scheduling system (CS) which learns habitual action sequences, hierarchically building more complex sequences from less complex ones. The highest layer, the Supervisory Attentional System (SAS), sets high-level goals for behaviour which CS carries out autonomously, and it can intervene directly to monitor behaviour or produce new behaviour sequences in novel or critical situations.

This control architecture is compared with a class of three-layer autonomous agent architectures primarily based on the work of Gat and colleagues (Gat, 1998), although related architectures have been developed by other groups. These architectures have been particularly successful in controlling implemented AI agents (eg NASA has used the approach in planetary rovers and an autonomous space probe). The motivations for the particular layered decomposition, the nature of the processes within the layers and the boundaries between them, and the nature of the interactions between the layers, are compared for both the AI and psychological models. The conclusion is that there is evidence for a core set of ideas which have been independently arrived at via the different approaches in each field.

2.The N&S framework compared with Maes' approach to action organisation.

The middle layer of the N&S framework, CS, is concerned with the problem of generating flexible sequences of action which further the agent's goals while accommodating to the particular environment in which the actions are being carried out. Alternative action sequences which can achieve the same goal compete on the basis of their activation level. Activation level is influenced by top-down input (specifying the current goals of the system) from the SAS, and by the state of the environment, so that methods of achieving goals are tailored to the objects available in the environment and the actions which they afford.

The particular approach taken in the CS system has much in common with that of Maes (1991). Maes' approach uses a network of nodes in which activation spreads both top-down, from goal nodes, and bottom-up, from nodes representing objects in the environment.

Maes' approach has been criticised as insufficient to provide reliable control of an agent carrying out realistic tasks (Tyrrell, 1993). However a number of related architectures have developed Maes' approach in ways which overcome these limitations (see Bryson, 2000, for a review). The N&S CS approach has been computationally implemented and applied to complex real-world tasks, and does not apparently suffer from the problems highlighted by Tyrrell with Maes' model. This is evidently in part because it extends the approach to allow hierarchical organisation of action sequences in which multiple behavioural routines may be active in parallel providing they do not compete for the same resources (such as effectors or items in the environment). The same approach is taken by more recent AI models based on Maes' system (eg Blumberg, 1996).

3.The N&S framework compared with the Domino agent model.

The second major element in the N&S framework, the Supervisory Attentional System, was initially somewhat poorly defined. Shallice and Burgess (1996) outlined the processes involved in the SAS and their relationships, based largely on neuropsychological evidence. However the picture remained unclear, with many processes under-specified. This is largely due to the difficulty of obtaining clear empirical data on "high-level" psychological processes which are relatively distant from the sensory and motor periphery.

Recently however progress has been made in specifying the operation of the SAS to the extent that

a preliminary computational implementation has been produced (Glasspool, 2000, Glasspool & Cooper, 2002, Shallice, 2002). This has been done by making an analogy between the SAS and an established model of executive function in AI, the “Domino” framework of Fox and Das (2000). The domino model provides a framework for processes of goal-setting, problem solving and plan execution which gives a promising initial fit to Shallice and Burgess's outline. A set of well understood and well specified formal semantics can be associated with the framework to render it computationally implementable. The domino thus provides an appropriate starting point for an SAS model. This has been developed by Glasspool (2000) and Glasspool and Cooper (2002) to provide a computational implementation of the most important processes in SAS and their interaction with the CS system.

This level of the N&S framework thus provides an example of the way in which a formal theory from AI can help organise an under-specified empirically-driven theory in cognitive psychology.

4. Discussion.

The discussion section reviews the work which has been done in producing computational implementations of the elements of the N&S framework and in modelling their interaction, and discusses its prospects as the basis of a useful architecture for engineering intelligent agents.

Some benefits and pitfalls of cross-disciplinary mapping of agent architectures are discussed. There is significant potential for insight in one field to apply directly in another, especially given the very different types of information used to motivate theories in AI and psychology. This is true at all levels of the cognitive system, but is perhaps particularly important in attempting to develop fully integrated theories of cognition as a whole. It is however important to guard against the dangers of finding parallels where none exist, or of over-generalising theories to make them “fit” to the extent that much which is important is lost.

References

Blumberg, B. M. 1996. *Old Tricks, New Dogs: Ethology and Interactive Creatures*. PhD thesis, MIT Media Laboratory.

Gat, E. 1998. On three layer architectures. In D. Kortenkamp, R. P. Bonasso and R. Murphey, eds. *Artificial Intelligence and Mobile Robots*. AAAI Press.

Bryson, J. 2000. Cross-Paradigm Analysis of Autonomous Agent Architecture. *Journal of Experimental & Theoretical Artificial Intelligence*, 12(2), 165-189.

Fox J and Das S. 2000. *Safe and Sound: Artificial intelligence in hazardous applications*. Cambridge, Mass.: MIT press.

Glasspool D. W. 2000. *The integration and control of behaviour: Insights from neuroscience and AI*. Paper in symposium "How to design a functional mind"; AISB Convention, 2000, and Technical Report 360, Advanced Computation Lab, Cancer Research UK.

Glasspool, D. W. and Cooper, R. 2002. Executive Processes. In R. Cooper (Ed.) *Modelling High Level Cognitive Processes*. New Jersey: Lawrence Erlbaum Associates. pp. 313-362.

Maes, P. 1991. The agent network architecture (ANA). *SIGART Bulletin*, 2(4), 115-120.

Norman, D. A. & Shallice, T. 1986. Attention to action: Willed and automatic control of behaviour. In R. J. Davidson, G. E. Schwartz, & D. Shapiro (Eds.) *Consciousness and self-regulation*, Vol. 4 (pp. 1-18). New York: Plenum Press. 1986.

Shallice, T. & Burgess, P. 1996 . The domain of supervisory processes and temporal organization of behaviour. *Philosophical Transactions of the Royal Society of London B*. 351, 1405-1412.

Shallice, T. 2002. Fractionation of the Supervisory System. In D. T. Stuss & R. T. Knight (Eds.) *Principles of Frontal Lobe Function*. Oxford University Press. pp. 261-277.