

**Simple Environments Fail as Illustrations of Intelligence:
A Review of R. Pfeifer & C. Scheier, *Understanding Intelligence*, Cambridge: MIT Press, 1999.**

Peter C. R. Lane and Fernand Gobet,
ESRC Centre for Research in Development, Instruction and Training,
School of Psychology, University Park,
NOTTINGHAM NG7 2RD, UK
{pcl, frg}@psychology.nottingham.ac.uk

The field of cognitive science has always supported a variety of modes of research, often polarised into those seeking high-level explanations of intelligence [2, 13] and those seeking low-level, perhaps even neuro-physiological, explanations [3, 9]. Each of these research directions permits, at least in part, a similar methodology based around the construction of detailed computational models, which justify their explanatory claims by matching behavioural data. We are fortunate at this time to witness the culmination of several decades of work from each of these research directions, and hopefully to find within them the basic ideas behind a complete theory of human intelligence. It is in this spirit that Rolf Pfeifer and Christian Scheier (hereafter, P&S) have written their book *Understanding Intelligence*. However, their aim is manifestly not to present an overview of all prior work in this field, but instead to argue forcefully for one particular interpretation – a synthetic approach, based around the explicit construction of autonomous agents. This approach is characterised by the Embodiment Hypothesis, which is presented as a complete framework for investigating intelligence, and exemplified by a number of computational models and robots to illustrate just how the field of cognitive science might develop in the future. We first provide an overview of their book, before describing some of our reservations about its contribution towards an understanding of intelligence.

Overview:

The book is a large one, exceeding 600 pages (this review refers throughout to the hardback version), and is divided into six parts. The first part sets the scene, describing what is meant by a ‘study of intelligence,’ and moves on to describe some classical approaches in AI and cognitive science. Chapter 1 begins by discussing a number of definitions of intelligence, covering topics as diverse as thinking and problem solving, natural language, abstract creativity (as characterised by the example of Albert Einstein), and the ability to survive within a natural environment. A number of criticisms of the classical information-processing approach to cognition are then presented in the next two chapters, relating to aspects such as the frame problem, symbol grounding and real-time behaviour.

The second part of the book introduces the basic concepts of ‘Embodied Cognitive Science.’ These stress the need for complete agents, operating in a ‘real’ environment; the agents should be autonomous, self-sufficient, embodied and situated (although at times, one or other of these requirements may be waived, such as self-sufficiency). This approach attempts to capture all pertinent interactions

between the agent and its environment by making the agent operate autonomously within that environment. Chapter 5 then introduces neural networks as the preferred choice for providing agents with adaptive behaviour.

The third part presents a number of typical examples and approaches exemplifying these concepts. This is one of the more interesting parts of the book, providing a good overview of topics such as: Braitenberg Vehicles, the subsumption architecture (mostly famous for its use in Brooks' robots, such as Genghis), artificial evolution and life, and dynamical systems.

The fourth (and by far the largest) part of the book looks toward a wider theory of cognitive science, with a discussion of five basic principles underlying intelligent systems, and a discussion of how these principles help to elucidate the phenomenon of memory. These principles (tentatively) set out implementation techniques for realising the construction of embodied cognitive systems, covering the need for: parallel processing, redundancy, a complete agent, sensory-motor coordination, a concept of value (or aim), ecological balance, and cheap design. This part of the book explores these principles in detail, and then presents a case study of their application to a high-level phenomenon, human memory.

The fifth part of the book considers issues of how an embodied system, once constructed, should be evaluated. This part is rather discursive, but is welcome in addressing this important issue; as P&S write, most early evaluations of embodied systems do not go beyond 'it works.'

Finally, the sixth part considers future directions, proposing an interesting test for embodied systems; to play a game of netball against a human team. This game stretches all the principles discussed in the text, requiring full embodiment and autonomy, but also a notion of team-playing; it would indeed be interesting to see when, and if, such a goal were met, and in what form. (We hope the outcome will be theoretically more satisfying than the Deeper Blue solution to the classical AI goal of beating the world chess champion.)

Treatment of prior literature:

Although already a large book, anything with the title 'understanding intelligence' must justify its particular selection of material carefully. To the current reviewers, the book's major contribution is its presentation of contemporary research on behaviour-based robotics. However, the book would have been much improved by making this its sole focus, because there are some major (and revealing) flaws in the treatment of issues relating more generally to intelligence. In this section, we indicate three shortcomings in the book's treatment of prior literature: first, the book's omission of the historical roots of behaviour-based robotics; second, the emphasis on a division between classical and embodied cognitive science; and third, the simplistic treatment of contemporary work in classical cognitive science.

Our first criticism is with respect to the historical roots of behaviour-based robotics. The book frequently claims that embodied cognition is a 'new field' (e.g. p. xiii, p. 297), initiated sometime in the 1980s as a reaction to the failings of classical AI. In actual fact, behaviour-based robotics *predated* classical AI, and it is surprising that the book makes no mention of such prior work. In the 1940s and

1950s many interesting embodied robots were constructed, before digital computers were a practicality, let alone widely available; at the time, the field was known as *cybernetics* [8]. Two of the first cybernetic models, known as *Cora* and *Speculatrix*, were constructed by Grey Walter [14]. *Cora* could learn conditioned reflex actions, such as associating a whistle with a succeeding light, and *Speculatrix* was a mobile ‘tortoise,’ employing trophic control mechanisms to search for sources of light. Further work in cybernetics explored topics such as: how groups of such robots would interact, classification directly from perception, maze running, and self-organised learning. Although it is to be expected that contemporary behaviour-based robotics should provide richer and more interesting results, a large amount of theoretical and empirical work was performed at this early date. The field then went into abeyance, as researchers began to explore classical AI using the newly available digital computer; programmable computers are considerably more flexible than hard-wired robots, and it was only natural to consider how they could generate complex behaviour over and above that achieved by their relatively simple cousins. The current resurgence of interest in behaviour-based robotics is therefore more a reflection of novel developments or technological interest in the field, rather than an intellectual revelation. It is tempting to draw a parallel with neural network research: initial investigation hits a technological barrier, but is revived once that barrier is crossed. The barrier in neural networks was the training algorithm (backpropagation) for networks with multiple layers. The barrier in behaviour-based robotics is less clear, but might be the presence of flexible robots for use in testing.

Our second criticism is more wide-ranging, and concerns P&S’s division between the embodied cognition approach and more classical AI/cognitive science approaches. The division seems largely based on the surface features of the model, e.g. whether it is a mobile robot as opposed to a computer simulation, rather than on any deeper methodological or functional principles. For instance, a specific behavioural trait or theory may be implemented (in a simulation program or robot) in at least three different ways, and all are exemplified both within this book, as well as within classical cognitive science:

1. We can understand the target behaviour, and then build a model to check this understanding. This approach is used in the *sahabot* robot, and frequently within classical AI, e.g. ACT-R models of memory [1].
2. We can build a model of a specific kind, and then observe the behaviour of the model. This approach is a favourite within dynamic systems research, and is exemplified by P&S with studies of Braitenberg vehicles. Within cognitive science, models such as Soar [13], constructed according to basic cognitive principles, are examined in novel domains to determine predictions of subject performance; these predictions may then be tested to validate or further develop the assumed principles.
3. We can build a model capable of learning or some other form of adaptive behaviour, and then train or evolve the model to exhibit a given behaviour. This more complex approach is designed to test our theories about learning, and not our understanding of the performance

itself (though the latter may be refined by the former). Examples in P&S's book include reinforcement learning algorithms, neural networks and genetic algorithms. Classical cognitive science also includes many examples, e.g. ACT-R [1], EPAM [6, 7], and Soar [13].

These considerations lead us to reexamine the novelty of the design principles for autonomous agents proposed by P&S in part four. Although these principles represent a synthesis of work within embodied cognition, they by no means differentiate the field from other work in cognitive science. For example, the three-constituents principle (p. 303) states that the definition of an autonomous agent must always include an ecological niche, a set of desired behaviours and tasks, and the design of the agent; this principle actually applies to any kind of working computational model. More discriminatory is the breakdown of the third constituent into seven separate principles, although again these are not specific to the kinds of model considered in this book. For instance, the role of parallelism is explicitly embodied within an architecture such as Soar [13], just as is the value principle, which enables the system to learn. What is unique to embodied cognition is a desire to combine all these principles into a single, operational model. Managing this combination only appears possible, based on the models described within this book, by restricting the type of behaviour or ecological niche within which the model can work. An insistence on this combination being present prevents the deeper exploration of basic concepts which perhaps rely only on one or other principle. For instance, the *sahabot* robot does not do food-foraging, and so is not a complete agent in the sense of doing all that an ant does; however, it is complete in the sense of successfully completing the task of navigation.

Our third criticism relates to the simplistic treatment of contemporary cognitive science, which at times borders on caricature. A good example is the topic of human memory, to which Chapter 15 is devoted. P&S propose that the storehouse metaphor, which dominates the field, is beset with six main problems (pp. 509-510): (a) the metaphor is sometimes used too literally; (b) observations of behaviour are often mixed with hypotheses about mechanisms; (c) research has largely neglected important issues about memory, such as ecological validity [12]; (d) a number of phenomena are difficult to explain within this framework, such as the effect of context, variability, and the encoding of skills; (e) the homunculus problem [4]; and (f) the fact that flowchart models, which do not indicate the mechanisms underpinning memory, are often used to describe human memory.

While these criticisms may apply to some non-computational psychologists, most are beside the point for a substantial group of researchers using computer models. Good examples of such models include ACT-R [1], EPAM [6, 7], and Soar [13]. Although these models are perhaps guilty of taking the storehouse metaphor seriously, each clearly separates behaviour and mechanisms; has been used to replicate phenomena with high ecological validity, such as car-driving, typing simulation, etc; readily addresses effects of context variability through a dynamic growth of memory structures; has no place for a homunculus; and has explicit mechanisms. Such models have been used as practical and effective components of intelligent tutoring systems [1], and for predicting performance on human-computer interfaces [10].

P&S's alternative is to see memory as recategorisation, while 'psychological research makes a clear distinction between categorization and memory' (p. 522). This overlooks the work of psychologists such as Estes [5] or Kintsch [11], who have developed mechanisms of information processing which simultaneously account for phenomena in both memory and perception. In addition, contemporary models such as ACT-R, EPAM or Soar all include such mechanisms. The real contribution of P&S is the emphasis on embodiment for studying memory and categorisation. Unfortunately, P&S do not offer any model or mechanism for explaining phenomena of interest to researchers in memory, while several information-processing models have led to important theoretical and practical contributions.

Robots or theories?

The attempt by P&S to encompass many aspects of cognitive science, AI and behaviour-based robotics into one book, though admirable, has been at the expense of clarity as to what issue is being explored. For instance, it is made clear that embodied cognition is about building robots exhibiting particular behaviours, and this approach is proposed as the way forward for cognitive science. This proposal conflicts with that adopted in a large number of alternative computational models of human learning and memory, models which have not been discussed by P&S, and so no direct comparisons have been made. The omission is striking, as such models have been widely tested in diverse applications, and we suggest that the desire to build fully autonomous robots has gained precedence over the desire to present a comprehensive set of principles underlying intelligence.

The distinction between a constructed robot and a plausible theory is exemplified by P&S's description of experiments with Braitenberg vehicles. For example, on pp. 185-6, an experiment is described in which three mobile robots with lights are free to move within an enclosed area; the robots are attracted towards each others' lights, and so tend to move together. An interesting behavioural characteristic of these robots is that at times one robot will become trapped against a wall by a second robot, prompting the third robot to come careering in, knocking the outside robot away, and so 'freeing' the first. How are we to understand this behaviour? The embodied cognition approach espoused by P&S requires the robots to be built and the described behaviour demonstrated. However, although an illustration of our reverse-engineering capabilities, this does not contribute much to our *conceptual understanding* of why the third robot 'freed' the first. The value of constructing the robots is to prevent any over-theorising, such as hypothesising the presence of mental states or intentionality beyond the needs of the phenomenon. However, that should not be a temptation in this simple case, as the robots do not exhibit more complex behaviour and so do not require a complex theory. Instead, the behaviour of the robots is readily understood in terms of the tendency of any robot to approach the brightest light source; when two robots are together, so are their lights, hence the third robot will tend to approach them. The freeing behaviour is thus a natural consequence of a low-level tendency to approach the lights. This theory also has predictive power, in that behaviour may be predicted for larger numbers of robots or robots with varying strengths of lights. In addition, it is independent of the particular construction of the

robots, and so could be applied to a broader class of robotic/animal behaviour.

Such an alternative level of understanding becomes increasingly imperative as the complexity of the behaviour being explained increases. In fact, most of the models described in this book adopt just this approach. For example, the *sahabot* robot (pp. 553-5) was constructed to model how a desert ant navigates its way back to its nest. So as to focus on the elements of navigation, a number of non-ant-like features are present: the robot is constructed with wheels, instead of legs; it is considerably larger than an ant; the robot only performs navigation, with no other food-searching or similar chores to perform; and certain known aspects of ant navigation are not included, such as path integration. Even so, due to the quality of the abstraction and the resulting conformance with the ant's behaviour, the robot is acceptable as a model of ant navigation. The model represents a confirmation that our understanding of ant navigation has covered the major factors, and takes into account the dynamic nature of the input stimulus and other interactions with the environment. It is important to appreciate that the robot itself is not the theory; if all we had as a theory of ant navigation was the finished robot, we would be in the same position of trying to understand its behaviour as we were with the ants.

Summary:

This book is useful in presenting an overview of contemporary research in behaviour-based robotics. In addition, it provides an introduction to a large number of topics within classical cognitive science and AI. However, our main criticism is that it tries to accomplish too much, and loses its way in transferring lessons derived from the models into theories of intelligence. If P&S are to be believed, a long history of work in traditional cognitive science and AI has contributed little to our understanding of intelligence. This view appears attractive when it is placed within simple dichotomies: the computer may have won the chess game or solved the algebra problem, which people require years to master, but try to make it walk across the campus... The idea that perceptual-motor skills must be understood so as to understand how a machine might walk across a campus, sit down at a chess board, and then play a game is counter to the abstractions evident in scientific methodology. What classical cognitive science (with non-mobile computers) has contributed is an understanding of how people interact with conceptual environments. Nothing in this book begins to challenge the role of such research. In future, it is to be hoped that the right way to integrate models of higher-level cognition will be uncovered so as to build useful (perhaps marketable) robots which utilise our theoretical understanding; perhaps the necessary techniques for this will be based on those within this book. Even so, the scientific understanding of a human's (or the robot's) intelligence is not going to be found in the engineering details, but instead at a higher, more conceptual, level.

References

1. J. R. Anderson, & C. Lebière (Eds.). *The atomic components of thought*. Mahwah, NJ: Erlbaum. (1998).
2. J. S. Bruner, J. J. Goodnow, G. A. Austin. *A study of thinking*. New York, NY: John Wiley & Sons, Inc. (1956).
3. P. S. Churchland, & T. J. Sejnowski. *The computational brain*. Cambridge, MA: The MIT Press. (1992).
4. D. Dennett. *Consciousness explained*. Boston: Little, Brown. (1991).
5. W. K. Estes. *Classification and cognition*. Oxford: Oxford University Press. (1994).
6. E. A. Feigenbaum, & H. A. Simon. A theory of the serial position effect. *British Journal of Psychology*, 53 (1962) 307-320.
7. E. A. Feigenbaum, & H. A. Simon. EPAM-like models of recognition and learning. *Cognitive Science*, 8 (1984) 305-336.
8. F. H. George. *The foundations of cybernetics*. Gordon and Breach: London. (1977).
9. D. O. Hebb. *Organization of behavior*. New York, NY: John Wiley & Sons, Inc. (1964).
10. B. John, A. H. Vera & A. Newell. Toward real-time GOMS: A model of expert behavior in a highly interactive task. *Behavior and Information Technology*, 13 (4) (1994) 255-267.
11. W. Kintsch. *Learning, memory, and conceptual processes*. New York, NY: John Wiley & Sons, Inc. (1970).
12. U. Neisser. *Memory observed. Remembering in natural contexts*. San Francisco: Freeman & Company. (1982).
13. A. Newell. *Unified theories of cognition*. Cambridge, MA: Harvard University Press. (1990).
14. W. G. Walter. *The living brain*. Duckworth. (1953).

Main text: 3040 words